Computer Science & Information Technology 140

David C. Wyld,
Natarajan Meghanathan (Eds).

Computer Science & Information Technology

8th International Conference on Computer Science and Information
Technology (CoSIT 2021),
March 27 ~ 28, 2021, Sydney, Australia.

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Natarajan Meghanathan (Eds),
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-925953-38-1
DOI: 10.5121/csit.2021.110401- 10.5121/csit.2021.110408

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 8th International Conference on Computer Science and Information Technology (CoSIT 2021), March 27 ~ 28, 2021, Sydney, Australia, 8th International Conference on Artificial Intelligence and Applications (AIAPP 2021), 8th International Conference on Signal and Image Processing (SIGL 2021), 7th International Conference on Cryptography and Information Security (CRIS 2021) and 2nd International Conference on Natural Language Processing and Machine Learning (NLPML 2021) was collocated with 8th International Conference on Computer Science and Information Technology (CoSIT 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CoSIT 2021, AIAPP 2021, SIGL 2021, CRIS 2021 and NLPML 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, CoSIT 2021, AIAPP 2021, SIGL 2021, CRIS 2021 and NLPML 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CoSIT 2021, AIAPP 2021, SIGL 2021, CRIS 2021 and NLPML 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Natarajan Meghanathan (Eds)

General Chair	Organization
David C. Wyld,	Southeastern Louisiana University, USA
Natarajan Meghanathan (Eds),	Jackson State University, USA

Program Committee Members

Abdel-Badeeh M. Salem,	Ain Shams University, Egypt
Abdelhak Merizig,	Mohamed Khider University, Algeria
Abdellatif I. Moustafa,	Umm AL-Qura University, Saudi Arabia
Abderrahim Siam,	University of Khenchela, Algeria
Abdulhamit Subasi,	Effat University, Saudi Arabia
Abhaskumarsingh,	Isroset, India
Abhishek Appaji,	B.M.S. College of Engineering, India
Abhishek Shukla,	R D Engineering College, India
Addisson Salazar,	Universitat Politècnica de València, Spain
Afaq Ahmad,	Sultan Qaboos University, Oman
Ahmad A. Saifan,	Yarmouk University, Jordan
Ahmed Elngar,	Beni-Suef University, Egypt
Ahmed Farouk Abdel Gawad,	Zagazig University, Egypt
AKhil Gupta,	Lovely Professional University, India
Alejandro Garces,	Jaume I University, Spain
Alexander Gelbukh,	InstitutoPolitecnico Nacional, Mexico
Ali Abdrhman Mohammed Ukasha,	Sebha University, Libya
Ali Asghar Rahmani Hosseinabadi,	University of Regina, Canada
Amal Azeroual,	Mohammed V University, Morocco
Amari Houda,	Networking & Telecom Engineering, Tunisia
Amel Ourici,	University Badji Mokhtar Annaba, Algeria
Amine Achouri,	University of Tunis, Tunisia
Amizah Malip,	University of Malaya, Malaysia
Amosa Babalola,	Federal Polytechnic Ede, Nigeria
Anirban Banik,	National Institute of Technology Agartala, India
Anita Dixit,	SDM College of Engineering and Technology, India
AntoanelaNaaji,	Western University of Arad, Romania
Aridj Mohamed,	Hassibabenbouli University, Algeria
Asif Irshad Khan,	King AbdulAziz University, Saudi Arabia
Atanu Nag,	Modern Institute of Engineering & Technology, India
Atul Garg,	Chitkara University, India
Azeddine WAHBI,	Hassan II University, Morocco
Benyamin Ahmadnia,	UC Davis, United States
Beshair Alsiddiq,	Prince Sultan University, Saudi Arabia
Bhagyashree S R,	Research ATMECE, India
Bilal Alatas,	Firat University, Turkey
Bin Cao,	Hebei University of Technology, China
Bouchra Marzak,	Hassan II University, Morocco
boukarinassim,	skikda university, algeria
Chang-Wook Han,	Dong-Eui University, South Korea

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

8th International Conference on Computer Science and Information Technology (CoSIT 2021)

8th International Conference on Artificial Intelligence and Applications (AIAPP 2021)

8th International Conference on Signal and Image Processing (SIGL 2021)

7th International Conference on Cryptography and Information Security (CRIS 2021)

2nd International Conference on Natural Language Processing and Machine Learning (NLPML 2021)

A DEEP LEARNING APPROACH TO NIGHTFIRE DETECTION BASED ON LOW-LIGHT SATELLITE

Yue Wang, Ye Ni, Xutao Li and Yunming Ye

Department of Computer Science,
Harbin Institute of Technology, Shenzhen, China

ABSTRACT

Wildfires are a serious disaster, which often cause severe damages to forests and plants. Without an early detection and suitable control action, a small wildfire could grow into a big and serious one. The problem is especially fatal at night, as firefighters in general miss the chance to detect the wildfires in the very first few hours. Low-light satellites, which take pictures at night, offer an opportunity to detect night fire timely. However, previous studies identify night fires based on threshold methods or conventional machine learning approaches, which are not robust and accurate enough. In this paper, we develop a new deep learning approach, which determines night fire locations by a pixel-level classification on low-light remote sensing image. Experimental results on VIIRS data demonstrate the superiority and effectiveness of the proposed method, which outperforms conventional threshold and machine learning approaches.

KEYWORDS

Night fire detection, pixel segmentation, low-light satellite image

1. INTRODUCTION

Wildfire is a severe threat to forests and human estate, which often takes place frequently due to the increase of dry fuel and the impact of extreme climatic conditions. A small wildfire could grow into a serious and big one if it is not detected in time. The problem is especially fatal at night, when the wildfires are less likely to be noticed. Hence, how to accurately detect night fires becomes an very important issue.

Low-light satellite, e.g., National Polar-Orbiting Partnership (NPP)/Visible Infrared Imaging Radiometer Suite (VIIRS), Defense Meteorological Satellite Program (DMSP)/ Operational Lines can System (OLS) and Luojia-1A Satellite, which can take remote sensing pictures at night, offers an opportunity to identify night fires. The low-light satellite image is a multichannel matrix, and its spectral bands span visible light, near infrared, short wave infrared and medium wave infrared. With the rich band information, night fires can be identified. Previous studies solve the problem based on threshold or conventional machine learning approaches, which are not robust and accurate enough. Recently, deep learning techniques have shown promising performance on conventional computer vision tasks. Hence, in this paper, we aim to develop a new deep learning approach, which can accurately detect night fires based on low-light satellite images.

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 01-09, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110401

In particular, we propose a pixel-level classification method based on recursive convolutional neural network for night fire detection. In our method, the spatial context is effectively exploited by the recursive convolution mechanism. Moreover, a squeeze excitation (SE) module is introduced to model the channel correlations for night fire detection. To validate the effectiveness of the proposed method, we conduct experiments on VIIRS satellite data. The experimental results show that the developed method performs better than conventional machine learning approaches, including Light GBM [1], random forest [2]. Moreover, it also outperforms existing deep learning techniques, e.g., multi-scale convolutional network and UNet [3].

2. RELATED WORK

Most of previous studies focus on the fire detection in small scale range. For example, in [4], a fire detection algorithm is developed based on the video data. In the method, the videos are from city monitors, where the covered area is quite small. Due to observation limitation, some studies attempt to detect fires based on low-light satellite images. The studies address the problem via a threshold segmentation. For example, in [5], an active fire detection algorithm is developed based on VIIRS. The method mainly analyzes the data characteristics and adopts some threshold rules to detect fires. However, the threshold based methods are not robust enough.

In addition to the threshold based methods, some machine learning approaches are also leveraged for fire detection upon low-light satellite data. For example, by extracting the multi-channel features, light GBM and random forest are applied to fire identification. However, the performance of the methods significantly relies on the manually constructed features.

Recently, with the rapid development and great success of deep learning techniques, its performance often beats that of the conventional machine learning algorithms. Moreover, it works in an end-to-end manner and does not need any manually constructed features. Though some deep learning algorithms have been developed for remote sensing tasks, none has ever touched the fire detection on low-light satellite. In this paper, we aim to develop a new fire detection algorithm based on the low-light satallite images.

3. THE PROPOSED APPROACH

In this section, we introduce the proposed deep learning approach to fire detection. Our notion is treating the fire location detection as a pixel level binary classification task. Hence, we build a deep learning approach to solve the problem. There are three key issues to be taken into account: (i) the spatial contexts are very important and should be effectively exploited; (ii) the channels contribute differently to the identification; (iii) the positive (fire pixels) and negative (non-fire pixels) examples are totally imbalanced. Next, we elaborate the proposed approach in the following four subsections. First, we introduce the developed multi-scale recursive convolution neural network unit. Second, the detection architecture is built upon the unit. Third, a squeeze excitation scheme is incorporated into the architecture. Finally, we apply the focus loss to learn the parameters in the architecture.

3.1. Multi-Scale Recursive Convolution Neural Network Unit

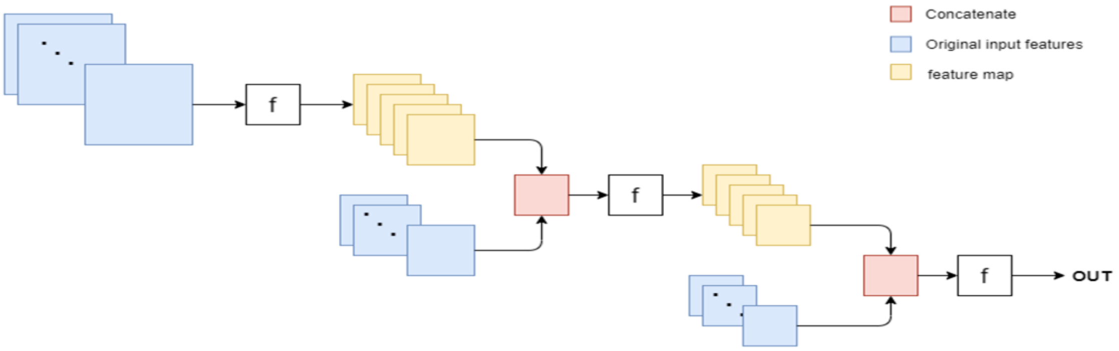


Figure 1. The Structure of Multi-scale Recursive Convolution Network Unit.

To detect the fire locations, we consider the task as a pixel-level classification problem. One of the key issues of the pixel-level classification is to effectively exploit the spatial contexts around the pixel. Inspired by [6], we develop the multi-scale recursive convolutional network unit. The main idea of the new unit can be illustrated as Fig. 1. We can see that given a multi-channel image patch, the new unit leverages three scales of convolutions to compute its output feature maps. In the first scale, a convolution filter is first utilized to compute the feature maps. Then, we crop the corresponding multi-channel input into the same size as the feature maps, and then combined it with the corresponding feature maps. In the second scale, the fused results are processed by the similar procedures to produce the output. Finally, the results in the second scale are further convolved with a filter to compute the third scale output, which is also the final output of the new unit.

The new unit can effectively exploit the spatial context due to its multi-scale recursive convolution. Moreover, in the new unit the contexts are gradually shrunk to the center point of each image patch. By doing so, the contexts can be exploited and the noise information is also effectively controlled at the same time, which is especially important for the pixel level based classification.

3.2. The Fire Detection Network Architecture

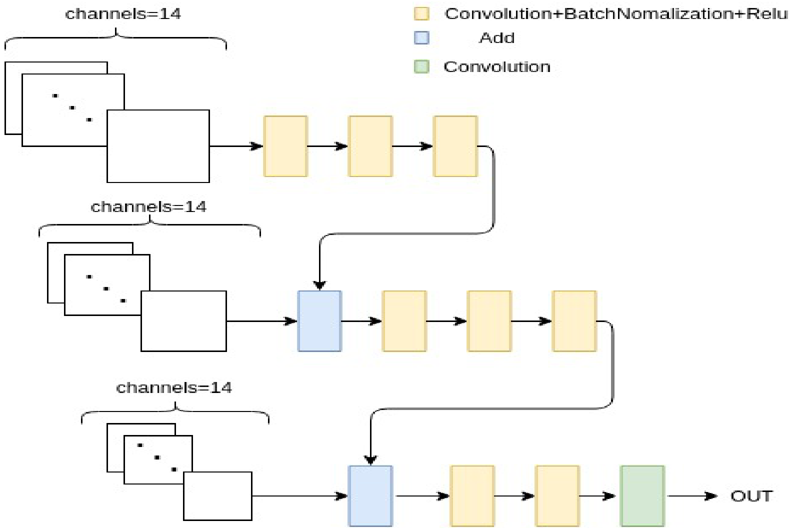


Figure 2. The Architecture of Fire Detection Network.

Upon the multi-scale recursive convolution network unit, we develop the fire detection network architecture. In our approach, we treat the fire location detection from low-light satellite images as an pixel based classification problem. For each pixel, we determine whether it is fire or not by extracting a small patch center at the pixel. With the small patch as input, a binary classification neural network architecture is established based on the multi-scale recursive convolution neural network unit, which is shown as in Fig. 2. We can see that given a 14-channel image patch as input, the patch is first convolved with three filter layers appended by a batch normalization layer and a rectified linear unit (ReLU) activation function. The input patch is cropped into appropriate size with the output feature maps and then added together in the second scale. In the third scale, the similar procedure is performed. With the output feature maps from the scale, a classification decision is made to determine whether the center point of the input patch is fire or not.

3.3. Incorporation of Squeeze Excitation Scheme

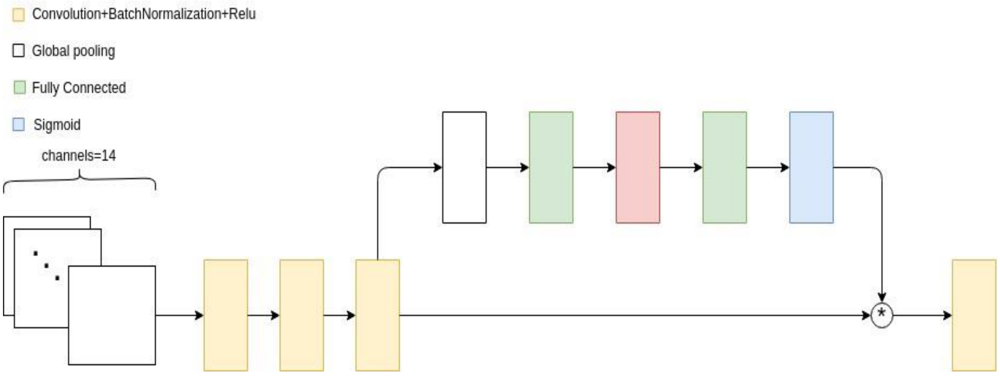


Figure 3. The structure of SE-block.

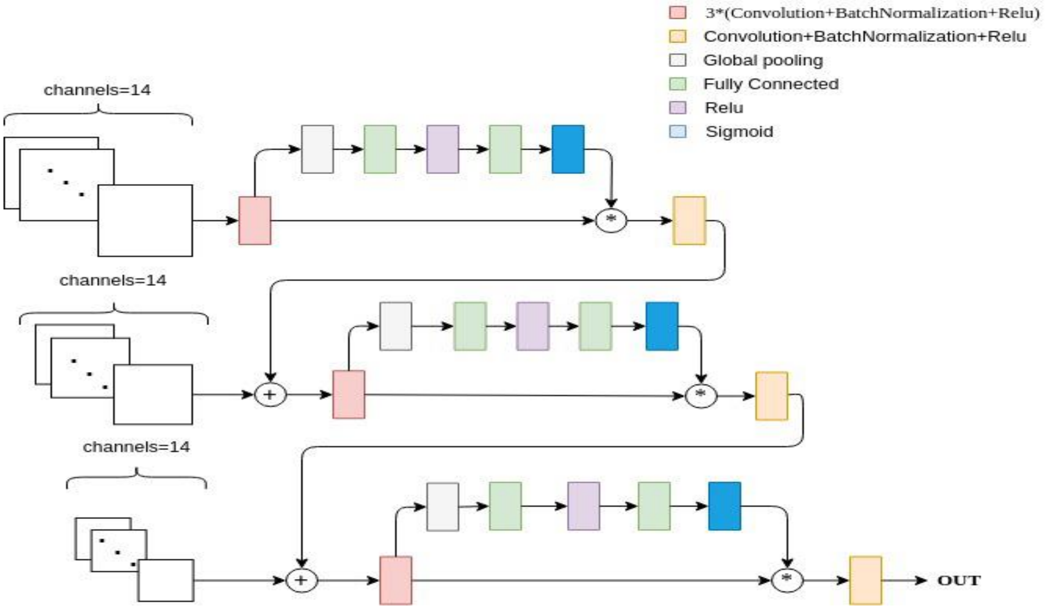


Figure 4. The structure of the multi-scale recursive convolution network model.

As shown in Fig. 3, the input low-light satellite image patch has multiple channels. In the channels, some are helpful for the fire detection while some are less important. Hence, the detection approach should effectively model the importance of different channels. Inspired by [7], we incorporate the squeeze excitation module into our fire detection network architecture.

The modified network architecture with squeeze excitation (SE) module is shown in Fig. 4. We can see the architecture is still a three-scale recursive convolution structure. Some modifications are made to incorporate the SE module. First, we use multiple convolution layers and activation function layers to extract the features. Then with the feature maps in each layer, we use a global pooling for sampling, and use multiple fully connected layers to characterize the contributions of different channels. The SE-module part outputs a multi-channel weighting matrix, which reserves the same size as feature maps. Finally, we perform the dot product with the feature maps and computed weighting matrix. As a result, the produced feature maps utility the weights computed, and channel importances are effectively modeled and fused.

Table 1 summarizes the detailed parameter settings of our fire detection network. We can see that the proposed fire detection network is mainly divided into three stages, and each stage consists of a similar network structure. In the stages, the model leverage inputs of different scales. The feature maps generated in the previous stage are combined with the low-light remote sensing image of the same scale to compute the network input of this stage. Finally, through this network we can obtain the fire detection results at the central point.

Table 1. The Structure of the Fire Detection Network.

Stage	Name	Layer	Filter	Stride	Output Size
	Input_1				31 * 31 * 14
First	Output_1	Conv1	5 * 5 / 32	1	27 * 27 * 64
	Output_2	Conv2	3 * 3 / 64	1	25 * 25 * 64
	Output_3	Conv3	3 * 3 / 128	1	23 * 23 *128
	Output_4	AvgPool	23 * 23		1 * 1 * 128
	Output_5	Fc1	128 * 8		1 * 1 * 8
	Output_6	Fc2	8 * 128		1 * 1 * 128
	Output_7 = Output_3 * Output_6				23 * 23 * 128
	Output_8	Conv4	3 * 3 / 64	1	21 * 21 * 64
	Output_9	Conv5	1 * 1 / 2	1	21 * 21 * 2
	Input_2				21 * 21 * 14 + 21 * 21 * 2
Second	Output_10	Conv6	5 * 5 / 32	1	17 * 17 * 64
	Output_11	Conv7	3 * 3 / 64	1	15 * 15 * 64
	Output_12	Conv8	3 * 3 / 64	1	13 * 13 * 64
	Output_13	AvgPool	13 * 13		1 * 1 * 128
	Output_14	Fc1	128 * 8		1 * 1 * 8
	Output_15	Fc2	8 * 128		1 * 1 * 128
	Outpu_16 = Output_12 * Output_15				13 * 13 * 128
	Output_17	Conv9	3 * 3 / 64	1	11 * 11 * 64
	Output_18	Conv10	1 * 1 / 2	1	11 * 11 * 2
	Input_3				11 * 11 * 14 + 11 * 11 * 2
Third	Output_19	Conv11	5 * 5 / 32	1	7 * 7 * 64
	Output_20	Conv12	3 * 3 / 64	1	5 * 5 * 64
	Output_21	Conv13	3 * 3 / 64	1	3 * 3 * 64
	Output_22	AvgPool	3 * 3		1 * 1 * 128
	Output_23	Fc1	128 * 8		1 * 1 * 8
	Output_24	Fc2	8 * 128		1 * 1 * 128
	Output_25 = Output_21 * Output_24				3 * 3 * 128
	Output_26	Conv14	3 * 3 / 64		1 * 1 * 64
	Output_27	Conv15	1* 1 / 2		1 * 1 * 2

3.4. The Loss Function

As a binary classification problem, one remainder issue is that the positive and negative samples are extremely imbalanced. In reality, the night fires take place at very few locations and time points. Hence, most of the low-light satellite images do not contain any positive examples. Only some of them include positive examples. The negative examples are very prevalent. Hence, we need to address the imbalance issue appropriately.

To tackle the issue, we adopt the focal loss [9] as our objective function. By a carefully modified cross entropy loss, the focal local loss can nicely deal with the skew positive and negative sample ratio. Specifically, it is formally computed as follows:

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\gamma} logy' \ , & y = 1 \\ 1-(1-\ \alpha)y'^{\gamma}\log(1-y')\,, & y = 0 \end{cases} \qquad (1)$$

Here α and γ are two positive parameters, and y and y' is the ground-truth class label and the prediction probability delivered by models, respectively. A ∈ (0,1)controls the class weights, which is utilized to adjust the imbalanced ratio between positive and negative examples. γ is a tunable parameter to adjust the loss. When γ is approaching 0, the loss resembles the cross entropy. When it becomes larger, the loss pays more attentions to indistinguishable part. The loss function is equivalent to cross entropy if we set α = 1/2 and γ = 0.

4. EXPERIMENTS

4.1. Experiments Setup

Data Sets. In the experiment, we leverage the M-band data from Suomi NPP VIIRS to test the models. The data includes 16 channels, which are emissive, reflective and temporal brightness channels. Each pixel in the image denotes a resolution of 750m. Missing value is very prevalent in Suomi satellite data. To tackle the missing values, we replace them with the average value of each channel. The ground truth fire locations are obtained as night fire product introduced in [9].

In the VIIRS Nightfire (VNF) the occurrence time and locations are recorded for each fire point at night. We leverage the information as our ground-truth labels and correspond them to the Suomi data. In the experiment, we utilize the VIIRS night low-light monitoring data from June 5 to 6, 2019 and October 2 to 6, 2019. As for a comparison, we leverage conventional machine learning approaches light BGM and random forest as our baselines. In addition, the deep learning techniques convolutional neural network (CNN), UNet are also compared.

Evaluation Metrics. To validate the performance of different methods, we leverage the widely utilized precision, recall and F1-score. The three metrics are computed as follows:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F_1 = \ 2 \times \frac{Recall \times \ Precision}{Recall \ + \ Precision} \quad (4)$$

Here *TP*, *TN*, *FP* and *FN* denote the numbers of true positive, true negative, false positive and false negative samples, respectively. In the three metrics, both precision and recall are biased while F1-score is a more comprehensive measure. Hence, F1-score is utilized to denote the overall evaluation. The higher the F1-score is, the better the performance is.

4.2. Experiment Results

Table 2 reports the experimental results of different models. We can see from the Table that UNet performs the worst, because the positive and negative examples are totally imbalanced. UNet fails to produce a promising segmentation. CNN is better but performs worse than conventional machine learning approaches light GBM and random forest. This is because: (i) the CNN model does carefully exploit the spatial contexts but utilizes them directly, where noisy contexts may hurt the performance; (ii) the imbalance issue of samples is not carefully considered. When equipped with the focal loss, the CNN model delivers better performance than lightBGM and random forest. Our proposed method yields better result than the CNN with focal loss, and its result is further improved when combined with the focal loss. All the results demonstrate the superiority and effectiveness of the proposed method.

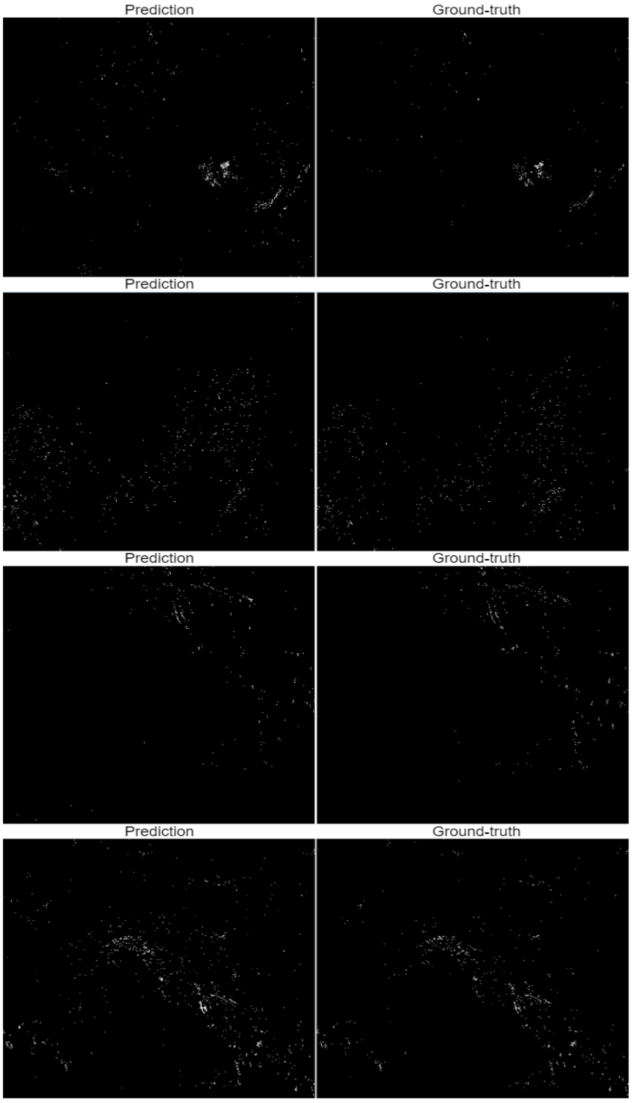


Figure 5. The night fire detection results of our method.

To visually examine the performance of our proposed method, we depict in Fig. 5 the detected locations of our model and the ground truth in four examples. We can see that the fire points detected by our method are quite consistent with the ground-truth locations, which further validates the effectiveness of the proposed method.

Table 2. The experimental results of different models on fire point detection tasks, where the symbol with * is the best result.

Method	Precision	Recall	F1-score
Random forest	0.756	1.0*	0.861
LightGBM	0.907	0.797	0.848
CNN	0.955	0.563	0.708
CNN + Focal Loss	0.893	0.871	0.882
UNet	0.294	0.165	0.211
Our method	0.983*	0.820	0.894
Our method + Focal Loss	0.949	0.929	0.939*

5. CONCLUSION

In this paper, we propose a new deep learning method to detect night fires based on the low-light satellite images. To effectively exploit the spatial contexts, a multi-scale recursive neural network unit is developed. Squeeze excitation module is incorporated in our method to characterize the channel importance. A focal loss objective function is adopted to tackle the sample imbalance issue. Experimental results on VIIRS low-light data set demonstrate the effectiveness and superiority of our method over existing techniques.

ACKNOWLEDGEMENT

The research was supported by the Shenzhen Science and Technology Program under Grant JCYJ20170811160212033.

REFERENCES

[1] black/white or gKe G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Advances in neural information processing systems. 2017: 3146-3154.

[2] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[4] Günay O, Taşdemir K, Töreyin B U, et al. Video based wildfire detection at night[J]. Fire Safety Journal, 2009, 44(6): 860-868.

[5] Schroeder W, Oliva P, Giglio L, et al. The New VIIRS 375 m active fire detection data product: Algorithm description and initial assessment[J]. Remote Sensing of Environment, 2014, 143: 85-96.

[6] Pinheiro P, Collobert R. Recurrent convolutional neural networks for scene labeling[C]//International conference on machine learning. 2014: 82-90.

[7] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[9] Elvidge C D, Zhizhin M, Hsu F C, et al. VIIRS nightfire: Satellite pyrometry at night[J]. Remote Sensing, 2013, 5(9): 4423-4449.

AUTHORS

Yue Wang received a bachelor's degree from South China University of Technology in Guangzhou, China in 2019, and started studying for a master's degree in Harbin Institute of Technology Shenzhen campus in 2019. His research direction is the application of semantic segmentation technology and satellite remote sensing.

Ye Ni received the B.Sc degree from Harbin Institute of Technology, Weihai, China, in 2019, and he is currently pursuing the M.Sc. degree in Harbin Institute of Technology, Shenzhen, China. His research interests include computer vision and remote sensing.

Xutao Li is now an Associate Professor in the Shenzhen Graduate School, Harbin Institute of Technology. He received the Ph.D. and Master degrees in Computer Science from Harbin Institute of Technology in 2013 and 2009, and the Bachelor from Lanzhou University of Technology in 2007. His research interests include data mining, machine learning, graph mining and social network analysis, especially tensor based learning and mining algorithms.

Yunming Ye received the Ph.D. in Computer Science from Shanghai Jiao Tong University. He is now a professor in the Shenzhen Graduate School, Harbin Institute of Technology. His research interests include data mining, text mining, and ensemble learning algorithms.

MANAGING THE COMPLEXITY OF CLIMATE CHANGE

Shann Turnbull

Principal: International Institute for self-governance, Sydney, Australia

ABSTRACT

This paper indicates how the knowledge of complex systems can be put into practice to counter climate change. A contribution of the paper is to show how individual behaviour, institutional analysis, political science and management can be grounded and integrated into the complexity of natural systems to introduce mutual sustainability. Bytes are used as the unit of analysis to explain how nature governs complexity on a more reliable and comprehensive basis than can be achieved by humans using markets and hierarchies. Tax incentives are described to increase revenues while encouraging organisations to adopt elements of ecological governance found in nature and in some social organisations identified by Ostrom and the author. Ecological corporations provide benefits for all stakeholders. This makes them a common good to promote global common goods like enriching democracy from the bottom up while countering: climate change, pollution, and inequalities in power, wealth and income.

KEYWORDS

Bytes, Climate Change, Common Good, Ecological Governance, Tensegrity

1. INTRODUCTION

The purpose of this paper is to present existing knowledge of how society might better counter the complexity of climate change. The causes and solutions of climate change are widely understood and accepted. The difficulty is how to motivate nations to take collective action. One way could be to share with citizens the knowledge required to constructively manage complex problems described as "the tragedy of the commons" [1].

These tragedies arise when different individuals or groups promote their self-interest by over exploiting common life-sustaining resources to eliminate them for everyone. The extermination of humanity on Easter Island is an example. Climate change introduces the risk of exterminating humanity.

1.1. Avoiding Tragedies of the Commons

For the first time the tragedy of the commons has become a global issue of human and institutional behaviour. Countless examples of such complex problems and their solutions have arisen over millenniums in the context of excessive hunting, fishing, grazing and irrigation. But their solutions have yet to be taught in graduate schools.

Political scientists Elinor Ostrom and her husband Vincent spent their lives studying how societies possessed societies have avoided the tragedy of the commons since pre-modern times. The solution did not depend upon either markets or State but by forming special types of complex

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 11-27, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110402

network relationships that introduce checks and balances on power elites whose actions could destroy the common good for everyone.

The Ostrom's used the language of political scientists to describe the nature of these networks as "polycentric republics" [2-9]. The decentralised and distributed communication and control architecture in such "polycentric republics" is also found in our brains [10] For this reason this form of governance can be described as "ecological" [11, 12, 13]. The knowledge on how to counter climate change becomes subjected to the natural science of governance in a way that also enriches democracy [14, 15].

The science of governance is grounded in contributions by Neumann [16] Shannon [17] and Ashby [18]. They indentified how to improve the reliability of data processing in respectively: decision-making, communications and control. This knowledge explains why and how nature creates complexity and how complexity can be best managed [20-27].

1.2. The Science of Governance

Governance science uses data as its unit of analysis. Data is routinely metered in bytes. Bytes are eight units of data called "bits". Bits are perturbations in matter and/or energy that make a difference. To minimise the materials and energy for living things to be created, developed, survive and reproduce in unknowable dynamic complex environments, evolution has developed processes for minimising the material and energy required.

According to [28] "The brain makes up 2% of a person's weight. Despite this, even at rest, the brain consumes 20% of the body's energy". The human brain is thousands of times more efficient than the most advanced computer chips that cannot match its performance even ignoring their dependence on external power sources [29, p. 9].

Unlike the social science of economics that seeks to minimise the undefinable social construct of cost, the science of governance is based on minimising materials and/or energy. In this way Transaction Byte Analysis (TBA) subsumes and extends the Transaction Cost Economics (TCE) developed by Coase [30] and Williamson [31] who limited the concerns only hierarchical organisations.

TBA provides a method for analysing any type of organisation and so any type of collective activity by humans or any other specie. This is because no collective action can occur in society or nature without data processing within and between coordinating entities.

Managing problems like climate change requires knowledge of how to manage complexity. This is common knowledge with natural scientists designing self-governing automobiles and space probes.

An introduction to this knowledge for social scientists is presented in the following sections. This knowledge provides a framework presented Section 3, for understanding why current forms of markets and hierarchy are ineffectual to counter climate change. Section four suggests how tax incentives can introduce ways to introduced ecological forms of organisations and different types of markets to counter climate change. Conclusions then follow in Section 5.

2. LANGUAGES AND ARCHITECTURE OF COMPLEXITY

2.1. Tensegrity

Words are the tools of thinking and special words are required to communicate special concepts to explain the complex communication and control architecture of ecological organisations. Mathews [32] identifies a number of special words in a review of the literature. But Mathews omitted an overarching concept of complexity called "Tensegrity". This feature is universal. It introduces inconsistent and paradoxical relationships in both physical and social structures. This allows novel relationships to arise to create new entities that are better suited in a new context while also reproducing paradoxical relationships to maintain evolutionary processes. A process inhibited by hierarchies, heterarchies or other types of relationships

Buckminster Fuller [33] coined the word "Tensegrity" by combing the words "tension" and "integrity". This concept has since been recognised by natural scientists but largely neglected by social scientists. One exception is Pound [34, 35, p.11] who recognised its need but not its name in stating: "always have an opposition viewpoint" and at p.18 "There must always be an opposition party and the prospect of insurgency".

2.2. The Architecture of Life and the Universe

Scientists like Harvard biologist Ingber [36] described tensegrity as "The Architecture of life" and quantum physicist Bohm [36] described the concept in different words as the architecture of the universe. Its relevance to social organisations was identified in the PhD dissertation of the author [38, pp. 8, 69, 134].

The science of governance explains why the laws of nature found in the physical world apply to individuals, society and its institutions. This explains the similarities noted between biology and economics tabulated in [38, p. 68]. Ashby [18, p. 1] explains why identical phenomena arise in both social and natural science by observing "The truths of cybernetics are not conditional upon them being derived from another branch of science. Cybernetics has its own foundations." The remit of cybernetics is "The science of communication and control in the animal and machine" [39]. The science of governance has subsumed the science of cybernetics by being the science of communication and control in the animal, machine and social organisations.

2.3. Holons and Holarchy

Mathews [32] identifies a key type of structure for creating or governing complexity that Koestler [39] called a "Holon". However, Hock [41] invented his own word "Chaord" from combining the words "Chaos" and "Order". In 1970, Hock became the founding CEO of the credit card company Visa International Inc. He created an organisation that meets the test of being composed of "polycentric republics". Visa was owned by its member banks with each bank having its own board of directors within a common legal entity. Each geographic board possessed the power to issue and manage its own Visa cards to create hundreds of "polycentric republics". Each "Republic" cooperated ~ competed with each other in their mutually owned legal entity.

Koestler coined the word "Holon" to describe an entity that is both a "Whole" of sub-systems and component of a larger system made up of Holons that he called a Holarchy. The Greek word for "whole" is "Holo with the suffix "on" being a component, like protons and electrons being components of an atom. Holons were what Smuts [42] and Simon [43] were describing with different words.

Holons possess quite different properties from hierarchies as revealed by Hock's [44, p. 30] description of a Chaord that he described in two different ways:

1. Any self-organizing, self-governing, adaptive, nonlinear, complex organism, organization, community or system, whether physical, biological, or social, the behavior of which harmoniously combines characteristics of both chaos and order.
2. An entity whose behavior exhibits observable patterns and probabilities not governed by the rules that govern or explain its constituent parts.

Hock described "chaordic" in three ways:

1. The behaviour of any self-governing organism, organization, or system, which harmoniously blends characteristics of order and chaos.

2. Patterned in a way dominated by neither chaos nor order.

3. Characteristic of the fundamental organizing principles of evolution and nature.

2.4. Other Cybernetic Approaches

Beer [45] pioneered the application of cybernetics analysis to management. He developed the Viable Systems Model (VSM) to describe any organizational structure that can produce itself and survive in a changing environment [46]. Because of their cybernetic heritage a number of VSM features are found in Holons, but the reverse does not apply.

Beer developed VSM before the concept of "corporate governance" became a discipline recognized by social scientists[1]. This made VSM subject to the discretion of management. It was not hard wired into organizational constitutions as found in organizations governed by polycentric subsystems in VISA, The John Lewis Partnership in the UK or the Mondragón Corporacion Cooperativa (MCC) in Spain.

Beer [47] was aware of the concept of Tensegrity and developed a synthetic form he described as "Syntegrity" [48]. But like VSM its introduction was at the grace and favor of management. Crucially VSM does not include the concept of Tensegrity that is a defining feature of Holons. While Mathews [32, pp. 52-53] does not use the word Tensegrity, he recognizes its existence and its special beneficial attributes by describing their contrary ~ complementary characteristics as a defining feature of a Holon. As examples, Mathews (pp. 41-44) refers to Holons as possessing: "Centralisation ~ de-centralisation", "Bottom-up ~ Top-down", "Autonomous ~ integrated", "Order ~ ambiguity", "Management ~ leader". This last feature does not communicate a contrarian relationship like the others. A better description would be to use the words: "Subordinate ~ leader" as arises for a Holon within a Holarchy.

2.5. Tensegrity Hidden from Management Scholars

Tensegrity naturally arises in mutual organizations from the conflicts arising within and between stakeholders. Tensions can arise between similar stakeholders, like the member banks of Visa,

[1] Beer met the author in Toronto on August 3, 1996, and a after reading a version of Turnbull [50] Beer advised that he had not extended his cybernetic insights to the governance of firms. Beer had been President of the World Organization of Systems and Cybernetic since 1987 and encouraged the author to publish in the Systems Science literature.

and/or between different stakeholders classes. Examples of the latter are: customers, distributers, suppliers, contractors, employees, executives, shareholders and host communities. Tensegrity is mostly extinguished in centralized command and control hierarchies. This could explain why management scholars and practitioners promote collegiate and cooperative relationships that obscure even further how Tensegrity is hard wired into human behavior who represent an organizational Holon. Some scholars are aware of the benefits of contested relationships like Pound and Jensen [50, p. 852] who reported on "The failure of internal control systems".
Different types of stakeholders may possess different interests in the firm that can create tensions, but the interests of such stakeholders are not typically formerly integrated into the governance architecture of firms. When they are, they like unlikely to possess meaningful power and/or influence to create serious tensions.

The MCC is an exception with multi-stakeholders interests participating in cooperative supervisory boards [51]. How these potential tensions can be organized to create internal challenges for continuous improvements and adaptations to new risks and opportunities are raised in the next section. This begins by considering the systemic problems inherent in simply hierarchies involved in complex activities.

3. WHY REPLACE HIERARCHIES?

The imperative to transform existing dictatorial command and control business hierarchies into polycentric networks of stakeholder republics arises because:

1. Hierarchies possess excessive exploitative powers that can corrupt their directors, managers, the business and society, [50; 852, 38, p. 115; 52, p. 9] and,
2. Humans possess limited physiological and neurological capacity to receive, process, store, process and communicates bytes, data, information, knowledge and wisdom to cope with complexity [31: p. 21], and,
3. Cybernetic laws of requisite variety that state it is impossible to:

    a) Reliably communicate complexity up or down a hierarchy either simply or reliably without a requisite variety of independent cross checking channels, [17] and,
    b) Reliably directly amplify control of complex variables without supplementary co-regulators providing a requisite variety of regulation [18, p. 265].

The above problems means that corporate governance codes supported by the World Bank, OECD, UK, US and around the world are promoting a system of exploitative governance, that is subject to failure because business and political leaders lack knowledge of governance science. The problems are not limited to publicly traded entities but also to private firms, government owned firms and even non-profit organizations.

How ecological governance can mitigate twenty systemic problems inherent in enterprises organized as simple command and control hierarchies are outlined in Table 1. "How mimicking nature can mitigate systemic problems in hierarchies". Details are provided in academic [12, 13, 14, 20-27, 38, 51, 53, 56-61, 63, 64, 66-70, 72, 74-76] and practitioner articles [11, 13, 52, 54, 55, 62, 65, 71, 73, 77].

3.1. Data Processing Limitations in Hierarchies

The practicality of firms transforming to polycentric republics is supported by their existence without any special laws in leading jurisdictions like the US, UK and Europe. A survey [78] of the internal architecture of stakeholder-governed corporations around the world by Bernstein [77] revealed that a common feature was a distribution of power to a number of boards and/or control centers. This suggests that the inherent conflicts of interest that can arise when workers can dismiss bosses and bosses can dismiss workers without a separation of powers do not allow such organizations to become sufficiently sustainable to be identified.

Table 1. How mimicking nature can mitigate systemic problems in hierarchies

	Toxic problems of hierarchies	Mitigation by mimicking nature
1	Society assumes top-down control is natural	Nature uses bottom/up control & top/down guiding
2	So no education about ecological governance with distributed control to simplify complexity	Complexity simplified with almost self-governing sub-systems dependent upon contrary guiding
3	Unitary boards obtain absolute power to identify and manage their own conflicts of interest to allow absolute corruption of directors, the business and society	Shareholders appoint one board to manage the business and a second to govern the corporation to establish tensegrity benefits for all stakeholders and society
4	Group think arises from directors captured by CEO to hide risks, misconduct & malfeasance	Governors/guardians of stakeholder voices obtain contested "requisite variety" of data for checks and balances
5	Corporations can lie and/or mislead themselves about director independence	Directors independence becomes irrelevant as Governors control minimized conflicts
6	Directors capture auditors who judge their accounts	Governors control auditors who judge directors accounts
7	Auditors lie that they are independent	Auditors kept independent by Governors
8	Accounting doctrines hide how investors get overpaid beyond their investment time horizons with surplus profits creating hidden sources of inequality and stakeholder exploitation	Ownership of surplus profits distributed by corporations issuing shares to citizen stakeholders that democratizes wealth and power. Reduces the need for corporate taxes and welfare programs
9	Directors control advisors to shareholders	Shareholder advisors controlled by Governors
10	Directors nominating themselves for election	Director nomination by shareholders & Governors
11	Directors control their own pay after setting and marking their own "exam papers" aka KPIs	Governors determine director pay from Stakeholder Key Performance Indicators (KPIs)
12	Directors control reports about corporate impact on the environment, stakeholders and community welfare and their own governance	Stakeholders provide guardians with reports for shareholders on Governors pay, corporate impacts on: stakeholders, the environment and society.
13	Directors control how they are held accountable to shareholders at AGMs and control the voting processes on own election and remuneration.	Stakeholder nominee controls conduct of AGMs. Governors determine AGM agenda, location, acceptance of proxy votes, vote counting, etc.
14	Directors ignorant of shareholder identities, etc.	All ultimate owners and/or controller made public
15	Share trading relationships and price manipulation hidden from directors and public	No shares traded without prior disclosure of any related derivatives and identity of counter parties
16	Shares traded covertly by third party exchanges and in "Dark pools"	Corporations directly execute all share transfers
17	Directors not held to account by various stakeholder groups who may have conflicting interest but on who directors rely upon to improve the quality, reliability, and efficacy of	Each common interest stakeholder group obtains rights to form their own non-profit associations to appoint advocates-supplementary regulators/ management

	continuous operational improvements	mentors that avoid directors and shareholders being kept in a cocoon of ignorance
18	Directors of simple command and control hierarchies lack systemic process to cross check management actions and misreporting	Directors obtain stakeholder communication and control channels independent of mangers to cross check integrity of operations and outcome reports.
19	Impossibility of controlling complexity directly	Complexity controlled indirectly by stakeholders
20	Self-regulation/governance is impossible	Self-governance shrinks costs & size of government & compliance costs.

Such separations of powers need not necessarily produce polycentric republics as possessed by Visa Inc, The John Lewis Partnership or the MCC. These organizations, and especially the MCC and the Citizen Utility Boards (CUBs) established by Ralph Nader [79] in the US provide working models for constructing Figure 1. Figure 1 does not represent any existing firm with polycentric republics. It is generic discussion model to illustrate some critical elements for introducing ecological governance.

A crucial essential feature is to introduce contestability and so Tensegrity into the governance architecture of business by introducing a requisite variety of a separation of powers. This also eliminates a number of systemic toxic conflicts of interest while at the same time decomposing decision-making and communications overload of directors. Such outcomes are supported by the groundbreaking work of Persson, Roland, and Tabellini [80]. They showed how an appropriate separation of powers provides net advantages to all constituent stakeholders.

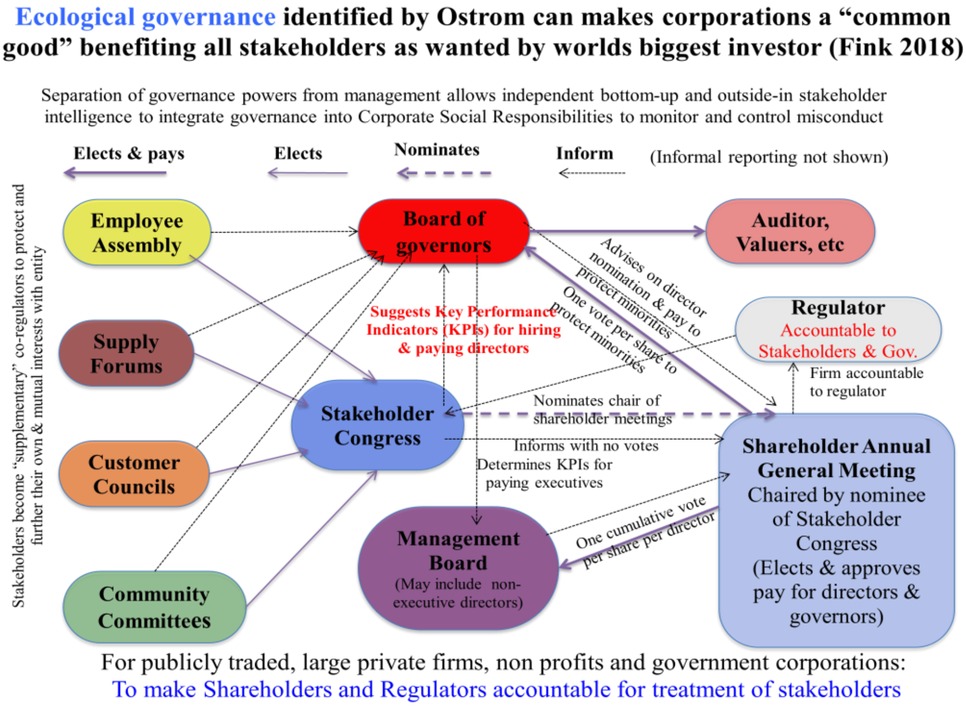

Ecological governance identified by Ostrom can makes corporations a "common good" benefiting all stakeholders as wanted by worlds biggest investor (Fink 2018)
Separation of governance powers from management allows independent bottom-up and outside-in stakeholder intelligence to integrate governance into Corporate Social Responsibilities to monitor and control misconduct
Elects & pays
Elects
Nominates
Inform
(Informal reporting not shown)
Stakeholders become "supplementary" co-regulators to protect and further their own & mutual interests with entity
Employee Assembly
Supply Forums
Customer Councils
Community Committees
Board of governors
Auditor, Valuers, etc
Regulator
Accountable to Stakeholders & Gov.
Firm accountable to regulator
Suggests Key Performance Indicators (KPIs) for hiring & paying directors
Advises on director nomination & pay to protect minorities
One vote per share to protect minorities
Stakeholder Congress
Nominates chair of shareholder meetings
Informs with no votes
Determines KPIs for paying executives
Management Board
(May include non-executive directors)
One cumulative vote per share per director
Shareholder Annual General Meeting
Chaired by nominee of Stakeholder Congress
(Elects & approves pay for directors & governors)
For publicly traded, large private firms, non profits and government corporations:
To make Shareholders and Regulators accountable for treatment of stakeholders

Figure 1

A complete separation of powers requires that no director who possesses the power to manage the enterprise is also involved in governing the corporate entity. The European two-tiered board would appear to achieve this objective but is compromised by supervisory boards also being accountable for management. Shareholders in Figure 1 are shown electing two boards and in different ways. The author has introduced this arrangement in two public companies he has founded [60, 81].

3.2. Separation of Powers Limits Corruption

A separation of the power to manage from the power to govern is typically introduced by Venture Capitalists (VCs) as a condition for their investment. This arrangement is also introduced in Leveraged Buy-Outs (LBOs). Jensen [50, p. 869] states that they are "proven models of governance structure" and "LBO associations and venture funds also solve many of the informational problems facing typical boards of directors." Legal scholar Dallas [82, 83] has presented arguments and proposals for a separation of corporate powers, as have Diermeier & Myerson [84].

Senator Murray [85] recommended to the Australian Parliament that all publicly traded companies should be required to separate the power to manage from the powers to govern by establishing the arrangements shown in Figure 1. Murray renamed the "Corporate Senate" [60] that only had veto power over conflicts of directors' interests to be come a "Corporate Governance Board" with executive powers over any conflicts of directors interests that included managing the AGM. Figure 1 suggests that the "Stakeholder Congress" appoints the chair of the AGM to avoid directors or governors being conflicted in their accountability to shareholders and other stakeholders.

3.3. Shareholder Primacy Subjected to Democracy by Other Stakeholders

The arrangement in Figure 1 allows shareholders to exercise control over the appointment and remuneration of both directors and governors with information provided independently of them. Shareholders obtain access to alternative views on the ability of the enterprise to provide benefits for all stakeholders [86]. Such contestability introduces Tensegrity into the conduct of AGMs.

There is no ethical commercial reason for managers of an organization to become both over-worked and conflicted by also being involved in the political process of governing the organization. US company director consultant, Ralph Ward promoted his business by publishing [65] that identifies how the arrangement described in Figure 1 can provide win-win outcomes for shareholders, directors, mangers, auditors, non-executive directors or governors and stakeholders.

Figure 1 describes, "A new way to govern: Organizations and society after Enron" [62] commissioned by the UK "Think and do tank" the New Economics Foundation as a public policy pocket book to identify an alternative to Thatcher privatization or State ownership. Columbia law professor Katharina Pistor [87] prescribed the pocket book for a course module at the Swiss International Law School [73]. The pocket book describes "A new model of corporate governance" [89] sought by the largest asset manager in the world to allow companies to provide "benefits for all stakeholders" [89]. All the 180 CEO's of the US Business Round Table who have Fink as a shareholder made a commitment the following year to provide benefits to all their stakeholders [86].

3.4. Shareholders Become Accountable for Stakeholders

However, the BRT made no mention for the need to also introduce "A new model of corporate governance". CEO's accountable to all stakeholders can become accountable to no one. This led some commentators [90] like Pistor to state that: "America's corporate leaders believe they can decide freely to whom they serve." A careful review of Figure 1 reveals that while stakeholder interests are represented in an influential manner, shareholder primacy is maintained to make shareholders accountable for protecting and furthering the interest of all other stakeholders. Achieving this objective would make corporations not only a common good but align their purpose in promoting the common good both locally and globally.

Figure 1 allows corporations to be become ethical [91]. In this way it creates the solutions identified in the right hand column of Table 1. Ecological governance also directly increases the wellbeing of individuals by allowing them to constructively use their embedded instincts to introduce checks and balances for spreading organisational self-regulation and self-governance [9, 75]. How such transformation might be achieved is next considered.

4. How Can Corporations Become a Common Good?

4.1. Tax Incentives

Elements of ecological governance could be introduced in various ways and stages to publicly traded corporations or other types of organizations in the private, public and/or non-profits sectors. These could arise from:

1. Governing bodies of organizations establishing a requisite variety of stakeholders advisory panels as shown in Figure 1 with any supporting geographical sub-units that besides increasing the depth and density of cross checking data from managers could also be used to promote Just In Time supplies, Total Quality Control for customers, Innovations by lead customers and/or users, employee voice (including anonymously to facilitate whistle blowers) and host community environmental and other feedback advice. Executives would be become accountable for making the entity a common good entity as proposed by the US BRT [86].
2. Governing bodies establishing the above processes in formal regulations of the organization to facilitate the establishment of contestability, challenged, feedback and organizational Tensegrity to promote continuous evolutionary improvements. Directors would become accountable for making the entity a common good entity.
3. Shareholders and/or members changing the constitution of their entity to introduce the arrangements outlined above to make shareholder and/or members accountable for transforming their entity to promote the common good with a "new model of corporate governance" such as proposed by Fink [89].
4. Regulators mandating changes in corporate constitutions as above and/or
5. The government providing tax, and/or other incentives to implement the above processes with and/or without ecological ownership to create common good enterprises that also reduces inequality by democratizing the wealth and control of nations bottom-up.

Ecological governance enriches democracy by directly engaging with broad constituencies of voting stakeholders. This is why it is important that only voters are recognized as stakeholders, not corporate entities with who they may be associated as employees, contractors or agents. It is by this means that democracy becomes enriched with a supplementary political process for individual voter to directly engage and participate in influencing institutions in the private, non-profit and government sectors. While Figure 1 maintains shareholders primacy, this can be extended to all stakeholders by introducing an ecological form of capitalism [12, 51, 53, 54, 55, 56, 59, 62, 71, 92, 93] that transfers ownership from shareholders to stakeholders after the time horizon of investors.

4.2. Democratising the Wealth and Control of Nations

The author's book, Democratising the wealth of nations, [92] was not about governance but ownership. It introduced the idea of Ownership Transfer Corporations (OTCs) as a way of creating an economic and political incentive for individual voters to become engaged in reforming capitalism[2].

At the suggestion of the founder of the UK conservative Think Tank ResPublica the author changed the name of OTC's to the more politically nuanced language of "Endowment Corporation" [71]. A short summary of the 1975 book was published by politically left Australian Think Tanks forty years after its publication [94]. Academic presentations of the idea of replacing exclusive, static and perpetual property rights with inclusive, dynamic and time limited rights are presented [12, 51, 93, 95] with the 1997 article republished in the Corporate Governance volume of This History of Management Thought [51].

The Central Research Institute for National Economy in Prague translated articles of the Author for his visits in 1991 and 1992 [52, 54, 55, 96]. In 1992 the State Commission for Reform of the Economic System hosted the author in Beijing to make presentations on using employee ownership as a technique for privation [53, 96].

The host was Professor Jiang Yiwei, an elected deputy to the National People's Congress and a member of its law committee. Yiwei had visited Yugoslavia in 1983 with his finding published the following year in China [98] on Yugoslavian worker self-management initiatives. Yiwei had promoted employee ownership in China in his 1988 book. The text of the book "From enterprise-based economy to economic democracy" was in both Chinese and English [99].

4.3. All Investments Except Land Have Limited Life

The majority of all business investments are time limited. Time limited property rights are not an alien, nor a provocative concept for venture capitalists and professional investors. A fact illustrated by the many Build Own, Operate and Transfer (BOOT) projects around the world. All intellectual property rights are time limited. A major intellectual problem is created by accountants assuming all business remain a "going a concern".

So unlike professional investors, accounting doctrines do not have time horizons. This means that investors can get overpaid receiving a cash return back after their time horizon. Any such return is by definition in excess of their incentive to invest to create a "surplus" profit. This is different and additional to any "super", "excessive" or "monopoly" profit that may be received and reported before their time horizon.

What is not reported is not managed, let alone taxed. As reported in the author's initial article [95] surplus profits can be many times greater than the value of the original investment. However, economists assume that there is no limit to greed. This denies them possessing a concept or word to describe profits in excess of the incentive to invest.

[2] The book was published by the politically right Company Directors Association of Australia and launched and reviewed by a socialist Dr. Jim Cairns in 1975. Cairns had a PhD in economics and at the time was the Deputy Prime Minister of Australia. His favorable review was published in the Journal of Australian Stockbrokers [114] providing evidence of the books bipartisan appeal. The publisher of the book had added an alternative title for their members being "New money sources and profits motives". Copies were sent to all members of parliament with a different covering letter that used the title most suited for the views of the recipient.

4.4. Surplus Profits Not Reported

Surplus profits are an invisible and insidious systemic source of the inequality created by capitalism not considered by economists like Picketty [100]. As surplus profits cannot be measured, the best way of making the economy more efficient and fairer is to introduce dynamic property rights to democratize the wealth of towns, cities, regions and nations [101-103]. Techniques for democratizing urban commons described in [92] and subsequent literature archived by the New Garden Cities Alliance [104].

As corporations are typically taxed at a lower rate than many citizens and have ways to shift profits to tax havens, the transfer of corporate equity to individuals can generate increased tax revenues and votes for political leaders. As shown in [51, 92, Appendix] only a relatively small tax incentive is required to provide investors with a bigger, quicker and less risky profits in return for gradually giving up ownership over the twenty year life a patent of say twenty years.
The endowment process can occur by shareholders agreeing to change their corporate constitutions to create a new class of Stakeholders shares. There is not limit to how many stakeholders shares can be issued but the percentage equity endowed each year can be constant. A process is established for any type of stakeholders shown in Figure 1 to become shareholders in a company that maintains shareholder primacy. The author's book [92] suggests a proportion of stakeholders shares are reserved to fund a minimum universal wellbeing income described as a "Social Dividend".

Endowment corporations would payout all their profits each year like many cooperatives. Business growth, management and investor succession could be provided for through dividend re-investment in "offspring" enterprises. Giant corporations would become replaced with nested networks of locally owned and controlled enterprises of human scale as illustrated by the MCC. Investors "fading out with a profit" [105, 106] would retain pre-emptive rights to continually be a shareholder in such networks of their choice.

5. CONCLUSIONS

Any incentive introduced to democratize the wealth of nations could also be used to democratize the control of corporate capitalism as indicated in Figure 1. This would enrich democracy on a bottom up basis with ecological governance. This would make self-governance practical to provide an additional way to reduce the size, cost and influence of governments [9, 107].

Perhaps the most challenging problem in countering climate change is educating both democratically elected leaders and those leading other nations in self-perpetuating command and control hierarchies that neither markets nor hierarchies are sufficient, or even necessary. This problem requires a compelling proportion of the general population to demand changes as outlined above. The challenge for universities it to initiate educational courses at all levels of society to facilitate decision-making.

The economics Nobel Prize Committee have done their part in recognizing Elinor Ostrom in 2009 for sharing the knowledge of organizing collective action to avoid tragedies of the commons. It is ironic that she shared the award with Oliver Williamson who spent his life researching "Markets and Hierarchies" [31].

Markets are the cause of the problem as noted by a former Chief Economist of the World Bank. As Lord Stern he reported to the UK government that: "climate change is the result of the biggest market failure the world has ever seen" [108]. This is because markets did not price the pollution cost of burning carbon. Carbon taxing and trading provide a way to counter market failure but stop markets creating counter productive messages.

5.1. Countering Climate Change by Transforming Markets and Hierarchies

To avoid markets creating messages to exacerbate climate change the definition of economic value needs to become by defined by the degree each bioregion of the world becomes sustainable for eternity. In other words the value of money in each region needs to be tethered to a Sustainable Index for each region [109-113] Hierarchies inhibit countering the problem of climate change as educational institutions and monotheistic religions unwitting reinforce the belief, that hierarchies are the natural of things. The opposite is the truth. As described by Hock [41, p.7]:

Industrial Age, hierarchical command and control pyramids of power, whether political, social, educational or commercial, were aberrations of the Industrial Age, antithetical to the human spirit, destructive of the biosphere and structurally contrary to the whole history and methods of biological evolution. They were not only archaic and increasingly irrelevant; there were a public menace.

Action research is required to test how best to introduce elements of ecological governance to institutions in the private, government and non-profit sectors. The immediate limitation is that no known graduate schools provide education in how to become a governance architect to lead implementation action. The biggest challenge is to disseminate this knowledge of how to sustain humanity on the planet.

ACKNOWLEDGEMENTS

The author would like to thank the Reviewers for their constructive suggestions.

REFERENCES

[1] Garratt Hardin, (1968) ‘The tragedy of the commons’, Science, Vol. 162, pp. 1243, 1244-1245.

[2] Vincent Ostrom, (1987) The Political Theory of a Compound Republic: Designing the American Experiment, Lincoln, NE, University of Nebraska Press.

[3] Elinor Ostrom (1990) Governing the Commons: The Evolution of Institutions for Collective Action, Cambridge University Press.

[4] Elinor Ostrom (1993) "Self-Governance, the Informal Public Economy, and the Tragedy of the Commons." In: Institutions of Democracy and Development. P.L. Berger, ed. San Francisco: ICS Press.

[5] Elinor Ostrom (1998) ‘Scales, polycentricity, and incentives: Designing complexity to govern complexity’, In: Protection of Global Biodiversity: Converging Strategies. L.D. Guruswamy & J.A. McNeely, eds. Durham, NC: Duke University Press.

[6] Elinor Ostrom (1998) ‘Self-Governance of Common-Pool Resources,’ In: The New Palgrave Dictionary of Economics and the Law, Vol. 3. P. Newman, ed. New York: Stockton.

[7] Elinor Ostrom (2010) ‘The Challenge of Self-Governance in Complex Contemporary Environments.’ The Journal of Speculative Philosophy 24(4): 316-332

[8] Elinor Ostrom (2012) ‘The Challenges of Achieving Conservation and Development.’ In The Annual Proceedings of The Wealth and Well-Being of Nations, 2011-2012, Volume IV: Self-Governance, Polycentrism, and the Social Order: Ideas and Influence of Elinor Ostrom. E. Chamlee-Wright, ed. Beloit, WI: Beloit College Press.

[9] Elinor Ostrom, Robert Walker & Roy Gardner (1992) The Political Theory of a Compound Republic: Designing the American Experiment, Plymouth, UK, Lexington Books

[10] Scott J. A. Kelso, Guillaume Dumas, & Emmanuelle Tognoli, (2013) ‘Outline of a General Theory of Behavior and Brain Coordination’, Neural News, 37, pp. 120-131.

[11] Shann Turnbull, (1991) Economics and the laws of nature, presented to Australian Conservation Foundation, 28 October. In: Allen Marston, ed. (1992) The Other Economy: Economics nature can live with, p. 39, Introduction pp. 3-60, Learn by doing publishers: Auckland, New Zealand.

[12] Turnbull, S. (2015) 'Sustaining society with ecological capitalism', Human Systems Management, 34 17-32, https://content.iospress.com/download/human-systems-management/hsm0831?id=human-systems-management%2Fhsm0831.

[13] Shann Turnbull, & Kent Myers, (2017) 'Shaping global cooperation with ecological governance', entry dated September 29th to the "Global Challenge - A new shape: Re-modelling global cooperation", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3103450.

[14] Shann Turnbull, (2002) The science of corporate governance. Corporate Governance: An International Review, 10(4): 256–272, http://ssrn.com/abstract_id=316939

[16] John, von Neumann, (1947) Theory of games and economic behaviour. CT: Yale Univerity Press.

[17] Claude, E. Shannon, (1949) The mathematical theory of communications, 1–94, Urbana, IL: The University of Illinois Press.

[18] W. Ross Ashby, (1957) An introduction to cybernetics. London: Chapman & Hall, <http://pespmc1.vub.ac.be/books/introcyb.pdf>.

[19] Ray Kurzweil, (1999) The age of spiritual machines: When computers exceed human intelligence. New York: Viking.

[20] Michael Pirson, & Shann Turnbull, (2011) Corporate governance, risk management, and the financial crisis - An information processing view. In: Corporate Governance: An International Review, 19(5): 459–470, available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8683.2011.00860.x/abstract>.

[21] Michael Pirson, & Shann Turnbull, (2011) Towards a more humanistic governance model: Network governance structures. Journal of Business Ethics, 99(3): 238-263.

[22] Michael Pirson, & Shann Turnbull, (2012) A new approach to fix broken governance. ISEI Insight 13: 28-35, available from: <http://www.ieseinsight.com/doc.aspx?id=1366&ar=3>.

[23] Michael Pirson, & Shann Turnbull, (2015) The future of corporate governance: Network governance - a lesson from the financial crisis. Human Systems Management, 34(1): 81-89.

[24] Michael Pirson, & Shann Turnbull, (2016) Decentralized governance structures are able to handle CSR induced complexity better, Business and Society, 1-31, DOI: 10. http://ssrn.com/abstract=2709413.

[25] Shann Turnbull, & James Guthrie, (2019) 'Holacracy – How Extinction Rebellion's success in organising protests is based on management science', Long Finance, November 1, https://www.longfinance.net/news/pamphleteers/holocracy-how-extinction-rebellions-success-organising-protests-based-management-science/.

[26] Shann Turnbull, & James Guthrie (2019) 'Simplifying the management of complexity: As found in nature', Journal of Behavioural Economics and Social Systems, 1(1): 51-73, https://papers.ssrn.com/abstract_id=3474786.

[27] Shann Turnbull, & Michael Pirson, (2019) 'The future of management: network governance', The European Financial Review, May 1, pp. 45-50, http://www.europeanfinancialreview.com/the-future-of-management-network-governance/.

[28] Factbook (2019) The physics factbook, 'Power of a Human Brain',9

[29] The Economist (2020) 'Technology Quarterly: Artificial intelligence and its limits', 13th June.

[30] Ronald, H. Coase (1937) 'The Nature of the Firm', Economica, 4(16), p. 403, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0335.1937.tb00002.x

[31] Oliver, E. Williamson, (1975) Markets and Hierarchies: Analysis and antitrust implications, Free Press, New York..

[32] John Mathews, (1996) Holonic organisational architectures. Human Systems Management, Vol. 15, pp. 27–54.

[33] Buckminister Fuller, (1961) 'Tensegrity', Portfolio and Art News Annual, No.4 http://www.rwgrayprojects.com/rbfnotes/fpapers/tensegrity/tenseg01.html

[34] John Pound, (1992) Beyond takeovers: Politics comes to corporate control. Harvard Business Review, March-April, pp. 83–93.

[35] John Pound, (1993) The rise of the political model of corporate governance and corporate control. New York University Law Review, Vol. 68, no. 5, pp. 1003–1071.

[36] Donald, E. Ingber, The architecture of life, Scientific American, pp. 30-39, January

[37] David, Bohm, (1980) Wholeness and the implicit order, Routledge & Kegan Paul: London.

[38] Shann Turnbull, (2000) The governance of firms controlled by more than one board: Theory development and examples. PhD Thesis, Macquarie Graduate School of Management, https://papers.ssrn.com/abstract_id=858244.

[39] Norbert, Wiener (1948) Cybernetics: Or Control and Communications in the Animal (and the Machine, Cambridge, MIT Press.

[40] Arthur Koestler, (1967) The ghost in the machine, London: Hutchinson.

[41] Dee Hock, (1995) The Chaordic Organization: Out of Control and Into Order, World Business Academy Perspectives, Vol. 9, No. 1, p. 7, https://www.ratical.org/many_worlds/ChaordicOrg.pdf

[42] Jan Smuts (1925) Holism and evolution. London & New York: Macmillan.

[43] Herbert, A. Simon, (1962) The architecture of complexity, Proceedings of the American Philosophical Society, Vol. 106, pp. 467–482.

[44] Dee Hock, (1999) Birth of the Chaordic Age, Berrett-Koehler Publishers, San Francisco.

[45] Stafford Beer (1959) Cybernetics and management. English University Press; London.

[46] Stafford Beer (1995) Brain of the firm, 2nd edn, John Wiley & Sons, Chichester: England.

[47] Stafford Beer (1994), Beyond dispute. John Wiley & Sons Inc: New York City.

[48] Angela Espinosa, & Robert Hardin (2007). Team syntegrity and democratic group decision-making: Theory and practice, Journal of the Operational Research Society, Vol. 58, pp.1056-1064, August, https://www.tandfonline.com/doi/abs/10.1057/palgrave.jors.2602261?journalCode=tjor20

[49] Shann Turnbull, (1995) Innovations in corporate governance: The Mondragón experience', Corporate Governance: An International Review, Vol. 3, No. 3, pp. 167-180, http://papers.ssrn.com/sol3/paper.taf?ABSTRACT_ID=6455

[50] Michael, C. Jensen, (1993) The modern industrial revolution: Exit and the failure of internal control systems, The Journal of Finance, Vol. 48, No.3, pp. 831-880.

[51] Shann Turnbull, (1997) Stakeholder Governance: A cybernetic and property rights analysis, Corporate Governance: An International Review, Blackwell, Vol. 5, No. 1, pp. 11-23, http://papers.ssrn.com/sol3/paper.taf?ABSTRACT_ID=11355

[52] Robert Monks, & Allan Sykes, (2002) Capitalism without owners will fail: A policy maker’s guide to reform, New York, Centre for the study of financial innovations, Vol. 57 November.

[52] Shann Turnbull, (1991) ‘Property Rights and Markets’, in: Economic Alternatives for Eastern Europe Briefing No. 6, New Economics Foundation, London.

[53] Shann Turnbull, (1991) 'Re-inventing Corporations', Human Systems Management, Vol. 10, No. 3, pp. 169-186.

[54] Shann Turnbull, (1991) 'Socialising Capitalism', in: Stuart M. Speiser, ed., Equitable Capitalism: Promoting Economic Opportunity Through Broader Capital Ownership, Chapter 9, pp 97–113, New Horizons Press: New York,

[55] Shann Turnbull, (1991) Statická Nebo Dynamnická Vlastnická Práva (Static or Dynamic Property Rights?) Central Research Institute for National Economy, Prague, Kveten,

[56] Shann Turnbull, (1993) 'Democratic Capitalism; Self-financing local ownership and control', Human Systems Management, 12(4): 333-348

[57} Shann Turnbull, (1993) 'Flaws and Remedies in Corporatisation and Privatisation', Human Systems Management, Vol. 12, No. 3, pp. 227-252.

[58] Shann Turnbull, (1995) ‘Best practices in the governance of GBEs’, in: J. Guthrie, ed. Making the Australian Public Sector count in the 1990s, pp. 99-109, IIR Conferences: Sydney.

[59] Shann Turnbull, S. (2000) 'Stakeholder Governance: A cybernetic and property rights analysis'. In: R. I. Tricker, ed. Corporate Governance: The history of management thought, pp. 401–413, Ashgate Publishing: London.

[60] Shann Turnbull, (2000) ‘Corporate Charters with Competitive Advantages', St. Johns Law Review, Vol. 74, No. 44, pp. 101–159, http://papers.ssrn.com/sol3/paper.taf?ABSTRACT_ID=10570.

[61] Shann Turnbull, (2001) ‘The competitive advantage of stakeholder mutuals’: In The New Mutualism in Public Policy, ed. J. Birchall Chapter 9, pp. 171–201, Routledge, London. http://ssrn.com/abstract=242779

[62] Shann Turnbull, (2002a) A new way to govern: Organisations and society after Enron. New Economics Foundation, London, https://papers.ssrn.com/abstract_id=319867

[63] Shann Turnbull, (2005) ‘The use of bytes to analyse complex organizations’. In: Kurt Richardson, ed., Managing the Complex: Philosophy, theory and applications, Chapter 9, pp. 152-165, Charlotte, NC: Information Age Publishing Inc., http://ssrn.com/abstract=645281

[64] Shann Turnbull, (200a) 'The Science of Governance: A Blind spot of risk managers and corporate governance reform', Journal of Risk Management in Financial Institutions, Vol. 1, No. 4, pp. 360–368.

[65] Shann Turnbull, (2012) 'Discovering the "natural laws" of Governance'. In: ed. Ralph Ward, The Corporate Board, March/April, Vanguard Publications Inc.: Okemos, MI, http://ssrn.com/abstract=2062579

[66] Shann Turnbull, (2012) 'The limitations in corporate governance best practices'. In Thomas Clarke & Douglas Branson, eds. Handbook of Corporate Governance, Chapter 19, pp. 428–449, Sage: London & Thousand Oaks, CA, http://ssrn.com/abstract=1806383

[67] Shann Turnbull, (2013a) 'Achieving environmental sustainable prosperity'. In: Karagiannis Nikolaos & John Marangos, eds, Toward a Good Society in the Twenty-first Century: Principles and Policies, Part II Sustainability, Ecology, and Good Society, Chapter 4, pp. 75–103, New York, NY: Palgrave Macmillan, http://ssrn.com/abstract=1769349

[68] Shann Turnbull, (2013) 'A sustainable future for corporate governance theory and practice' in S. Boubaker, Bang D. Nguyen & Duc K. Nguyen, eds. Corporate Governance: Recent Developments and New Trends, pp. 347–368, Springer-Vertag, Heidelberg, http://ssrn.com/abstract=1987305

[69] Shann Turnbull, (2013) 'A transaction byte paradigm for researching organizations'. In: Giulia Mancini & Mariarosalba Angrisani, eds., Mapping Systemic Knowledge, pp. 243–267, Lambert Academic Publishing, Germany, http://www.volsu.ru/download.php?id=00000031271-1.pdf

[70] Shann Turnbull, (2013) 'How can non-profit organizations enhance performance and legitimize their operations? Presented to: 9th European Institute for Advanced Studies in Management, University of Lund, Sweden, June 13-14, http://ssrn.com/abstract=2223032

[71] Shann Turnbull, (2014a) 'A proposal for self-governing corporations'. In: Philip Blond, ed., The Virtue of Enterprise: Responsible Business for a New Economy, pp. 52-54, January, ResPublica: London, http://www.respublica.org.uk/wp-content/uploads/2014/01/jae_The-Virtue-of-Enterprise.pdf

[72] Shann Turnbull, (2014b) Designing resilient organisations: With operating advantages for public, private, non-profit and government entities and their stakeholders, Lambert Academic Publishing: Saarbrücken, Germany. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2473393

[73] Shann Turnbull, (2015) Introduction to the Practices, Science, and Art of Designing Corporate Charters, https://vimeo.com/137118382/1d7e82ce27

[74] Shann Turnbull, S. (2016) 'Defining and achieving good governance', in Güler Aras & Coral Ingley (eds.), Corporate Behavior and Sustainability: Doing Well by Being Good, Chapter 13, pp. 232-249, Ashgate Publishing Ltd., UK, http://ssrn.com/abstract=2571724

[75] Shann Turnbull, (2018) 'A vision for an eco-centric society and how to get there', The Ecological Citizen, Vol. 1, No. 2, pp. 141-142, http://www.ecologicalcitizen.net/pdfs/Vol%201%20No%202.pdf.

[76] Shann Turnbull, (2019) 'Causes and solutions for misconduct in financial services industry,' Law and Financial Markets Review, Vol. 13, Nos. 2-3, ppl 99-113, 16 April, https://doi.org/10.1080/17521440.2019.1602694

[77] ShannTurnbull, (2019) 'How shareholders, corporations and directors can become ethical', The European Financial Review, September 1, pp. 28-32, https://www.europeanfinancialreview.com/how-shareholders-corporations-and-directors-can-become-ethical/.

[78] Paul Berstein, (1980) Workplace democratization: Its internal dynamics, Transaction Books: New Brunswick, New Jersey.

[79] Beth Givens, (1991) Citizen utility boards: Because utilities bear watching, Centre for public interest law, University of San Diego, School of law, California.

[80] Torsten, T. Persson, Gerard Roland & Guido Tabellini, (1996) Separation of powers and accountability: Towards a formal approach to comparative politics. Innocenzo Gasparini Institute for Economic Research (IGIER), Working Paper, No. 100, July, Milan.

[81] Shann Turnbull, (2002) 'Watchdog Boards: Past, Present and Future?' Working Paper, February, http://papers.ssrn.com/abstract_id=608244

[82] Lyn, L. Dallas (1988) 'Two models of corporate governance: Beyond Berle & Means', Journal of Law Reform, University of Michigan, Vol. 22, No. 1, pp. 19–116.

[83] Lyn, L. Dallas (1997) 'Proposals for reform of corporate boards of directors: The dual board and board ombudsperson', Washington and Lee Law Review, Winter Vol. 54, No. 1, pp. 92–146.

[84] Daniel Diermeier, & Roger B. Myerson, (1999) 'Bicameralism and its consequences for the internal organisation of legislatures., The American Economic Review, Vol. 89, No. 5, pp. 1182–1196, https://www.aeaweb.org/articles?id=10.1257/aer.89.5.1182

[85] Andrew Murray, (1998) Minority report on the company reform review Bill 1997, Parliamentary Joint Committee on Corporations and Securities, March, The Parliament of the Commonwealth of Australia, https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Corporations_and_Financial_Services/Completed_inquiries/1996-99/companylaw/report/d01

[86] BTR (2019) 'Business Round Table redefines the purpose of a corporation to promote an economy that serves all Americans', August 9th, https://www.businessroundtable.org/business-roundtable-redefines-the-purpose-of-a-corporation-to-promote-an-economy-that-serves-all-americans

[87] Katharina Pistor, (2015) 'Module corporate law', Swiss International Law School, https://www.swissintlawschool.org/sils-ll-m/modules/module-corporate-law/

[88] Shann Turnbull, (2015) Introduction to the Practices, Science, and Art of Designing Corporate Charters, https://vimeo.com/137118382/1d7e82ce27

[89] Lawrence Fink, (2018) 'A sense of purpose', BlackRock letter to CEO's, https://www.blackrock.com/corporate/investor-relations/2018-larry-fink-ceo-letter

[90] Katharina Pistor, (2019) 'Why America's CEOs have turned against shareholders', Project syndicate, 26 August, https://www.project-syndicate.org/commentary/american-ceos-turn-against-shareholder-primacy-by-katharina-pistor-2019-08?barrier=accesspaylog

[91] Shann Turnbull, (2019) 'How shareholders, corporations and directors can become ethical', The European Financial Review, September 1, pp. 28-32, https://www.europeanfinancialreview.com/how-shareholders-corporations-and-directors-can-become-ethical/

[92] Shann Turnbull, (1975) Democratising the wealth of nations, Company Directors Association of Australia, Sydney, https://papers.ssrn.com/abstract_id=1146062

[93] Shann Turnbull, (1998) 'Should ownership last forever?' Journal of Socio-Economics, Vol. 27, No 3, pp. 341-363, http://papers.ssrn.com/paper.taf?abstract_id=137382 http://papers.ssrn.com/paper.taf?abstract_id=132108

[94] Shann Turnbull, (2015) 'Winning government with policies for reducing inequality?' Evatt Foundation 25 March http://www.evatt.org.au/news/winning-government-reducing-inequality.html. Republished 2016, Australian Fabian Society Newsletter, 30 April, http://d3n8a8pro7vhmx.cloudfront.net/australianfabians/mailings/322/attachments/original/WGBRIWLT.pdf?1461754892 and in 2020, June, Search Foundation, https://www.search.org.au/wining_government_by_reducing_inequality_with_less_taxes

[95] Shann Turnbull, (197) 'Time Limited Corporations', Abacus: A Journal of Business and Accounting Studies, Sydney University Press, Vol. 9, No. 1. pp. 28-43, June.

[96] Shann Turnbull, (1990) 'Re-inventing corporations', Journal of employee ownership, law and finance, Vol. 2, no. 4, pp. 109-136. (Re-published in Czech as Podniková Organizace, 1991).

[97] Shann Turnbull, (1991) 'Property Rights and Markets', in: Economic Alternatives for Eastern Europe Briefing No. 6, New Economics Foundation, London.

[98] Jiang Yiewei (1984) The self-determination system and current economic difficulties of Yugoslavia, Industrial Economic Management No. 3.

[99] Jiang Yiwei (1988) From enterprise-based economy to economic democracy, (Trans. Li Zhenguo etc.) Beijing.

[100]Thomas Picketty, (2014) Capital in the Twenty First Century, Arthur Goldhammer (translator), The Belknap Press of Harvard University Press,

[101]Shann Turnbull (2009) 'Affordable housing policy: Not identifiable with orthodox economic analysis', The Icfai University Journal of Urban Policy, Vol. 4, No. 1, pp. 21-43 http://papers.ssrn.com/abstract_id=1027864

[102] Shann Turnbull, 2007, 'A framework for designing sustainable urban communities', Kybernetes: The international journal of systems & cybernetics, Vol. 36, Nos. 9-10, pp. pp. 1543–1557, October 23, http://papers.ssrn.com/abstract_id=960193

[103]Shann Turnbull, (2017) 'Democratising the wealth of cities: Self-financing urban development', Environment and Urbanisation, Vol. 29, No. 1, pp. 237-250, April, http://journals.sagepub.com/doi/full/10.1177/0956247816685985

[104]NGCA New Garden Cities Alliance 2021, https://gardencities.info/reference-documents/01-principles/community-land-bank-clb/community-land-bank-bibliography/

[105]Shann Turnbull, (1974) 'Multinationals: Fading out with a Profit', Development Forum, United Nations, Geneva, p.3, June.

[106]Shann Turnbull, S. (1975) 'Fading Out with a Profit-Planned Corporate Obsolescence', The Canadian Forum, Vol. 55, No. 651, June, pp.14-16.

[107]Shann Turnbull, 2020, Do we need “A new model of corporate governance?” Working paper, https://papers.ssrn.com/abstract_id=3735205

[108]Lord Stern, (2006) The Economics of Climate Change: The Stern Review, Cabinet Office, HM Treasury, London, http://www.sternreview.org.uk

[109]Shann Turnbull, (1983b) ‘Selecting a local currency’, Options, June, Australian Adam Smith Club, Sydney. (Republished 1997 as ‘Creating a community currency’. In: Morehouse (ed.) pp. 167-177), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1128862

[110]Shann Turnbull, (2011) ‘Options for Reforming the Financial System’, in: The IUP Journal of Governance and Public Policy, Vol. 6, No. 3, pp. 7-34, September, http://ssrn.com/abstract=1322210

[111]Shann Turnbull, (2016) ‘Terminating currency options for distressed economies’, Athens Journal of Social Science,Vol. 3, No. 3, p. 205 http://www.athensjournals.gr/social/2016-3-3-3-Turnbull.pdf

[112]Shann Turnbull, (2018) ‘Sustainable Value Money: Why it is needed, how to get it?’ In: Sabri Boubaker, & Duc, K. Nguyen, (eds.), Corporate Social Responsibility, Ethics and Sustainable Prosperity, pp. 413-443, World Scientific Publishing, Singapore, https://ssrn.com/abstract=3022277.

[113]Shann Turnbull, S. (2020) ‘Reforming money to rescue economies and the planet’, 8 May, Online Opinion, https://www.onlineopinion.com.au/view.asp?article=20885

[114]Jim Cairns, (1976) ‘Review of Shann Turnbull’s Book - Democratising the wealth of nations’, JASSA, The Journal of the Securities Institute of Australia, No. 1, pp. 9–13, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3346835

AUTHOR

Dr Shann Turnbull is the Principal of the International Institute for Self-governance. He qualified as an Electrical Engineer in Tasmania, and has a BSc from Melbourne University, and a MBA from Harvard. His PhD from Macquarie University Sydney established the science of governance in any specie. He has worked as a business entrepreneur founding a number of enterprises including two public mutual funds and three firms that became publicly traded. In 1975 he co-authored the world’s first educational qualification for company directors and wrote “Democratising the Wealth of Nations”. In 1977 he was commissioned by the Australian Government to undertake the first economic analysis of Aboriginals, and in 1991 he advised on employee privatisation in the Czech Republic, Slovakia and China. He is prolific author on reforming the theories and practices of capitalism by adopting the practices of nature to achieve sustainability.

OPTIMIZATION OF RANDOM FOREST MODEL FOR ASSESSING AND PREDICTING GEOLOGICAL HAZARDS SUSCEPTIBILITY IN LINGYUN COUNTY

Chunfang Kong[1,2,3,4], Kai Xu[1,2,3,*], Junzuo Wang[1], Yiping Tian[1,3,4],
Zhiting Zhang[1,3,4] and Zhengping Weng[1,3,4]

[1]School of Computer, China University of Geosciences, Wuhan, China
[2]Hubei Key Laboratory of Intelligent Geo-Information Processing,
Wuhan, China
[3]Innovation Center of Mineral Resources Exploration Engineering Technology
in Bedrock Area, Ministry of Natural Resources, Guiyang, China
[4]National-Local Joint Engineering Laboratory on Digital Preservation and
Innovative Technologies for the Culture of Traditional Villages
and Towns, Hengyang, China

ABSTRACT

The random forest (RF) model is improved by the optimization of unbalanced geological hazards dataset, differentiation of continuous geological hazards evaluation factors, sample similarity calculation, and iterative method for finding optimal random characteristics by calculating out-of-bagger errors. The geological hazards susceptibility evaluation model based on optimized RF (OPRF) was established and used to assess the susceptibility for Lingyun County. Then, ROC curve and field investigation were performed to verify the efficiency for different geological hazards susceptibility assessment models. The AUC values for five models were estimated as 0.766, 0.814, 0.842, 0.846 and 0.934, respectively, which indicated that the prediction accuracy of the OPRF model can be as high as 93.4%. This result demonstrated that the geological hazards susceptibility assessment model based on OPRF has the highest prediction accuracy. Furthermore, the OPRF model could be extended to other regions with similar geological environment backgrounds for geological hazards susceptibility assessment and prediction.

KEYWORDS

Geological Hazards, Susceptibility Evaluation, Random Forest (RF), Optimized RF (OPRF), Geographical Information Systems (GIS).

1. INTRODUCTION

The geological hazards system is a nonlinear, dynamic and open complex giant system with multiple levels of structure, multiple control parameters, multiple time scales, and diverse processes [1]. Geological hazards is one of the most serious disasters that can cause not only great economic losses and ecological damage, but can also critically threaten the survival of human beings and the construction of major projects [2-4]. Therefore, the selection of a suitable geological hazards susceptibility assessment method is an important part of geological hazards research, which is of great significance to disaster reduction and prevention [4-6].

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 29-49, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110403

To date, various models and methods have been developed and applied for assessing geological hazards susceptibility in many areas of the world. Among them, the qualitative evaluation method based on expert experience is one of the commonly used methods in the early years. Such as fuzzy comprehensive evaluation model [7-9], analytical hierarchy process [2,3,7,9-12], and weighted linear combination [12], and so on. These methods determine the weight of each evaluation factor through expert scoring, being less time-consuming; However dependence on the subject experience and analysis judgment of the individual experts leads to lack of consistency and portability.

Deterministic model is another commonly used method to evaluate the susceptibility of geological hazards, such as the limit equilibrium method. The mode has high reliability based on the mechanical models of the relationship between relating factors and geological hazards [13], but it requires absolute detailed parameters such as physical, geological environment, tectonic lithology, hydrology and so on. Therefore, the availability of data limits the applicability of this kind of model to the evaluation of local-scale geological hazards.

In addition, quantitative evaluation is the most widely used method in evaluating the susceptibility of geological hazards, such as information value model [2,11,14-16], mathematical statistics method [2,6,10,14,16], certainty factor [17,18], logistic regression (LR) [19,20], artificial neural network (ANN) [3,4,6,21-23], decision tree (DT) [19,24,25], support vector machines (SVM) [6,15,19,26-30], and so on. These methods mainly use mathematical model to establish the quantitative relationship between geological hazards and evaluation factors, which can quantitatively describe the sensitivity of each evaluation factor in different intervals. Meanwhile, the data availability is high and the prediction accuracy is good. However, it is difficult to determine the relationship between each factor and geological hazard point in high dimensional space, and it is easy to overfit. It’s also hard for these methods to effectively deal with the multi-source, multi-class, multi-quantity, multi-modal, and multi-temporal geological hazard data accumulated in the long-term geological survey.

Recently, ensemble learning improves the accuracy and generalization ability of the model by integrating multiple weak classifiers into a single strong classifier. As a typical and representative ensemble learning method, random forests (RF) exhibits robust performance in data classification and pattern recognition problems [31]. At the same time, this method does not require the background knowledge of the sample and does not need to choose variables, omitting the tedious work of data pre-processing. Also, it integrates multiple decision trees by random sampling and predicts by majority voting mechanism. Compared with traditional machine learning methods such as ANN and SVM, RF has the advantages of fast classification speed, strong noise resistance and high prediction accuracy. Moreover, the introduction of randomness makes the model not easy to overfit. However, the voting selection mechanism in the RF model will lead to some decision trees with low training accuracy have the same voting ability, reducing the voting accuracy. Moreover, the number of decision trees and other parameters in RF models may also greatly impact the final classification results of RF.

Therefore, the main objective of the current study is to establish a geological hazards susceptibility evaluation model based on optimized random forest (OPRF) with strong processing ability for high-dimensional and large data sets by combining multiple decision trees. For this purpose, the RF model is optimized by optimizing the non-equilibrium data sets, differentiating continuous attributes, and improving the similarity calculation. Next, the out-of-bag (OOB) error estimation is calculated iteratively to find the best random feature and number. After that, taking Lingyun County as the case study, geological hazards susceptibility is divided into four levels for Lingyun County by using the OPRF model. Finally, to evaluate the effectiveness of the proposed OPRF, field investigation and the area under characteristic (AUC) values of the receiver operating curves (ROC) were used for comparison to the traditional ML classifiers. The evaluation results can provide reference for geological hazard prediction and disaster prevention and mitigation, and also provide decision support for land use development and rational utilization of resources and environment in Lingyun County.

2. Study Area and Data Treatment

2.1. Study area

Lingyun County is located between longitude 106°23'E to 106°55'E and latitude 24°06'N to 25°37'N in the northwest part of Guangxi, with a total area of about 2048.40km2 and a total population of 193,600, as shown in Figure 1.

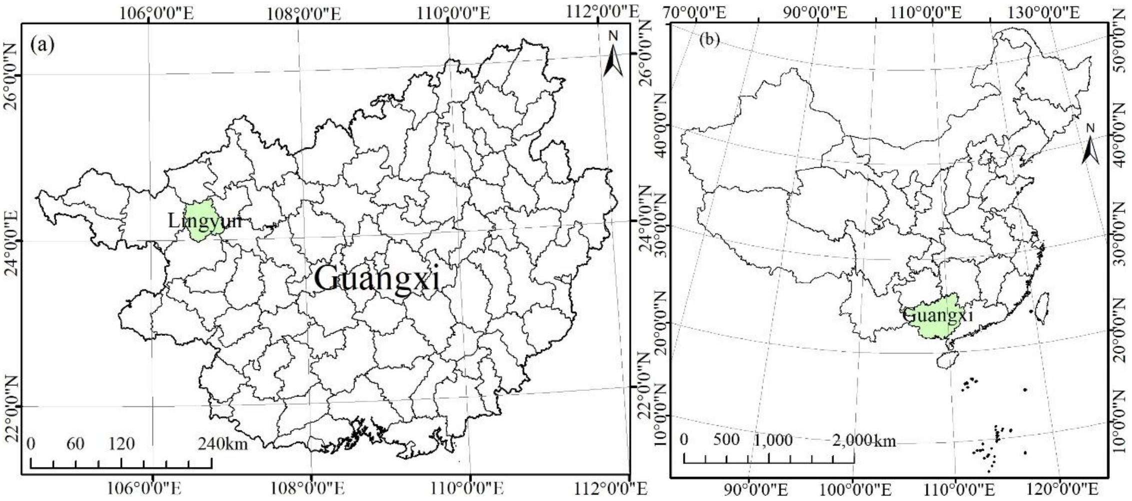


Figure 1. Location of Lingyun County in Guangxi Province (a) and China (b)

It is situated in the transitional zone of the Yungui plateau and the hilly mountainous area of Guangxi. The terrain in the northwest is high and low in the southeast, where in the west it is mostly a clastic rock geomorphology area, and in the east it is mainly a carbonate rock geomorphology. It belongs to a mountainous area with intervening deep valleys, with the mountain area accounting for 93.32% of the total area of the County. There are two main streams and 11 tributaries in the county, which belong to the Youjiang River and Hongshui River. Due to the strong influence of the southern subtropical monsoon and Karst landform, it is under the control of a tropical warm air mass for about half a year. Therefore, heavy rainfalls usually take place during the monsoon season (May to September); it has become one of the heaviest rains centers in Guangxi, and flood disasters occur from time to time in Lingyun County [32].

The intricate tectonic framework formed due to the occurrence of three obvious stages of tectonic evolution in Lingyun County, such as the Caledonian, Indosinian-Yanshan, and Himalayan periods. The exposed strata are mainly clastic rocks of Triassic and Cretaceous, carbonate rocks of Devonian, Carboniferous and Permian, accounting for 29.35% and 31.82% of the total area, respectively. In addition, there is also 16.30% clastic rock intercalated with siliceous rock, 11.35% sandstone, shale, conglomerate, 7.35% clastic rock intercalated with limestone. Late Cretaceous feldspar quartz porphyry veins with striped distribution, and the thin thickness of the Quaternary residual layer is distributed in structural erosion middle-low mountain areas, Karst depressions and valleys.

In general, it is a fragile geological environment zone and is prone to geological hazards in Lingyun County. According to inventory data from the Guangxi Geological Survey Bureau, there are 209 geological hazards in Lingyun County (Figure 2), including landslides, unstable slopes, collapses, dangerous rocks, and so on.

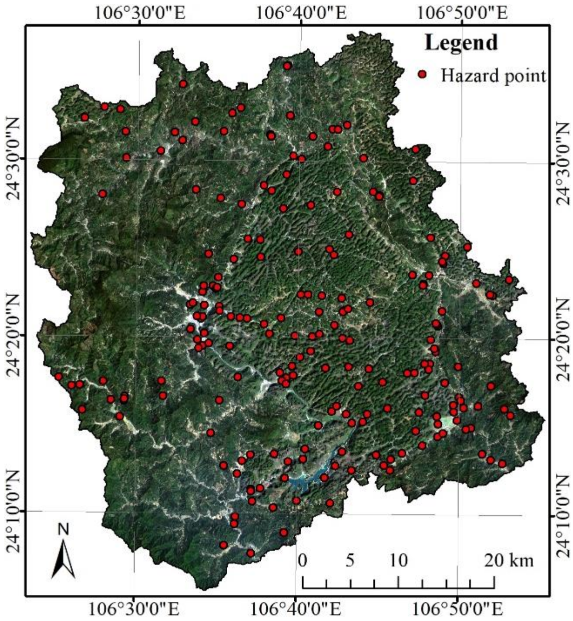


Figure 2. Image of Lingyun County and distribution of geological hazards

2.2. Data source

According to the characteristics of geological hazards and field investigation in Lingyun County, it is found that geological hazards susceptibility is closely related to the characteristics of natural geography, basic geology, ecological environment, human activities, and so on. In the current study, a total of ten geological hazards impacting elements were selected based on the field expedition of Guangxi Geological Survey Bureau as model input variables. They are slope, aspect, topographic curvature, normalized difference vegetation index (NDVI), annual precipitation, strata lithology, tectonic complexity, residential density, road network density, and land use and land cover (LULC). The data adopted in the current study are gathered mainly from the Guangxi Geological Survey Bureau and Guangxi Meteorological Bureau, as shown in Table 1.

Table 1. Data sources of geological hazards impacting elements.

No.	Factors	Data sources and scale
1	Slope	Digital elevation model (DEM) data of 90m
2	Aspect	
3	Topographic curvature	
4	NDVI	Landsat 8 OLI image
5	Annual precipitation	Meteorological data
6	Strata lithology	Geological map with scale 1:50,000
7	Tectonic complexity	
8	Residential density	Topographic map with scale 1:10,000
9	Road network density	
10	LULC	Landsat TM images

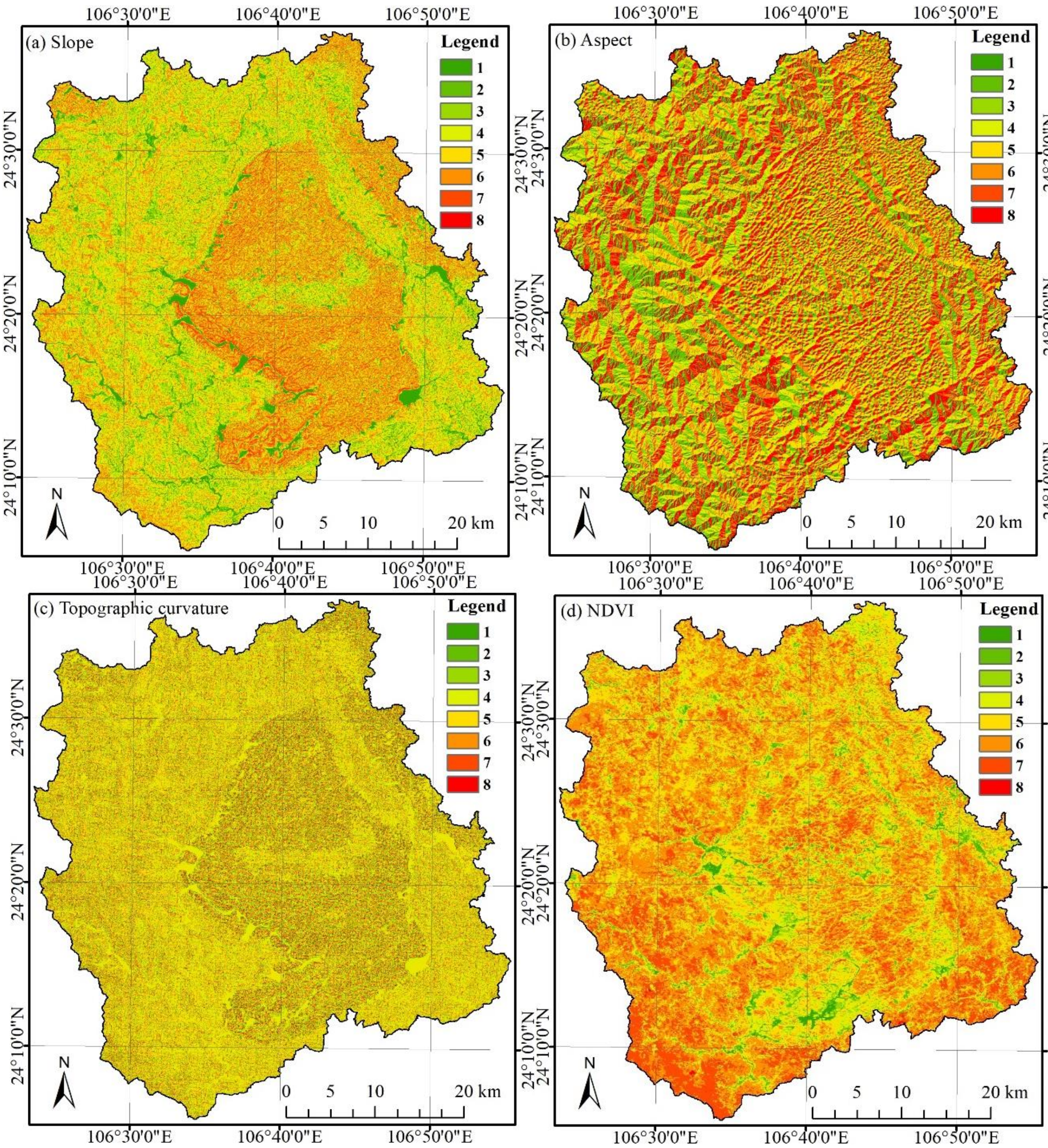
According to the size of geological hazards, this paper adopts a grid with a resolution of 30m×30m as the basic unit for the geological hazards susceptibility assessment, with a total of 2,275,996 evaluation units in Lingyun County.

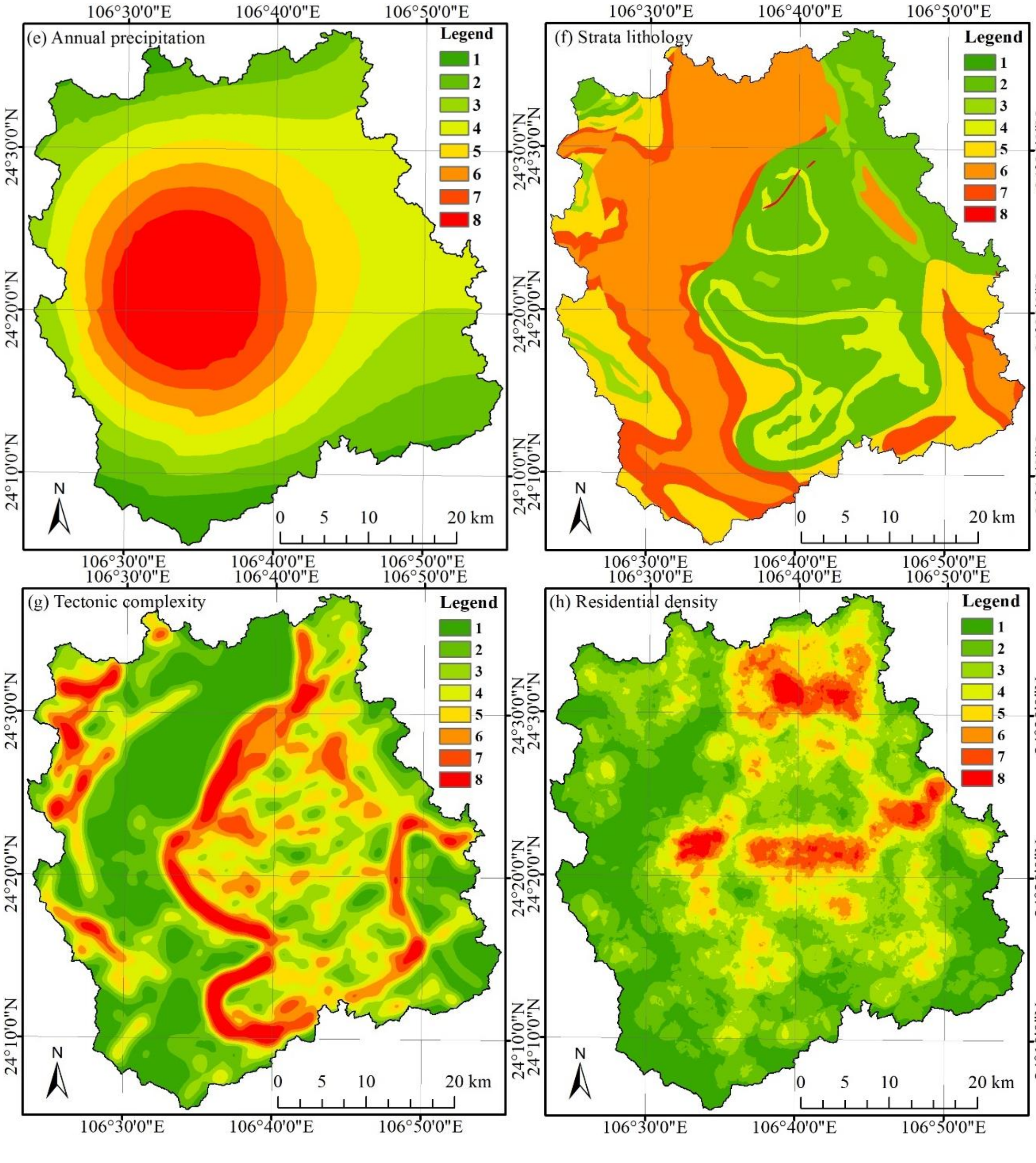
2.3. Treatment and analysis of geological hazards assessment factors

The classification of geological hazards impacting elements is closely related to the evaluation results of geological hazards susceptibility grade. In order to more objectively evaluate the susceptibility of geological hazards, the geological hazards impacting elements have been classified into different levels (Table 2) according to geological hazards characteristic and evaluation criterion developed by Guangxi Geological Survey Bureau for Lingyun County. At the same time, the geological hazards impacting elements of Lingyun County were differentiated, and the distinct effect is shown in Figure 3(a)-(j).

Table 2. Geological hazards impacting elements and their Classification.

No.	**Evaluation factor**	**Classification**
(a)	Slope (°)	1-[0,7); 2-[7,13); 3-[13,19); 4-[19,25); 5-[25,34); 6-[34,50); 7-[50,70); 8-[70,79)
(b)	Aspect (°)	1-[337.5,22.5); 2-[22.5,67.5); 3-[67.5,112.5); 4-[112.5,157.5); 5-[157.5,202.5); 6-[205.2,247.5); 7-[247.5,292.5); 8-[292.5,337.5)
(c)	Topographic curvature	1-[-25,-5); 2-[-5,-2.5); 3-[-2.5,-1); 4-[-1,0); 5-[0,1); 6-[1,2.5); 7-[2.5,5); 8-[5,25)
(d)	NDVI	1-[0,0.01); 2-[0.01,0.09); 3-[0.09,0.17); 4-[0.17,0.25); 5-[0.25,0.33); 6-[0.33,0.4); 7-[0.4,0.5); 8-[0.5,0.57)
(e)	Annual precipitation	1-[0,1930); 2-[1930,1990); 3-[1990,2050); 4-[2050,2110); 5-[2110,2170); 6-[2170,2230); 7-[2230,2290); 8-[2290,2350)
(f)	Strata lithology	1-Quaternary; 2-carbonate rock; 3-carbonatite with clastic rock; 4-clastic rock intercalated limestone; 5-clasolite intercalated with siliceous rocks; 6-clastic rock; 7-sandstone, shale, conglomerate; 8-granite or basal rocks
(g)	Tectonic complexity	1-[0,1.4); 2-[1.4,2.7); 3-[2.7,3.8); 4-[3.8,4.9); 5-[4.9,6); 6-[6,7.3); 7-[7.3,8.9); 8-[8.9,14.4)
(h)	Residential density	1-[0,1.2); 2-[1.2,2.4); 3-[2.4,3.5); 4-[3.5,4.5); 5-[4.5,5.8); 6-[5.8,7.1); 7-[7.1,8.6); 8-[8.6,12)
(i)	Road network density (km/km^2)	1-[0,3.2); 2-[3.2,4.7); 3-[4.7,6.1); 4-[6.1,7.8); 5-[7.8,9.7); 6-[9.7,11.7); 7-[11.7,13.9); 8-[13.9,15.3)
(j)	LULC	1-cultivated land; 2-woodland; 3-grassland; 4-river and lake; 5-construction land


(a) Slope
(b) Aspect
(c) Topographic curvature
(d) NDVI
Legend
1
2
3
4
5
6
7
8
106°30'0"E
106°40'0"E
106°50'0"E
24°30'0"N
24°20'0"N
24°10'0"N
N
0 5 10 20 km


(e) Annual precipitation
(f) Strata lithology
(g) Tectonic complexity
(h) Residential density
Legend
1
2
3
4
5
6
7
8
N
0 5 10 20 km
106°30'0"E
106°40'0"E
106°50'0"E
24°30'0"N
24°20'0"N
24°10'0"N

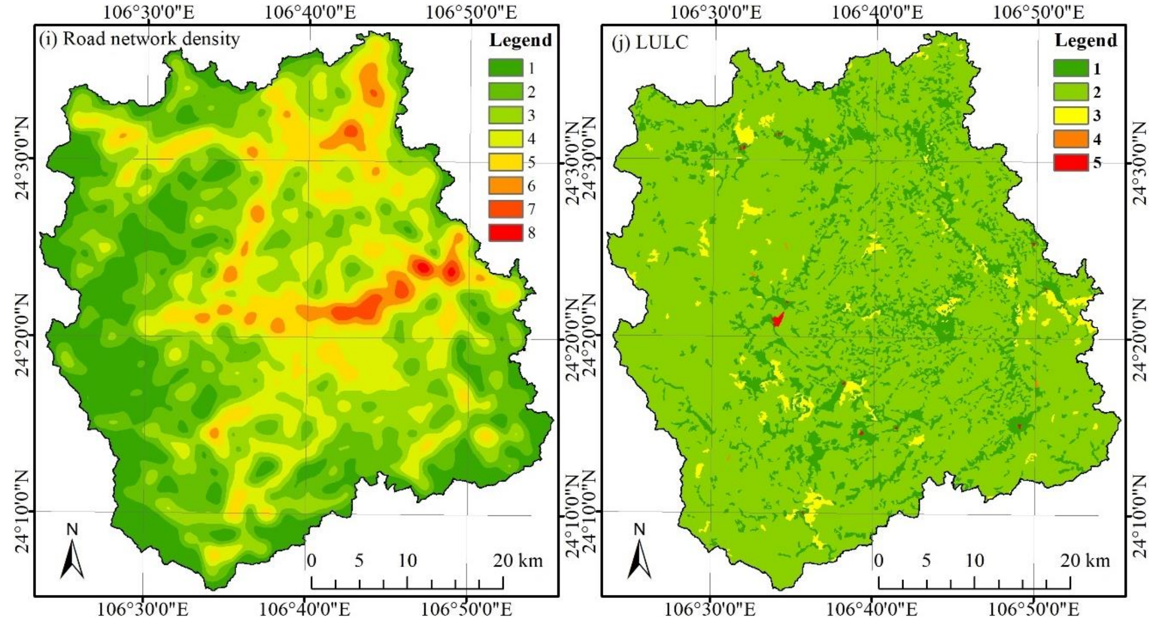


Figure 3. Attribute value of geological hazards evaluation factors [(a) slope, (b) Aspect, (c) Topographic curvature, (d) NDVI, (e) Annual precipitation, (f) Strata lithology, (g) Tectonic complexity, (h) Residential density, (i) Road network density, (j) LULC]

Meanwhile, the information values of each geological hazards impacting element was used to measure the impact of each element on geological hazards; the greater the information value, the greater the impact on geological hazards, which indicates that the higher the probability of occurrence of geological hazards in the region, the higher the susceptibility level [33,34]. The information values of each geological hazards impacting element in Lingyun County are shown in Figure 4.

Slope is an important indicator in the geological hazards survey process to measure the probability of movement of the slope deposits or Quaternary cover [21]. In the current study, the slope, aspect and topographic curvature was extracted from the digital elevation model (DEM) with 30m resolution by ArcGIS, as shown in Figures 3(a)-(c). At the same time, their information value is calculated, as shown in Figures 4(a)-(c). Figure 4(a) shows that the information value of the slope decreases first and then increases with the increase of the slope. This indicates that the impact of the slope with the occurrence of geological hazards also decreases first and then increases, and the impact of the slope with the occurrence of geological hazards is the most significant in the range of 50-70 degrees, followed by 35-50 degrees. Figure 4(b) shows that the information value of the aspect decreases first and then increases and then decreases and then increases with the increase of the aspect, which presents that the impact of the aspect on the occurrence of geological hazards is relatively complex, with the least impact in the range of 67.5-112.5, and the most significant in the range of 292.5-337.5. Figure 4(c) shows that the information value of topographic curvature decreases first and then increases with the increase of the topographic curvature, which states that the effect of the topographic curvature on the occurrence of geological hazards also decreases first and then increases, with the least effect in the range of -1 to 0, and the most significant in the range of 5 to 25.

The vegetation types are diverse, and the forest coverage rate is 71% in Lingyun County because it is a subtropical monsoon forest vegetation area, where the climate is mild, it is wet and rainy, and natural soil fertility is good. The NDVI of Lingyun County were extracted by Landsat 8 OLI image (2017/5/3, 127/043) and ArcGIS, as shown in Figure 3(d). At the same time, its information value is calculated, as shown in Figure 4(d). Figure 4(d) shows that the information values of NDVI decrease with the increase of NDVI, indicating that the effect of NDVI with the occurrence of geological hazards decrease with the increase of NDVI. That is to say, the better the vegetation cover, the less likely geological hazards will occur.

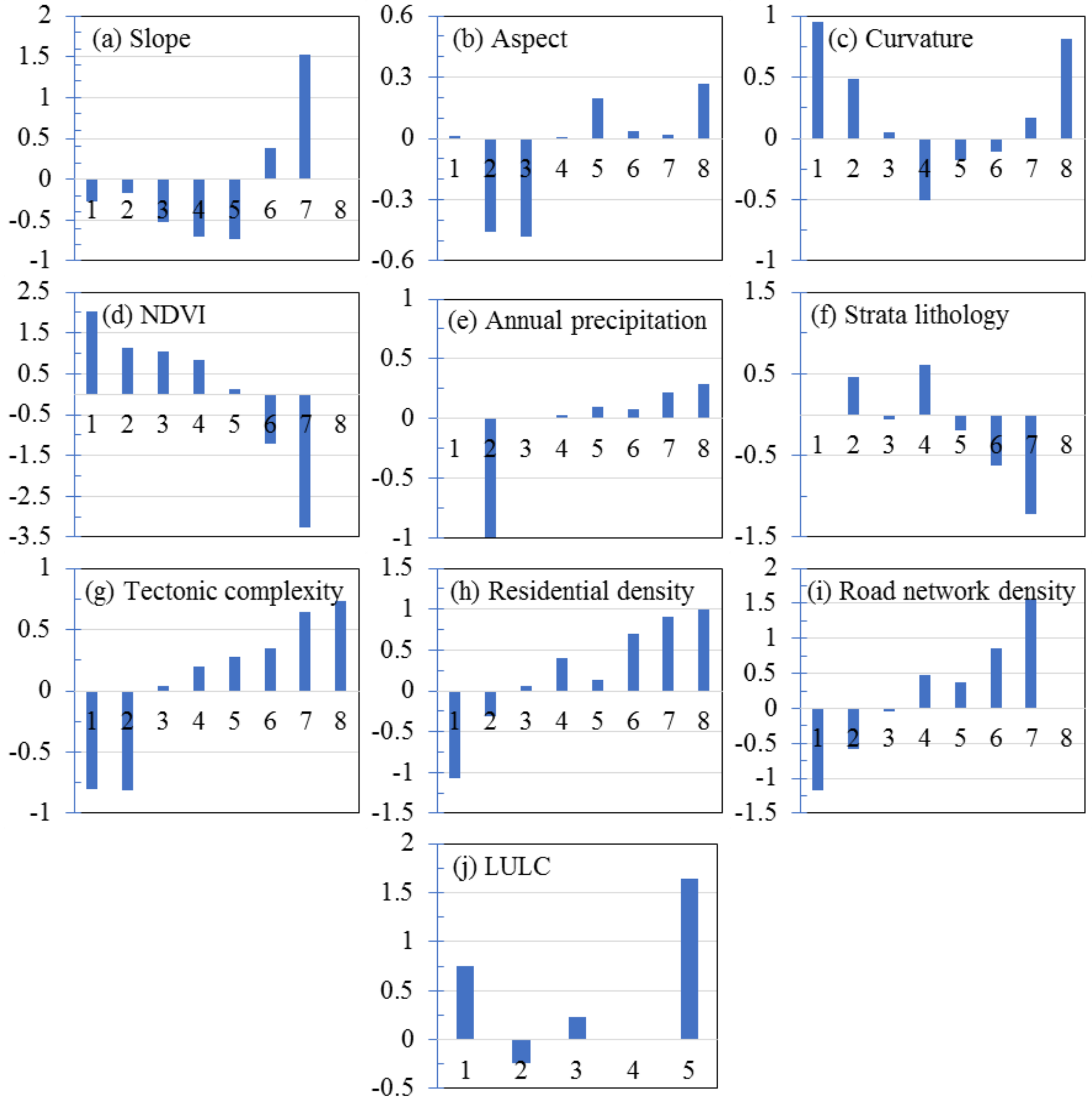


Figure 4. Information values distribution of main geological hazards impacting elements [(a) slope, (b) Aspect, (c) Topographic curvature, (d) NDVI, (e) Annual precipitation, (f) Strata lithology, (g) Tectonic complexity, (h) Residential density, (i) Road network density, (j) LULC]

Precipitation, especially heavy rain or continuous precipitation is the external dynamic factor that induces geological hazards [32]. There is plenty of precipitation in Lingyun County, and the average annual precipitation is 1235 mm. Under the action of precipitation infiltration, scour, and erosion, geological hazards occur from time to time. Meanwhile, the geological hazards and frequent periods of heavy rain are basically the same, indicating that the formation of geological hazards is closely related to heavy rain in Lingyun County [32]. Figure 3(e) is the annual precipitation map and Figure 4(e) is the information value of annual precipitation. Figure 4(e) indicates that the information value of precipitation increases with the increase of the precipitation, illustrating that the greater the precipitation, the greater the information value, and the greater the impact on the occurrence of geological hazards.

The strata exposed in Lingyun County are mainly carbonate rock and clastic rocks; also there is the clasolite intercalated with siliceous rocks, sandstone, shale, conglomerate, and clastic rock intercalated limestone, and so on. In the carbonate geomorphology area, rock joints and fissures developed, coupled with long-term weathering and dissolution, and rock collapse is easy to occur. In Karst depressions and valleys, it is easy to produce collapse under the action of groundwater [3], because shallow Karst develops and the thin Quaternary is overburdened. The landform of clastic rock is mainly composed of soft mud and shale, alternating with hard sandstone and siltstone. The mud shale is easy to weather and soften when it meets water, so it is easy to form a weak structural surface, resulting in geological hazards such as landslide, collapse and debris flow which are easy to occur. The strata and lithology of Lingyun County is exhibited in Figure 3(f), and the information value is expressed in Figure 4(f). Figure 4(f) indicates that the information value of clastic rock intercalated limestone is the largest, followed by carbonate rock, indicating that clastic rock intercalated limestone and carbonate rock are the most advantageous for the occurrence of geological hazards.

Fault is a zone with fragile structure and is prone to geological hazards [21]. Different periods and different forms of folds and faults with different properties have been formed after the occurrence of three strong crustal movements. At the same time, the later crustal rise suffered erosion and denudation, which caused some early-formed faults to reoccur, resulting in more complex geological structures in Lingyun County. Figure 3(g) states the tectonic complexity and Figure 4(g) states the information value of tectonic complexity in Lingyun County. Figure 4(g) also states that the information value of tectonic complexity increases with the increase of the tectonic complexity, illustrating that the greater the tectonic complexity, the greater the information value, and the greater the impact on the occurrence of geological hazards.

The geological engineering conditions of Lingyun are more complex because they are in the geological engineering environment composed of carbonate rocks, clastic rocks and loose accumulated rocks. As the scope of human activities continues to expand and strengthen, human activities such as steep slope cultivation and engineering construction strongly disturbed the topography and geomorphology in Lingyun County, which led to the occurrence of geological hazards, such as landslide, collapse, collapse, ground fissure, flood, water inrush, leakage, soil erosion, and so on. Residential density and road density of Lingyun were calculated, as exhibited in Figures 3(h)-(i). Meanwhile, their information values were also calculated, as exhibited in Figures 4(h)-(i). Figures 4(h)-(i) exhibit that the greater the density of settlements and roads, the greater the information value, the greater the impact on the occurrence of geological hazards.

In addition, there is 303.83km^2 of arable land and 1687.95km^2 of forest in Lingyun County. Figure 3(j) reveals the LULC map of Lingyun County, and its information value is exhibited in Figure 4(j). Figure 4(j) reveals that the information value of woodland is the smallest, while that of the construction land is the largest, indicating that woodland has the least influence on the occurrence of geological hazards, while construction land has the greatest influence on the occurrence of geological disasters; this further illustrates that the impact of human activities on the occurrence of geological disasters is relatively far-reaching.

2.4. Set up the geological hazards susceptibility assessment database

On the basis of the above, the database of the geological hazards susceptibility evaluation factors in Lingyun County was established, with a total of 2,275,996 grid evaluation units and 209 geological hazards points. Among them, 70% of the geological hazards points (146) were randomly selected as the geological hazards training samples, the rest 30% of the geological hazards points (63) were selected as the geological hazards testing samples. Accordingly, the non-hazards sample points of 10 times the number of geological hazards points (1460) were randomly selected as the geological hazards training samples, and 630 non-hazards sample points were selected as the geological hazards testing samples. The aim is to reduce the imbalance and spatial autocorrelation between the data of geological hazards points and non-hazards points.

3. Methods

3.1. RF model

RF is an ensemble learning method that generates a large number of independent training sets and multiple classification and regression trees (CART) by combining bagging [35,36]. The expression of the model is:

$$\{h(X,\theta_k), k = 1,2,3, \dots \} \quad (1)$$

where $h(X,\theta_k)$ is the classification regression tree without pruning generated by the CART algorithm; X is the input vector; $\{\theta_k\}$ is the random vector of independent distribution.

In geological hazards susceptibility assessment, first, 146 geological hazards and 1460 non-hazards sites samples were randomly selected by the bagging method as independent spatial training sets. Secondly, 10 geological hazards impacting elements were randomly selected for internal node branching without pruning to separately set up the CART tree for each training set [35, 36]. Thirdly, the other unselected 63 geological hazards and 630 non-hazards sites data as OOB data were to estimate the OOB error for each tree, and the OOB error for all trees is averaged to the RF [37]. Finally, the class with the most votes is taken as the geological hazards assessment result by synthesizing all decision trees. The specific implementation process is shown in Figure 5.

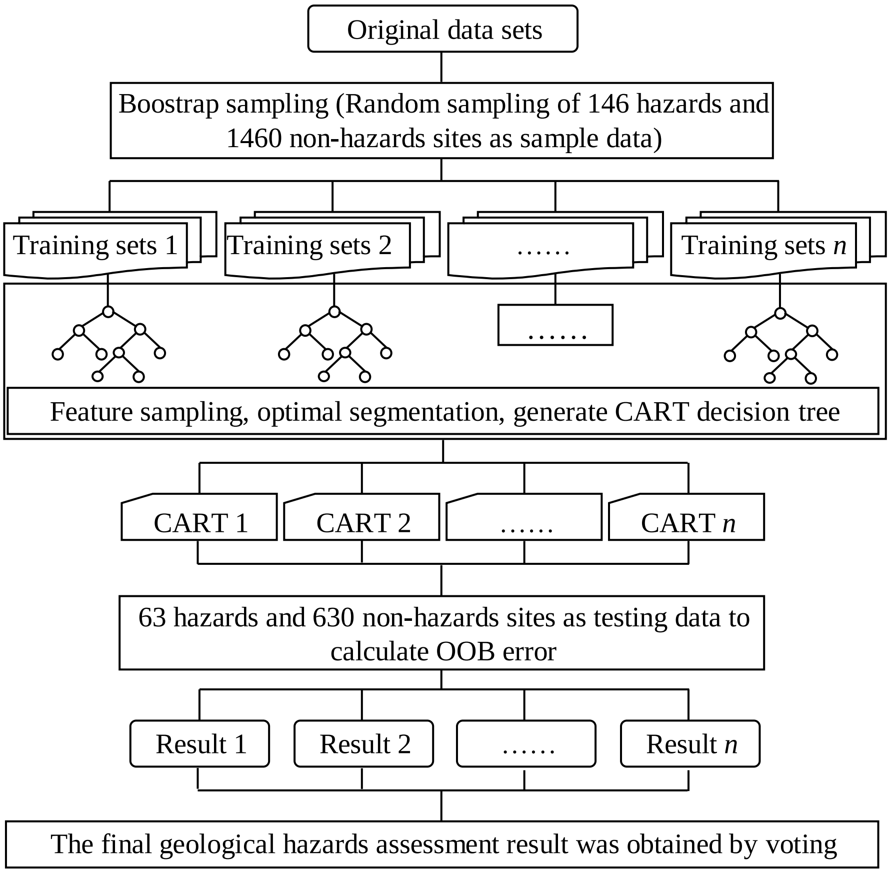

Original data sets
Boostrap sampling (Random sampling of 146 hazards and 1460 non-hazards sites as sample data)
Training sets 1
Training sets 2
……
Training sets n
……
Feature sampling, optimal segmentation, generate CART decision tree
CART 1
CART 2
……
CART n
63 hazards and 630 non-hazards sites as testing data to calculate OOB error
Result 1
Result 2
……
Result n
The final geological hazards assessment result was obtained by voting

Figure 5. Diagram of RF Algorithm

OOB error consists of unbiased estimates, approximated by cross-validation errors, and is bounded by generalization errors in RF [38]:

$$P^* \le \bar{\rho}\frac{(1-s^2)}{s^2} \quad (2)$$

where P* is the generalization error of the RF; $\bar{\rho}$ is the average of the correlation between CART trees; s is the average intensity of the decision tree.

Formula (2) illustrates that to enhance the generalization ability of RF, it can weaken the correlation between decision trees or increase the intensity of decision trees. For this purpose, this study introduces randomness to the feature selection of CART trees to weaken the correlation between decision trees.

The specific steps are as follows: (1) m features were randomly selected, (m≤10); (2) according to the principle of minimum non-purity of nodes, the optimal features are selected from these m characteristics to split the nodes; (3) the intensity and correlation of the CART tree are affected by m [38]. When m is too small, the intensity of the CART tree is weak; when m is too large, the intensity of CART tree increases, but the correlation between CART trees also increases.

In addition, this study further optimizes the RF model by optimizing the non-equilibrium data sets, differentiation of continuous attributes, and improving the similarity calculation of RF samples.

3.2. Optimization of non-equilibrium data sets

The sample data in the geological hazards susceptibility evaluation of Lingyun County are typically unbalanced data, because the number of geological hazards sites is far less than that of non-hazards sites, based on field investigations by the Guangxi Geological Survey Bureau.
In order to improve the evaluation accuracy, the C_SMOTE algorithm was applied to solve the non-equilibrium problem of the sample data. The steps are as follows:

(1) Calculate the central of the hazards sites, recorded as X_{center}. The formula is as follows:

$$X_{center} = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}, \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i2}, \dots\dots, \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ir} \right) \quad (3)$$

(2) Synthesis of "artificial" samples. The formula is as follows:

$$p_j = X_i + rand(0,1) \times (X_{center} - X_i) \quad (4)$$

where n_1 is the total sample number of the hazards sites; r is the attribute of each sample; $X_i\ (i = 1, 2, \dots\dots, n_1)$ is the hazards sites sample; X_{center} is the center of the hazards sites sample; $p_j\ (j = 1, 2, \dots\dots, m)$ is the synthetic "artificial" sample; and $rand(0,1)$ is a random number within the interval (0, 1).

(3) If the synthetic hazards sites sample number exceeds the actual required sample number, then use the under-sampling method to remove some samples far away from the center, finally, make the synthesized sample number reach the required equilibrium rate.The flow chart is shown in Figure 6:

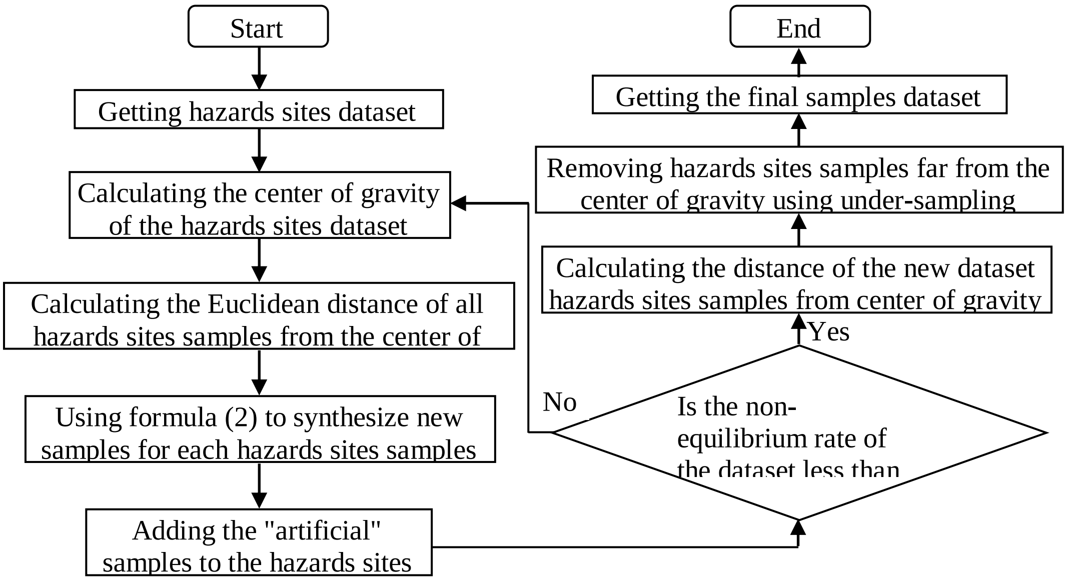

Start
Getting hazards sites dataset
Calculating the center of gravity of the hazards sites dataset
Calculating the Euclidean distance of all hazards sites samples from the center of
Using formula (2) to synthesize new samples for each hazards sites samples
Adding the "artificial" samples to the hazards sites
Is the non-equilibrium rate of the dataset less than
No
Yes
Calculating the distance of the new dataset hazards sites samples from center of gravity
Removing hazards sites samples far from the center of gravity using under-sampling
Getting the final samples dataset
End

Figure 6. Flow chart of the C_SMOTE algorithm

3.3. Differentiation of continuous attributes

There are 2 discontinuous attribute elements and 8 continuous attribute elements for geological hazards susceptibility evaluation in Lingyun County. To improve the accuracy of the RF model, this study adopts the entropy based on minimal description length principle (Ent-MDLP) to differentiate the attribute values of continuous evaluation factors. The steps are as follows:

(1) Dichotomy recursion to find breakpoints. First of all, find the adjacent points of different classes, and takes the midpoint between them as the candidate breakpoint T; secondly, each candidate breakpoint T can divide the sample set R into two subsets, calculate the information entropy of the two subsets respectively, then weight the summation to obtain the classification information entropy E(A,T,R); Finally, take the breakpoint T that makes the classification information entropy minimum as the final selected breakpoint.

(2) Determine the recursive downtime condition. The minimal description length principle (MDLP) is introduced here [38], and the downtime condition is that the information gain G should be satisfied:

$$\begin{aligned}\mathrm{G(A,T,R)} = \mathrm{E(R)} - \mathrm{E(A,T,R)} = \mathrm{E(R)} - \frac{|R_1|}{N \times E(R_1)} - \frac{|R_2|}{N \times E(R_2)} \\ > \frac{\log_2(N-1)}{N} + \log_2(3^k - 2) - [k \times E(R) - k_1 \times E(R_1) - k_2 \times E(R_2)]\end{aligned} \quad (5)$$

where A is an input variable, T is a breakpoint, R is a sample set, N is the total sample size, k is the number of categories; E(R) is the entropy of the sample set R; $E(R_1)$ and $E(R_2)$ are the entropy of the instance set R_1 and R_2 in each subinterval; and k_1 and k_2 are the number of categories in each subinterval.

3.4. Improving the similarity calculation of RF samples

It is an outstanding advantage of RF over other classifiers that RF can calculate the degree of similarity between samples and obtain the similarity matrix between samples. The similarity between the two samples can be measured by the frequency at which the two samples appear on the same leaf node on each tree, or by the probability that the two samples belong to the same class.

Assuming that the number of samples is N, the calculation process of the similarity matrix is as follows: First, the sample similarity matrix prox(i,j) is initialized as the all-zero matrix of N row N column. Then, all samples are discriminated with each tree generated, and each sample falls on one of the leaf nodes of the tree. Finally, for samples i and j, if they all land on the same leaf node of the tree, add 1 at row i and column j corresponding to the prox(i,j) matrix. Meanwhile, for the similarity between samples falling on different leaf nodes, the prox(i,j) matrix is improved by calculating the distance (d) between different leaves. The formula is as follows:

$$prox(i,j) = prox(i,j) + \frac{1}{d^m} \quad (6)$$

where d is the distance between the leaf nodes of sample i and j, and m is any positive real number.

The above procedure is repeated for each tree in the RF, traverse each tree to get a total addition value. Then each element is divided in the prox(i,j) matrix by the total number of trees to get the final prox(i,j) matrix. It is a symmetric matrix of N row N columns for prox(i,j) matrix, in which the diagonal elements are all 1, and the element of prox(i,j) in line i column j is defined as the similarity between sample i and sample j.

3.5. Set up the geological hazards susceptibility assessment model based on OPRF

In order to find the optimal random feature number, the OOB error of OPRF with a different random feature number was calculated by the cyclic iterative method, as shown in Figure 7.

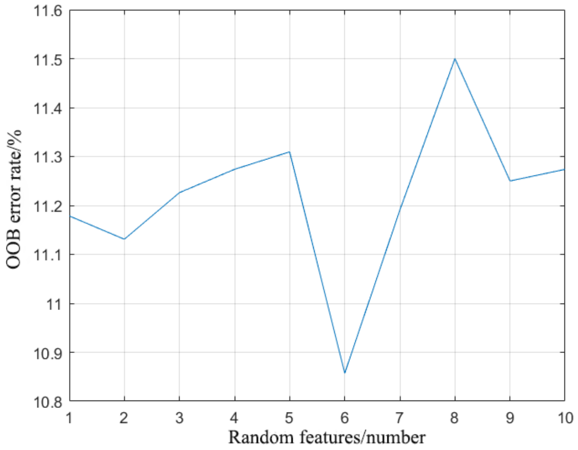


Figure 7. OOB Error distribution of OPRF with different numbers of random feature

Figure 7 indicates the variation characteristic of the OOB error with the increase of random feature number. When the number of random features is 6, the OOB error is the smallest, which indicates that the prediction accuracy of the geological hazards susceptibility evaluation model based on the OPRF established in the present study is the highest. Currently the number of decision trees in this OPRF is 81 and the maximum depth is 20.

4. Results and Discussions

4.1. Model evaluation metrics

Model precision and validation analysis is one of the essential steps for geological hazards susceptibility assessment and prediction [39]. Here, to test and verify the improvements and scientific significance of the proposed method in the current study, the proposed OPRF model was recommended and compared for comprehensive performance comparison with the RF and three other models, including LR, ANN, and SVM. The remaining 30% testing samples were used to test the five models, and the receiver operating characteristics (ROC) curves and the area under the curve (AUC) of each model prediction result were calculated (Figure 8), which is a widely used independent performance valuator [40-42]. The prediction performance is assessed by the AUC compared with the total plot area. If the AUC is equal to 1, it represents excellent prediction capability, while the AUC close to 0.5 represents a poor prediction capability [6, 8, 24, 30, 43, 44]. Figure 8 exhibits the ROC curves of the LR, ANN, SVM, RF and OPRF models in the current study.

Figure 8 shows that the AUC values of the LR, ANN, SVM, RF and OPRF models are 0.766, 0.814, 0.842, 0.846 and 0.934, respectively, which states that the prediction accuracy of five models for geological hazards susceptibility assessment in Lingyun County are 76.6%, 81.4%, 84.2%, 84.6%, and 93.4%, respectively. This result demonstrates that the geological hazards susceptibility assessment model based on OPRF has the highest prediction accuracy. which is mainly owing to the large number of elements selected in present study, the OPRF model, a type of ensemble learning, presented superiorities over a traditional method by not only accounting for different types of elements but also assessing the relative importance of the elements in terms of geological hazards stability [25]. At the same time, the result also demonstrates that the improvements proposed in the current study increase the performance of the RF model in evaluating and predicting the geological hazards susceptibility. Consequently, the OPRF model can be applied to the geological hazards susceptibility assessment under the same natural ecological environment.

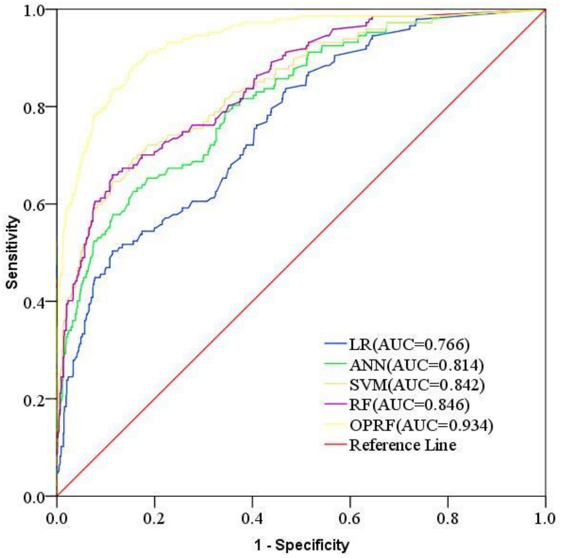


Figure 8. ROC curves and AUC values of test set for LR, ANN, SVM, RF and OPRF models

4.2. Evaluation results

The geological hazards susceptibility index of Lingyun County is calculated between 0 and 1, using the OPRF model, corresponding to the geological hazards susceptibility from low to high. At the same time, the Ent-MDLP method was used for the grading treatment, which was divided into four grades: [0-0.6776], (0.6776-0.7074], (0.7074-0.7372], and (0.7372-1], corresponding to the non-prone region, low-prone region, middle-prone region, and high-prone region, as shown in Figure 9.

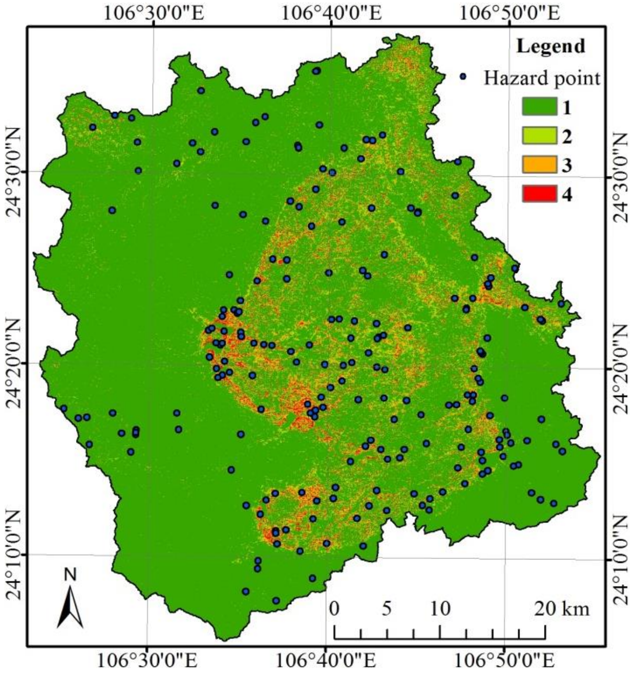


Figure 9. Evaluation results of geological hazards susceptibility in Lingyun County

Figure 9 shows that the high-prone region of geological hazards is 59.93 km^2, accounts for 2.93% of the total area in Lingyun County, mainly distributed in the regions of the carbonate rocks, where the slope is steep at 34-70 degrees, the aspect is between 292.5-337.5 and 157.5-202.5 degrees, the topographic curvature is between ±5-25°, vegetation coverage is low, the geological tectonic is complex, and the density of residents and road network is large. These regions are affected by multi-stage tectonic movement, which makes the joint fissure of rock mass develop, and the rock differentiation is strong, causes the frequent occurrence of disasters such as dangerous rocks, unstable slope, landslide, and collapse, indicating that carbonate rocks have a profound influence on the stability of geological hazards in the region. At the same time, there are many towns and traffic lines in these regions, indicating that these regions are strongly influenced by human activities.

The middle-prone region of geological hazards is 93.44 km^2, accounts for 4.56% of the total area in Lingyun County, mainly distributed in the regions of clastic rocks, clastic rock intercalated with limestone, and clastic rock intercalated with siliceous rock. Here the slope is from 7 to 34 degrees, vegetation coverage is low, and moderate density of population and road network. These regions have poor rock stability and strong weathering erosion, which provide a good material basis for the development of geological hazards.

The low-prone region of geological hazards is 139.31km^2, accounts for 6.8% of the total area in Lingyun County, mainly distributed near rural settlements where the rock mass is stable, the vegetation covers well, and is less disturbed by human activities.

The remaining region is the non-prone area of geological hazards, accounts for 85.71% of the total area in Lingyun County, where the rock mass is stable, the vegetation coverage is high, and is rarely affected by human activities to maintain its original natural ecological environment.

Figure 9 also indicates that the occurrence of geological hazards has a strong correlation with the vegetation index, road network density, and residential density, indicating the far-reaching impact of human activities on the occurrence of geological hazards in Lingyun County. It also indirectly illustrates that the construction of human engineering strongly interferes with the natural ecological environment of the region and leads to the frequent occurrence of geological hazards. Therefore, the research results of the current study also suggest that the stability and carrying capacity of the regional natural environment system should be fully considered in human engineering construction.

5. CONCLUSIONS

Geological hazards susceptibility evaluation is considered as an important task of geological hazards survey and is also the first important step in geological hazards risk assessments. Therefore, it is essential to accurately assess and predict geological hazards susceptibility regions with high performance-based models. Since performance of all kinds of proposed methods and techniques for simulating geological hazards is still being discussed, explorations of new methods for the evaluation of geological hazards are highly essential. These explorations will help obtain enough background knowledge to achieve some rational conclusions. The rapid development of advanced machine-learning allows for systems such as RF with high accuracy and better overall performance; use of these is recommended in disaster assessment and prediction. In this current study, the geological hazards susceptibility evaluation model based on OPRF was set up to assess and divide the hazards levels for Lingyun County. Meanwhile, field investigation and ROC curve were used to verify the evaluation results. The following conclusions have been reached in this study:

(1) The C_SMOTE algorithm is re-sampled on the line between the negative sample of the geological hazards point and the gravity center of the data set, so that the newly generated "artificial" sample is always between the center point and the negative sample of the geological hazards point; its position is determined by a random number, so it will not deviate from the geometric space of the negative sample set of the geological hazards point, and so it will not produce the tendency of marginalization, but will be directed towards the center point, thus reducing the randomness and blindness.

(2) The Ent-MDLP can better solve the differentiation problem when continuous geological hazards factors are increased and there is a lack of enough experience in the geological hazards susceptibility evaluation. At the same time, the discrete results show obvious trend characteristics and avoid the inconvenience of RF randomness to continuous factor analysis.

(3) When calculating the similarity between samples, for the similarity between samples falling at different leaf nodes, the loss of the sample similarity measure caused by “one-size-fits-all” is avoided by calculating the path distance d between different leaves to improve similarity matrix prox.

(4) The optimal random characteristic number is determined by finding the smallest OOB error of OPRF under different random characteristic numbers, which is calculated by iterative method.

(5) AUC values of the ROC curves and field investigation proved that the prediction accuracy of the geological hazards susceptibility evaluation model based on OPRF is higher than the original RF and the other three models.

In general, the improvements proposed in the current study aim to improve the accuracy and overall performance of the RF model for the geological hazards susceptibility evaluation. The RF model is improved in three aspects: optimization of unbalanced geological hazards data sets, differentiation of continuous geological hazards evaluation factor and the sample similarity calculation. On this basis, the geological hazards susceptibility evaluation model was set up based on OPRF. At the same time, the geological hazards susceptibility evaluation model was optimized by iteratively calculating the OOB error to find the best number of random features. Finally, geological hazards susceptibility is assessed by using the OPRF model, and the geological hazards susceptibility levels of Lingyun County are divided. Meanwhile, the accuracy and overall performance of evaluation results is verified by field investigation and ROC curves. The results indicate that the optimization strategies proposed in the current study are effective for the RF model. Furthermore, the OPRF can be expanded to the geological hazards susceptibility evaluation under the same natural ecological environment.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No.41201193), Guizhou Science and Technology Planning Project: Research and Application of Three-dimensional Prediction System for Gold Deposit in Southwest Guizhou Based on Geological Big Data ([2020]4Y039); Research and Development of Big Data Management and Intelligent Processing System for Manganese Ore Exploration ([2017]2951); Evaluation and Demonstration of Three-Dimensional Geological Survey of Deep Gold Mine in Guizhou Province (2020196700); Open fund project of National-Local Joint Engineering Laboratory on Digital Preservation and Innovative Technologies for the Culture of Traditional Villages and Towns (CTCZ19K01), and Open research project of key laboratory of Tectonics and Petroleum Resources (China University of Geosciences), Ministry of Education (No.TPR-2019-11). The authors would like to thank the anonymous reviewers for providing valuable comments on the manuscript.

REFERENCES

[1] Huang, R., Xu, X., Tang, C., & Xiang, X. (2008) Geological Environmental Assessment and Geological Hazard Management, Beijing, Science press.

[2] Sharma, S., & Mahajan, A. K., (2019) "A comparative assessment of information value, frequency ratio and analytical hierarchy process models for landslide susceptibility mapping of a Himalayan watershed, India", B. Eng. Geol. Environ, Vol.78, pp2431–2448.

[3] Sun, P., Cai, R., Xie, C., & Yi, Z. (2019) "Slope stability evaluation based on genetic optimization neural network", Mod. Electron. Tech, Vol.42, pp75–78.

[4] Wang, Y., Fang, Z., Wang, M., Peng, L., & Hong, H. (2020) "Comparative study of landslide susceptibility mapping with different recurrent neural networks", Comput. Geosci, Vol.138, pp104445.

[5] Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., & Al-Katheeri, M.M. (2016) "Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi Tayyah Basin, Asir region, Saudi Arabia", Landslides No.13, pp839–856.

[6] Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016) "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree", Landslides, No.13, pp361–378.

[7] Myronidis, D., Papageorgiou, C., & Theophanous, S. (2016) "Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP)", Nat. Hazards Vol.81, pp245–263.

[8] Ciurleo, M., Mandaglio, M. C., & Moraci, N. (2019) "Landslide susceptibility assessment by TRIGRS in a frequently affected shallow instability area", Landslides No.16, pp175–188.

[9] Sezer, E. A., Nefeslioglu, H. A., & Osna, T. (2017) "An expert-based landslide susceptibility mapping (LSM) module developed for Netcad Architect Software", Comput. Geosci, Vol.98, pp26–37.

[10] Bourenane, H., Guettouche, M. S., Bouhadad, Y., & Braham, M. (2016) "Landslide hazard mapping in the Constantine city, Northeast Algeria using frequency ratio, weighting factor, logistic regression, weights of evidence, and analytical hierarchy process methods", Arab. J. Geosci, No.9, pp1–24.

[11] Achour, Y., Boumezbeur, A., Hadji, R., Chouabbi, A., Cavaleiro, V., & Bendaoud, E.A. (2017) "Landslide susceptibility mapping using analytic hierarchy process and information value methods along a highway road section in Constantine, Algeria", Arab. J. Geosci, No.10, pp194–209.

[12] Hung, L.Q., Van, N.T.H., Duc, D.M., Ha, L.T.C., Son, P.V., Khanh, N.H., & Binh, L.T. (2016). "Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper lo river catchment (vietnam)", Landslides, Vol.13 No.5, pp1285-1301.

[13] Wang, X., Zhang, L., Wang, S., & Lari, S. (2014). "Regional landslide susceptibility zoning with considering the aggregation of landslide points and the weights of factors", Landslides, Vol.11, No.3, pp399-409.

[14] Liao, L., Zhu, Y., Zhao, Y., Wen, H., Yang. Y., Chen. L., Ma. S., & Xu. Y. (2019) "Landslide integrated characteristics and susceptibility assessment in Rongxian county of Guangxi, China", J. Mt. Sci, No16, pp657–676.

[15] Mokhtari, M., & Abedian, S. (2019) "Spatial prediction of landslide susceptibility in Taleghan basin, Iran", Stoch. Environ. Res. Risk Assess, Vol.33, pp1297–1325.

[16] Chen, W., Fan, L., Li, C., & Pham, B. T. (2020) "Spatial prediction of landslides using hybrid integration of artificial intelligence algorithms with frequency ratio and index of entropy in Nanzheng county, China", Appl. Sci, No10, pp29.

[17] Li, Y., Mei, H., Ren, X., Hu, X., & Li, M. (2018) "Geological disaster susceptibility evaluation based on certainty factor and support vector machine", J. Geo-info. Sci., Vol.20 No12, pp1699-1709.

[18] Zheng, Y., Chen, J., Wang, C., & Cheng, T. (2020) "Application of certainty factor and random forests model in landslide susceptibility evaluation in Mangshi City, Yunnan Province", B. Geol. Sci. Tech., Vol.39, No.6, pp131-144.

[19] Wang, F., Yin, K., Gui, L., & Chen L. (2018) "Landslide hazard analysis under different daily rainfall conditions in Wanzhou District", J. Geo-info. Sci., Vol.37 No.1, pp190-195.

[20] Hu, T., Fan, X., Wang, S., Guo, Z., Liu, A., & Huang, F. (2020) "Landslide susceptibility evaluation of Sinan County using logistics regression model and 3S technology", B. Geol. Sci. Tech., Vol.39, No2, pp113-121.

[21] Xu, K. Guo, Q., Li, Z., Xiao, J., Qin, Y., Chen, D., & Kong, C. (2015) "Landslide susceptibility evaluation based on BPNN and GIS: a case of Guojiaba in the Three Gorges Reservoir Area", Int. J. Geogr. Inf. Sci, Vol.29, pp1111–1124.

[22] Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., & Pourghasemi, H. R. (2018) "Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the Three Gorges Reservoir area, China", Comput. Geosci, Vol.112, pp23–37.

[23] Lee, D.H., Kim, Y.T., & Lee, S.R. (2020) "Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions", Remote Sens, No12, pp1194.

[24] Hong, H. Liu, J., Tien Bui, D., Pradhan, B., Acharya, T.D., Pham, B.T., Zhu, A.X., Chen, W., & Ahma, B.B. (2018) "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)", Catena Vol.163, pp399–413.

[25] Zhang, K., Wu, X., Niu, R., Yang, Y., & Zhao, L. (2017) "The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China", Environ. Earth Sci, Vol.76, pp405.

[26] Li, X., Cheng, X., & Chen, W. (2015) "Identification of forested landslides using LiDar data, object-based image analysis, and machine learning algorithms", Remote sens., Vol.7, No.8, pp9705-9726.

[27] Chen, Q., Liu, G., Ma, X., Zhang, J., & Zhang, X. (2019) "Conditional multiple-point geostatistical simulation for unevenly distributed sample data", Stoch. Env. Res. Risk A, Vol.33, pp973–987.

[28] Nguyen, H., Bui, X.N., Choi, Y., Lee, C.W., & Armaghani, D.J. (2020) "A novel combination of whale optimization algorithm and support vector machine with different kernel functions for prediction of blasting-induced fly-rock in quarry mines", Nat. Resour. Res, https://doi.org/10.1007/s11053-020-09710-7.

[29] Yu, X., & Gao, H. (2020) "A landslide susceptibility map based on spatial scale segmentation: A case study at Zigui-Badong in the Three Gorges Reservoir Area, China", PLOS ONE Vol.15, ppe0229818.

[30] Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2016) "A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India)", Environ. Modell. Softw, Vol.84, pp240–250.

[31] Chen, W., Li, X., Wang,Y., Chen, G., & Liu, S. (2014) "Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China", Remote Sen Environ., Vol.152, pp291-301.

[32] Zhang, L., Shi, S., & Liu, Q. (2016) "Spatial-temporal distribution characteristics and genetic analysis of geological disasters in Guangxi", Guangxi Water Resour. Hydropower. Eng, No.6, pp64–67.

[33] Murat, E., & Candan, G. (2004) "Use of fuzzy relations to produce landslide susceptibility map of landslide prone area (West Black Sea Region, Turkey)", Eng. Geol, Vol.75, pp229–250.

[34] Chen, L., Ye, J., Wei, C., & Xu, Y. (2016) "Application of ArcGIS and information method to landslide susceptibility evaluation", J. Guangxi Univ, Vol.41, pp141–148.

[35] Breiman, L. (1996) "Bagging predictors", Mach. Learn, Vol.24, pp123–140.

[36] Breiman, L. (2001) "Random forests", Mach. Learn, Vol.45, pp5–32.

[37] Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013) "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues", Nat. Hazards Earth Syst. Sci, Vol.13, pp2815–2831.

[38] Liu, J., Li, S., & Chen, T. (2018) "Landslide susceptibility assessment based on optimized random forest model", Geomat. Inf. Sci. Wuhan Univ, Vol.43, pp1085–1091.

[38] Frattini, P., Crosta, G., & Carrara, A. (2010) "Techniques for evaluating the performance of landslide susceptibility models", Eng. Geol, Vol.111, p 62–72.

[40] Hanley, J.A., & McNeil, B.J. (1983) "A method of comparing the areas under receiver operating characteristic curves derived from the same cases", Radiology, Vol.148, pp839–843.

[41] Swets, J. A. (1988) "Measuring the accuracy of diagnostic systems", Science, Vol.240, pp1285–1293.

[42] Fielding, A.H., & Bell, J.F. (1997) "A review of methods for the assessment of prediction errors in conservation presence/absence models", Environ. Conserv, Vol.24, pp38–49.

[43] Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C., & Peruccacci, S. (2010) "Optimal landslide susceptibility zonation based on multiple forecasts", Geomorphology, Vol.114, pp129–142.

[44] Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Tien Bui, D., Duan, Z., & Ma, J. (2017) "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", Catena, Vol.151, pp147–160.

AUTHORS

Chunfang Kong, Ph.D., Associate Professor, My current research interests in remote sensing of the resource environment, data mining and processing, machine learning, and GIS applications.

CONTEXT-AWARE SHORT-TERM INTEREST FIRST MODEL FOR SESSION-BASED RECOMMENDATION

Haomei Duan and Jinghua Zhu

School of Computer Science and Technology,
Heilongjiang University, Harbin, China

ABSTRACT

In the case that user profiles are not available, the recommendation based on anonymous session is particularly important, which aims to predict the items that the user may click at the next moment based on the user's access sequence over a while. In recent years, with the development of recurrent neural network, attention mechanism, and graph neural network, the performance of session-based recommendation has been greatly improved. However, the previous methods did not comprehensively consider the context dependencies and short-term interest first of the session. Therefore, we propose a context-aware short-term interest first model (CASIF).The aim of this paper is improve the accuracy of recommendations by combining context and short-term interest. In CASIF, we dynamically construct a graph structure for session sequences and capture rich context dependencies via graph neural network (GNN), latent feature vectors are captured as inputs of the next step. Then we build the short-term interest first module, which can to capture the user's general interest from the session in the context of long-term memory, at the same time get the user's current interest from the item of the last click. In the end, the short-term and long-term interest are combined as the final interest and multiplied by the candidate vector to obtain the recommendation probability. Finally, a large number of experiments on two real-world datasets demonstrate the effectiveness of our proposed method.

KEYWORDS

Recommendation, Session, Context, Neural Network, Interest, Graph

1. INTRODUCTION

In this era of the explosive growth of data, it is difficult for people to select the items they are interested in from a large number of items, and the recommendation system can recommend items that may be of interest to users. The recommendation system has been applied in many fields, such as e-commerce, music, and social media. It can be seen that the importance of the recommendation system. Most of the existing recommendation systems are based on the user's historical interactive information and make full use of the user's personal information. However, in some cases, users anonymously visit the site, thus the user's personal information is not available, besides, the user's historical interaction sequence is also very few. If the traditional method is used, it obviously is impossible to accurately recommend items for users. To solve this problem, session-based recommendation [1] is proposed to predict the next item that a user may click based on the sequence of the user's previous behaviors in the current session.

Due to the great practical value of session-based recommendation, it has been paid more and more attention, then various related recommendation algorithms are proposed. Markov Chain

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 51-65, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110404

(MC) is a classic case, which assumes that the next action is based on the previous ones [2]. The main problem with Markov methods, however, is that they assume too strongly the independence of more than two consecutive actions, so a great number of important sequential information cannot be well exploited for sessions with more than two actions. The proposal of recurrent neural network (RNN) has obtained promising results in the session-based recommendation system, which has been proved to be effective in capturing users' preference from a sequence of historical actions [3, 4, 5]. NARM [6] is designed to capture the user's sequential pattern and main purpose simultaneously by employing a global and local RNN. However, RNN-based methods hold that there is a strict sequential relationship between two adjacent items in a session, which restricts the extraction of the characteristics of session dynamic changes.

After the successful application of Transformer [7] in Natural Language Processing (NLP), many attention-based models have been designed, which has been shown that comparable performance with RNNs in many sequence processing tasks. For the session-based recommendation tasks, a short-term attention/memory priority (STAMP) model [8] has been proposed to learn users' current interest and general interest in a session. However, it only takes the future mean value of all items in the session as the context, without taking the dynamic variations and local dependencies of the sequence into account. A position-aware context attention (PACA) model[9] for session-based recommendation has was proposed in 2019, which takes into account both the context information and the position information of items. However, this method only trains an additional implicit vector for each item and has little performance improvement for session-based recommendation tasks.

To tackle the above problems, we propose a context-aware short-term interest first model (CASIF) for session-based reccommendation, which takes into account both the locally dependent context information, long-term interest, and short-term interest on the whole. Due to graph neural network (GNN) [10] is capable of providing rich local contextual information by encoding edge or node attribute features, we dynamically construct a graph structure for session sequences and capture context dependencies via GNN. Based on the session graph, the proposed CASIF can capture transitions of neighbor items and generate the latent factor vectors for all nodes included in the graph. Then we build the short-term interest first module, which can capture the user's general interest from the session in the context of long-term memory, at the same time get the user's current interest from the item of the last click. The implicit vectors representing general interest and short-term interest pass through the multilayer perceptron respectively. In the end, the short-term and long-term interest are combined as the final interest and multiplied by the candidate vector to obtain the recommendation probability.

The main contributions of this work are summarized as follows.

- To represent the session characteristics more accurately, we present a novel context-aware short-term interest first model(CASIF) for session-based recommendation. CASIF fully utilizes the complementarity between short-term interest first attention and graph neural network to enhance the recommendation performance.
- The module based on graph neural network is used to model local graph-structured dependencies of separated session sequences, while short-term interest first module is designed to capture contextualized global representations.
- We conduct extensive experiments on two baseline datasets. Our experimental results show the effectiveness and superiority of CASIF, comparing with the state-of-the-art methods via comprehensive analysis.

The rest of the article is structured as follows. We will state relevant work in Section 2. Detailed our proposed context-aware short-term interest first model in Section 3, Section 4 presents our detailed experimental results and analysis, and finally, we conclude this paper in Section 5.

2. RELATED WORK

Session-based recommendation tasks are performed based on the user's anonymous historical behavior sequence and implicit feedback data, such as clicks, browsing, purchasing, etc., rather than rating or comment data. The primary aim is to predict the next behavior based on a sequence of the historical sequence of the session. The related works of session-based recommendation are summarized as follows.

2.1. Conventional methods

The algorithms based on decision rules [11, 12] or to train the prediction model with shallow features [13] are the simplest methods. However, their recommendation performance is poor. Matrix factorization [14, 15]is a general method of recommending systems, which basic aim is to decompose the user-item rating matrix into two low-rank matrices, each of which represents the latent factors of the user or item. However, because session-based recommendation is a sequential recommendation problem without a user profile, these traditional underlying factor models may not well suited for session-based recommendations. Many sequential recommendation methods based on the Markov chain (MC) [16] model predict the next item based on the previous one through computing transition probabilities between two consecutive items. FPMC [2] models the sequence behavior between every two adjacent clicks and provides a more accurate prediction for each sequence by factoring the user's personalized probability transfer matrix. However, the main disadvantage of Markov chain-based models is that the assumption of independence is too strong, which limits the accuracy of recommendations.

2.2. Deep learning methods

The successful application of deep learning in other fields has led many people to introduce related methods into session-based recommendation tasks. The most typical example is that Hidasi et al. [3] introduction of recurrent neural network (RNN) into session-based recommendation for the first time, which is called GRU4REC, and achieve significant progress over conventional methods. Because of their excellent performance, many followers began to try this method. Such as, Tan et al. [4] further propose two techniques to improve the performance of session-based recommendation. Although these methods have improved the performance of session-based recommendation, they are all restricted by the constraints of RNNs that both the offline training and the online prediction process are time-consuming, due to its recursive nature which is hard to be parallelized [18]. Recently, attention mechanisms have shown significant improvement in many machine learning tasks, such as machine translation [7], knowledge graph [17]. For the session-based recommendation task, more and more methods are proposed to utilize the attention mechanism to improve performance. Li et al. [6] propose a hybrid RNN-based encoder with an attention layer, which is called neural attentive recommendation machine (NARM), employs the attention mechanism on RNN to capture users’ features of sequential behavior and main purposes. Then, a short-term attention priority model (STAMP) [8] using simple MLP networks and an attentive net, is proposed to efficiently capture both users’ general interest and current interest. Xu et al. [31] proposed Graph Contextualized Self Attention Network for Session-based Recommendation, which is a combination of GNN and attention mechanisms, and further improves the accuracy of the recommendation.

2.3. Neural network on graphs

In recent years, neural networks have been used to generate representations of graphically structured data, such as social networks and knowledge bases [19, 20]. Besides, classic neural networks CNN and RNN are also deployed on graph structure data [21]. Previously, in the form of recurrent neural networks, graph neural networks (CNN) [22] proposed to operate on digraph. Gated GNN [23] is a modification of GNN that USES gated recursive units and USES time backpropagation (BPTT) to calculate gradients. In recent years, GNN has been widely applied to different tasks, such as script event prediction [24], scene recognition [25], image classification [26]. Wu et al. [27] have applied graph neural network (GNN), to extract item embedding from a session graph. Furthermore, items' embeddings are inputted into an attentive network to generalize the final representation for item prediction. Wang et al. [30] proposed a novel Multirelational Graph Neural Network model for Session-based target behavior Prediction, which obtained excellent results by modeling multi-relational item graph.

In a word, these deep neural networks and graph neural network models have a strong ability to extract comprehensive features of depth, which greatly improves the performance of the recommendation.

3. THE CASIF MODEL

In this section, we introduce the proposed context-aware short-term interest first model for session-based recommendation (CASIF). We first formulate the problem of session-based recommendation and then describe the architecture of our model in detail (As shown in Figure 1.)

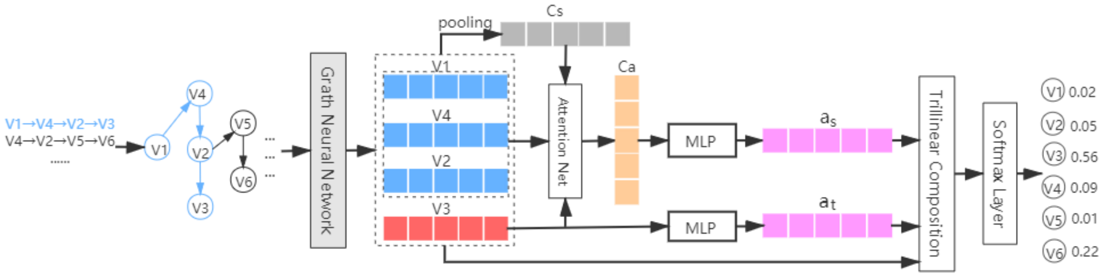


Figure 1. The general architecture of the proposed model.

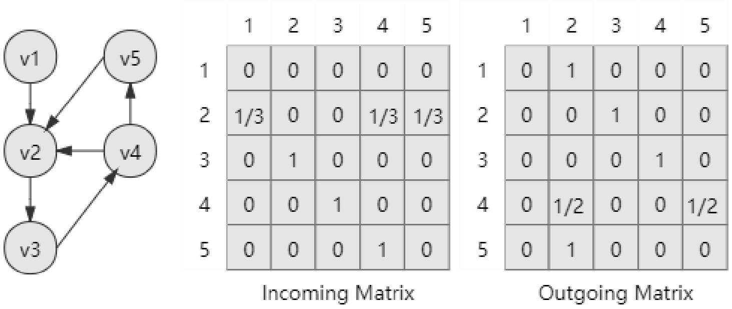
3.1. Problem formulation

Session-based recommendation aims to predict which item a user will click next, solely based on the user’s current sequential session data without access to the long-term preference profile. Here we give a formulation of this problem as below.

In session-based recommendation, let $V = \{v1, v2, \ldots, v_m\}$ denote the set consisting of all unique items involved in all sessions, where m represents the total number of items. An anonymous session sequence S can be represented by a list $S = \{s_1, s_2, \ldots, s_n\}$ ordered by timestamps, where $s_t \in V$ represents a clicked item of the user at time step$t$. Finally, the session-based recommendation aims to predict the next possible click (i.e., s_{t+1}) for a given prefix of the action sequence truncated at time t, $S = \{s_1, s_2, \ldots, s_{t-1}, s_t\}$. Specifically, our model returns a score list $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m\}$, where $\hat{y}$ represents the predicted scores respect to the item set v_i. Usually, top-k items will be chosen as recommendation items.

3.2. Structuring dynamic graph

The first part of the graph neural network module is to construct the meaningful graph from all the sessions. We embed every item $v \in V$ into a unified embedding space, which is represented as s. Given a session $S = \{s_1, s_2, \dots, s_n\}$, we treat each item s_i as a node and (s_{i-1}, s_i) as an edge which represents a user clicks item s_i after s_{i-1} in the session S. Therefore, every session sequence can be modeled as a directed graph. For example, considering a session $S = \{s_1, s_2, s_3, s_4, s_5\}$, the corresponding incoming matrice $\mathbf{M}^I \in \mathbb{R}^{5\times5}$ and outgoing matrice $\mathbf{M}^O \in \mathbb{R}^{5\times5}$ are shown in Figure2. Due to some items that may appear in the sequence repeatedly, we normalized weighted of all edges, which is calculated as the occurrence of the edge divided by the outdegree of that edge's start node. Note that our model can support various strategies of constructing a session graph and generate the corresponding connection matrices. Then we can apply the two weighted connection matrices with graph neural network to capture the local information of the session sequence.



	1	2	3	4	5
1	0	0	0	0	0
2	1/3	0	0	1/3	1/3
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0

Incoming Matrix

	1	2	3	4	5
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	0	1/2	0	0	1/2
5	0	1	0	0	0

Outgoing Matrix

Figure 2. An example of a session graph structure and the connection matrices $\mathbf{M}^I$ and $\mathbf{M}^O$

The node vector $\mathbf{h} \in \mathbb{R}^d$ indicates the latent vector of the item s learned via graph neural networks, where d is the dimensionality. The process of obtaining latent feature vectors of nodes as follows:

$$\mathbf{m}_i = Concat\begin{pmatrix} \mathbf{M}_i^I([s_1, \dots, s_n]\mathbf{W}^I + \mathbf{b}^I) + \mathbf{b}_i, \\ \mathbf{M}_i^O([s_1, \dots, s_n]\mathbf{W}^O + \mathbf{b}^O) + \mathbf{b}_o \end{pmatrix}, \qquad (1)$$

Where $\mathbf{M}_i^I \in \mathbb{R}^{1\times n}$ and $\mathbf{M}_i^O \in \mathbb{R}^{1\times n}$ represent the ith row of blocks in incoming matrices $\mathbf{M}^I \in \mathbb{R}^{n\times n}$ and outgoing matrices $\mathbf{M}^O \in \mathbb{R}^{n\times n}$ respectively corresponding to node s_i. $\mathbf{m}_i$ extracts the local contextual information of neighborhoods for node s_i.$\mathbf{W}^I$, $\mathbf{W}^O \in \mathbb{R}^{d\times d}$ are the parameter matrices. $\mathbf{b}^I$, $\mathbf{b}^O \in \mathbb{R}^d$ are the bias vectors. Eq.(1) is used for information propagation between different nodes. Next, $\mathbf{m}_i$ and s_i as input of graph neural network.

$$\mathbf{z}_i = \sigma(\mathbf{W}_z \mathbf{m}_i + \mathbf{P}_z s_i), \qquad (2)$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r \mathbf{m}_i + \mathbf{P}_r s_i), \qquad (3)$$

$$\tilde{\mathbf{h}}_i = \tanh\left(\mathbf{W}_o \mathbf{m}_i + \mathbf{P}_o(\mathbf{r}_i \odot s_i)\right), \qquad (4)$$

$$\mathbf{h}_i = (1 - \mathbf{z}_i) \odot s_i + \mathbf{z}_i \odot \tilde{\mathbf{h}}_i, \qquad (5)$$

Where $\mathbf{W}_z$, $\mathbf{W}_r$, $\mathbf{W}_o \in \mathbb{R}^{2d\times d}$, $\mathbf{P}_z$, $\mathbf{P}_r$, $\mathbf{P}_o \in \mathbb{R}^{d\times d}$,are learnable parameter matrices. $\sigma(\cdot)$ represents the logistic sigmoid function and $\odot$ is element-wise multiplication operator. $\mathbf{z}_i$ and $\mathbf{r}_i$ are the reset and update gates respectively, which determines whether some information is retained or forgotten. Finally, output $\mathbf{h}_i \in \mathbb{R}^d$ of the GNN layer is the latent future vector corresponding to s_i.

Note that for the construction of dynamic graph, in practice, the graph structure containing different information can be built according to the actual situation, such as the type of item, price and other ancillary information.

3.3. Short-term Interest First Module

After obtaining latent future vectors from the graph neural network, we input them into short-term interest first module. As can be seen in Figure 1. the pooling operation is used to generate the session feature m_s, which represents the main interest of the current session. Due to the mean average of all items feature in the whole session represents a central feature of the session, there we choose the mean pooling, the specific formula is as Eq.(6).

$$\mathbf{c}_s = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_i, \qquad (6)$$

After capturing the user's general interest m_s, we use an attention network layer to obtain the attention coefficient based on short-term interest first. Attention coefficient$s$ are computed as follows:

$$\alpha_i = \mathbf{W}_0\, \sigma(\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{h}_n + \mathbf{W}_3 \mathbf{c}_s + \mathbf{b}_a\ ) \qquad (7)$$

Where $\mathbf{h}_i \in \mathbb{R}^d$ denotes the ith latent vector corresponding to ith item s_i, $\mathbf{h}_n$ denotes the latent vector of the last click item s_n, $\mathbf{W}_0 \in \mathbb{R}^{1\times d}$ is a weighting vector, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3 \in \mathbb{R}^{d\times d}$ are weighting matrices, $\mathbf{b}_a \in \mathbb{R}^d$ is a bias vector, and $\sigma(\cdot)$ denotes the sigmoid function. α_i represents the attention coefficient of latent vector $\mathbf{h}_i$ within the current session prefix $S = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$. From Eq.(7) one can see that the attention coefficients are calculated based on the latent vector $\mathbf{h}_i$, the last-click $\mathbf{h}_n$ and session representation $\mathbf{c}_s$, thus, it can capture the correlations between the target item and the long/short term memory of the user's interest. Note that in Equation 7, the short-term interest is represented the last click item, which is explicitly considered, and this is why the proposed module is called the short-term interest first module.

So after the previous calculation, we can obtain the attention coefficients $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ with corresponding to the current session $S = \{\mathbf{h}_1,\ \mathbf{h}_2, \dots,\ \mathbf{h}_n\}$, the attention-based user's global interest $\mathbf{c}_a$ can be calculated as follows:

$$\mathbf{c}_a = \sum_{i=1}^{n} \alpha_i \mathbf{h}_i\ , \qquad (8)$$

3.4. MLP layer

Next, we choose the latent feature vector of the last click item $\mathbf{h}_n$ as current interest. The general interest $\mathbf{c}_a$ and current interest $\mathbf{h}_n$ are processed with two MLP networks for feature abstraction. The two MLP has the same structure except for different parameters. The specific definition is as follows:

$$a_s = \tanh(\boldsymbol{w}_s \boldsymbol{c}_a + \boldsymbol{b}_s)\ , \qquad (9)$$

$$a_t = \tanh(\boldsymbol{w}_t \boldsymbol{c}_a + \boldsymbol{b}_t)\ , \qquad (10)$$

Where $a_s, a_t \in \mathbb{R}^d$ denotes the final state of global interest and current interest respectively, $\boldsymbol{w}_s, \boldsymbol{w}_t \in \mathbb{R}^{d \times d}$ are learnable weighting matrix, and $\boldsymbol{b}_s, \boldsymbol{b}_t \in \mathbb{R}^d$are the bias vector. The tanh is non-linear activation function.

3.5. Model training and making recommendation

Finally, for a given candidate item $s_i \in V$, the score function is defined as:

$$\hat{z}_i = s_i^T (a_s \odot a_t), \qquad (11)$$

And then we apply a softmax function to get the output vector of the model $\hat{y}$:

$$\hat{\mathrm{y}} = softmax(\hat{\mathrm{z}}), \qquad (12)$$

Where $\hat{\mathrm{z}} \in \mathbb{R}^{|V|}$ denotes the recommendation scores with respect to the item set V, and $\hat{y}$ represents a probability distribution over the items $s_i \in V$, each element $\hat{y}_i \in \hat{y}$ denotes the probability of the event that item s_i is going to appear as the next-click in this session.

For each session prefix $S = \{s_1, s_2, s_3, s_4, s_5\}$, the loss function is defined as the cross-entropy of the prediction and the ground truth:

$$\mathcal{L}(\hat{\mathrm{y}}) = -\sum_{i=1}^{|V|} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i), \qquad (13)$$

Where y denotes a one-hot vector of the ground truth item. For example, if s_{n+1} denotes the ith element v_i in item dictionary V, then $y_k = 1$, if $i == k$, and $y_k = 0$.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Model training and making recommendation

4.1.1. Datasets

Our experiment constructs on two real-world datasets: Yoochoose from RecSys Challenge 2015 and Diginetica from CIKM Cup 2016. To facilitate comparison, we preprocess the two datasets in the same way as [6,8]. First, we process items in all sessions in chronological order. Second, we filter out those sessions that have only one item and that have items appearing less than 5 times. Third, we generate the sequences and the corresponding lables by splitting the input sequence. Specifically, we set the sessions for the next few days as test sets for Yoochoose and the sessions for the next few weeks as test sets for Diginetiva. we also use the most recent fractions 1/64 and 1/4 of the training sequences of Yoochoose. The statistics of datasets are presented in Table 1.

Table 1. Statistics of datasets used in the experiments

Statistics	Yoochoose1/64	Yoochoose1/4	Diginetica
all the clicks	557,248	8,326,407	982,961
train sessions	368,859	5,917,745	719,470
test sessions	55,898	55,898	60,858
all the items	16,766	29,618	43,097
average length	6.16	5.71	5.12

4.1.2. Baselines

We compare the proposed models with four traditional methods (POP, IKNN, BPR-MF, FPMC) and three recent deep learning models (GRU4REC, NARM, STAMP, SR-GNN).

- Pop is a simple baseline that recommends top rank items based on popularity in training data.
- Item-KNN [30] recommends the item similar to the items that have been clicked in the current session, where the cosine similarity is used.
- BPR-MF [28] is a learning-to-rank method. It is the most advanced non-sequential recommended method, which utilizes pair-sort loss optimization matrix decomposition.
- FPMC [2] combines the Markov chain model and matrix factorization for the next-basket recommendation.
- GRU4REC [4] is an RNN-based deep learning model for session-based recommendation. It utilizes a session-parallel mini-batch training process and applies a ranking-based loss function for training.
- NARM [6] adopts recurrent neural network as its basic component and utilizes an attention mechanism to extract users' main purpose.
- STAMP [8] is a novel short-term memory priority model. The attention mechanism is used to capture the user's general interest, and finally the click item represents the user's current interest.

- SR-GNN [27] models session sequences as graph structure data and uses graph neural networks to obtain item latent vectors, which are input to a traditional attentive neural network for learning session representation.

4.1.3. Performance metrics

We use the following performance metrics to compared these algorithms, which have been widely used in session-based recommendation systems.

$Recall@k$: Be widely used as a measure of predictive accuracy in all kinds of recommendation systems. It represents the proportion of correctly recommended items amongst the top-k items.

$$Recall@k = \frac{n_{hit}}{N}, \qquad (14)$$

Where N is the number of test sessions in the testing set, n_{hit} denotes the number of sessions which have hit items among top-K ranking list.

$MRR@k$: MRR (Mean Reciprocal Rank) is the average of reciprocal ranks of desired items. The reciprocal rank is set to zero if the rank is larger than K.

$$MRR@k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \qquad (15)$$

4.1.4. Parameter settings

Following previous methods [6, 8, 28] the dimension of embedding D is set to 100 for both datasets. All parameters are initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. The proposed CASIF model uses Adam optimizer to optimize these parameters and the batch size is set as 128. For the Yoochoose, the learning rate is set to 0.001. For the Diginetica, the learning rate is 0.003, and all learning rates will decay by 0.1 after every 3 epochs. Besides, the L2 penalty is 10e-5. Lastly, our model is implemented with Pytorch on GeForce GTX 1660Ti GPU and all the experimental results are the average value of five times testes.

4.2. Experiment results

4.2.1. Comparison with Baseline Methods

To state the performance of our CASIF model for session-based, all experimental results were evaluated by Recall@20 and MRR@20. As shown in Table 2, the best results have been highlighted in boldface. Results analysis is mainly divided into three parts, the first is the traditional method, the second is the deep learning related models, and the last is our model.

For traditional models, POP, as the simplest algorithm, has the worst recommendation performance. By analyzing the users individually and optimizing the paired ranking loss function, the performance of BPR-MF is better than that of POP. This suggests the importance of personalization in the recommendation task. Although FPMC integrates Markov chain and matrix decomposition, the overall result is not as good as Item-KNN. Please note that Item-KNN only uses the similarity between items, and does not consider the sequence information. This shows that the assumption of independence of continuous terms that the traditional MC method relies on is not realistic.

Obviously, the performance of all neural network models is better than that of traditional methods. This also verifies the powerful role of deep learning in this field, because they can extract some deep-seated and representative potential features from the temporal relationship of items in historical sessions. GRU4REC uses the recursive structure GRU as a special form of RNN to capture the general preferences of users. It improves the performance of the session-based recommendation greatly. But NARM and STAMP are better than it, which shows the effectiveness of attention mechanism and short-term behavior in predicting the next project problem.

Our model just uses a deep learning method and attention mechanism. Therefore, compared with the baseline model, our method achieve the best results. First, we use graph-structured data to input into the neural network, which captures the local dependence of the session. Secondly, in the attention layer, we first consider the short-term behavior and then integrate global behavior, which captures more accurate context information. Finally, more potential features are extracted by MLP. Our model is especially suitable for recommendation tasks with a large amount of data.

Table 2. Performance comparison for different methods over the three datasets.

Dataset	**Yooochoose 1/64**		**Yooochoose 1/4**		**Diginetica**	
Measure	**Recall@20**	**MRR@20**	**Recall@20**	**MRR@20**	**Recall@20**	**MRR@20**
POP	6.71	1.65	1.33	0.30	0.91	0.23
BPR-MF	31.31	12.08	3.40	1.57	15.19	8.63
IKNN	51.60	21.81	52.31	21.70	28.35	9.45
FPMC	45.62	15.01	--	--	31.55	8.92
GRU4REC	60.64	22.89	59.53	22.60	43.82	15.46
NARM	68.32	28.63	69.73	29.23	62.58	27.35
STAMP	68.74	29.67	70.44	30.00	62.03	27.38
SR-GNN	70.57	30.94	71.36	31.89	63.03	27.42
CASIF	**70.70**	**31.21**	**72.01**	**32.11**	**63.59**	**28.33**

4.2.2. Further comparison with excellent baselines

In order to further study the performance of our proposed CASIF and the state-of-the-art methods NARM, STAMP and SR-GNN in the real application environment, where only a few items can be recommended to users at once. Therefore, we evaluate our model in terms of Recall@10, MRR@10, Recall@5, and MRR@5 to further measure recommendation accuracy. The results on two datasets Yoochoose1/64 and Diginetica are summarized in table 3. As we can see, Our model is comparable to SR-GNN in terms of smaller data sets Yoochoose1/64 and Recall@5, which mainly because the data set is limited and the precision is not easy to improve on the strict performance metrics. However, overall, our model is comparable performs well on this case and shows obvious advantages, especially on MRR@10, which indicates that our model is more accurate in ranking candidate items and demonstrates the effectiveness of taking into account both the context information and shor-term intersts first for session-based recommendation.

Table 3. Recommendation performance of STAMP, SR-GNN and our CASIF on Recall@k, MRR@k, where k=10, 5

Dataset	Measure	NARM	STAMP	SR-GNN	CASIF	Improve
Yoochoose1/64	**Recall@10(%)**	57.50	58.07	60.01	61.21	+1.99%
	MRR@10(%)	27.97	28.92	29.56	30.58	+3.45%
	Recall@5(%)	44.34	45.69	47.11	46.88	-0.49%
	MRR@5(%)	26.21	27.26	28.18	28.29	+0.03%
Diginetica	**Recall@10(%)**	51.91	52.07	52.70	53.18	+0.91%
	MRR@10(%)	26.53	26.90	27.12	27.86	+2.72%
	Recall@5(%)	40.67	41.04	41.45	41.52	+0.16%
	MRR@5(%)	25.02	25.21	25.42	26.04	+2.43%

4.2.3. Effectiveness of short-term interest first module

To verify the effectiveness of short-term interest first module, We design a variant of our model, that is, after GNN, as shown in Eq. (16, 17, 18), we directly obtain the attention coefficient in a simple attention, and then sum weighted by the corresponding latent vectors. We mark this variant as CASIF-S. From Fig. 3. , we can see the experimental results, whether on the evaluation metrics Recall@20 or MRR@20, the performance of CASIF is higher than CASIF-S. the experience verifies the importance of the short-term interest first module in our proposed model, and it also demonstrates that it is not enough to only consider GNN, because GNN captures only the local dependencies of the session and does not obtain the main intent of the session from global. Especially, the pooling layer in our model can capture the main interst of session. And only consider attention, easy to deviate from the overall.

$$\alpha_i = \mathbf{W}_0 \, \sigma(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1 \,), \qquad (16)$$

$$\mathbf{h}_a = \sum_{i=1}^{n} \alpha_i \, \mathbf{h}_i, \qquad (17)$$

$$\hat{z}_i = \, s_i^T \mathbf{h}_a, \qquad (18)$$

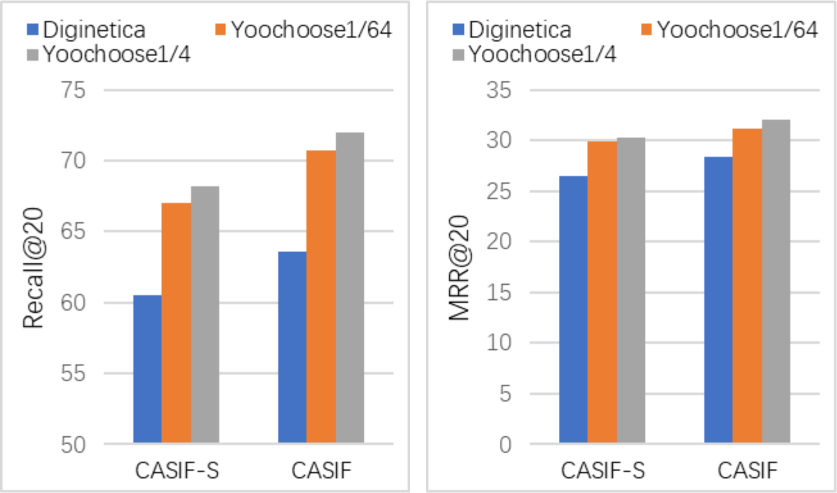


Figure 3. Recommendation performance of CASIF-S and CASIF in terms of Recall@20 (left) and MRR@20 (right) on three real-world datasets.

4.2.4. Analysis on Session Sequence Length

In order to compare the performance of the model on different session length datasets. We divide Yoochoose1/64 and Diginetica into two groups respectively. One is a short session datasets, which is marked as "Short", and the session length is less than or equal to 5. The other group of datasets has each session length greater than 5, which is named as "Long". We chose representative, advanced model (STAMP, SR-GNN) as baselines. The experimental results under Recall@20 and MRR@20 are shown in table 4 and table 5 respectively. The best results have been highlighted in boldface. As we can see, our model performs excellent on both the "Long" dataset and the "Short" dataset. On the "Short" dataset, our model is comparable to SR-GNN, which thanks to the power of graph neural networks. On the whole, our method is superior, it is mainly due to the fact that our model is mainly composed of GNN module and short-term interest first module. GNN can perform well on long session sequences, while short-term interest first module can perform well on short session sequences. It can be seen that our model is suitable for both recommendation tasks for long sessions sequences and recommendation tasks for short sessions sequences.

Table 4. The performance of different methods with different session lengths evaluated in terms of Recall@20

Method	**Yoochoose1/64**		**Diginetica**	
	Short	**Long**	**Short**	**Long**
STAMP	71.44	64.73	47.26	40.39
SR-GNN	70.69	70.70	50.49	21.27
CASIF-S	69.52	70.68	50.32	20.89
CASIF	**71.56**	**70.88**	**51.12**	**51.36**

Table 4. The performance of different methods with different session lengths evaluated in terms of MRR@20

Method	Yoochoose1/64		Diginetica	
	Short	Long	Short	Long
STAMP	32.60	24.31	26.26	25.33
SR-GNN	**31.15**	30.93	27.49	26.27
CASIF-S	29.52	28.78	26.45	25.33
CASIF	31.02	**31.28**	**28.12**	**29.83**

5. CONCLUSIONS

In this paper, we propose context-aware short-term interest first model for session-based recommendation, which combines GNN with short-term attention priority. GNN is used to capture the local dependencies of the session, while the short-term interest first module fully considers the current interest and long-term interest, which can extract the main intention of the session. We make full use of the complementarity of the two, and a large number of experimental results on two data sets demonstrate the superiority of our proposed model. However, the running time of this model is not optimized compared with previous methods. In the following work, we will continue to try to improve the structure of the model to make the model less complex and more efficient.

ACKNOWLEDGEMENTS

I would like to thank the tutor for his guidance on my research direction and the conference organizers and reviewers!

REFERENCES

[1] J.B. Schafer, J.A. Konstan, J. Riedl, Recommender systems in e-commerce, in Proceedings of the 1st ACM conference on Electronic commerce, 1999, pp. 158–166.

[2] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In WWW, pages 811–820. ACM, 2010.

[3] B. Hidasi, A. Karat-zoglou, L. Baltrunas, and D. Tikk. Session based recommendations with recurrent neural networks. ICLR, 2016.

[4] Yong Kiam Tan, Xinxing Xu, and Yong Liu. Improved recurrent neural networks for session-based recommendations. In RecSys, pages 17–22. ACM, 2016.

[5] Pengpeng Zhao, Haifeng Zhu, Yanchi Liu, Jiajie Xu, Zhixu Li, Fuzhen Zhuang, Victor S. Sheng, and Xiaofang Zhou. Where to go next: A spatio-temporal gated network for next poi recommendation. In AAAI, 2019.

[6] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In CIKM, pages 1419–1428. ACM, 2017.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, pages 5998–6008, 2017.

[8] Q. Liu, Y. Zeng, R. Mokhosi, H. Zhang. STAMP: Short-term attention/memory priority model for session-based recommendation, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1831–1839.

[9] Yi Cao, Weifeng Zhang, Bo Song, Weike Pan, Congfu Xua. Position-aware context attention for session-based recommendation. Neurocomputing Volume 376, 1 February 2020, Pages 65-72

[10] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al.

Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, 2018.

[11] G. Linden, B. Smith, J. York, Amazon. com recommendations: Item-to-item collaborative filtering, IEEE Internet Comput. (1) (2003) 76–80.

[12] B.M. Sarwar, G. Karypis, J.A. Konstan, J. Riedl, et al., Item-based collaborative filtering recommendation algorithms, WWW 1 (2001) 285–295.

[13] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the twentyfifth conference on uncertainty in artificial intelligence, UAI'09, AUAI Press, 2009, pp. 452–461.

[14] Mnih, A., and Salakhutdinov, R. 2007. Probabilistic matrix factorization. In Advances in neural information processing systems, 1257–1264.

[15] Koren, Y.; Bell, R.; and Volinsky,C. 2009. Matrix factorization techniques for recommender systems. Computer 42(8):30–37.

[16] Shani, G.; Brafman, R. I.;and Heckerman, D. 2002. An mdp-based recommender system. InUAI, 453–460.

[17] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C.D. Manning, Position-aware attention and supervised data improve slot filling, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 35–45.

[18] C. Zhou, J. Bai, J. Song, X. Liu, Z. Zhao, X. Chen, J. Gao, Atrank: An attention-based user behavior modeling framework for recommendation, Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[19] Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line: Large-scale information network embedding. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, 1067–1077. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

[20] Grover, A., and Leskovec, J. 2016. Node2vec: Scalable feature learning for networks. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 855–864. New York, NY, USA: ACM.

[21] Kipf, T. N., and Welling, M. 2016. Semisupervised classification with graph convolutional networks. In Proceedings of the 2016 International Conference on Learning Representations, ICLR '16.

[22] Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The graph neural network model. IEEE Transactions on Neural Networks 20(1):61–80.

[23] Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. S. 2015. Gated graph sequence neural networks. In Proceedings of the 2015 International Conference on Learning Representations, volume abs/1511.05493 of ICLR '15.

[24] Li, Z.; Ding, X.; and Liu, T. 2018. Constructing narrative event evolutionary graph for script event prediction.

[25] Li, R.; Tapaswi, M.; Liao, R.; Jia, J.; Urtasun, R.; and Fidler, S. 2017b. Situation recognition with graph neural networks. In 2017 IEEE International Conference on Computer Vision (ICCV), 4183–4192.

[26] Marino, K.; Salakhutdinov, R.; and Gupta, A. 2017. The more you know: Using knowledge graphs for image classification. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 00,20–28.

[27] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, T. Tan, Session-based recommendation with graph neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33 of AAAI '19, 2019, pp. 346–353.

[28] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the twenty fifth conference on uncertainty in artificial intelligence, UAI'09, AUAI Press, 2009, pp. 452–461.

[29] B.M. Sarwar, G. Karypis, J.A. Konstan, J. Riedl, et al., Item-based collaborative filtering recommendation algorithms, WWW 1 (2001) 285–295.

[30] W. Wang, W. Zhang, S. Liu, B. Zhang, L. Lin, H. Zha, Clicks: Modeling Multi-Relational Item Graph for Session-Based Target Behavior Prediction, WWW '20 , 2020.

[31] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. 2019. Graph Contextualized Self Attention Network for Session-based Recommendation. In IJCAI. 3940–3946.

AUTHORS

Haomei Duan is currently a master student in the College of Computer Science and Technology, Heilongjiang University, Harbin, China. She received the B.S. degree in College of Computer Science and Technology from Shandong University of Technology. Her research focuses on recommendation algorithms and systems.

Jinghua Zhu received the B.S. degree in computer software and the M.S. degree in computer science in 1999 and 2002, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology in 2009. She has been a Professor with the School of Computer Science and Technology, Heilongjiang University, China, since 2016. She has published many high quality conference and journal research papers. Her research interests include social networks, data mining, uncertain databases, and wireless sensor networks.

A COLOR IMAGE BLIND DIGITAL WATERMARKING ALGORITHM BASED ON QR CODE

Xuecheng Gong and Wanggen Li

School of Computer and Information,
Anhui Normal University, Anhui Wuhu, China

ABSTRACT

With the rapid development of network technology and multimedia, the current color image digital watermarking algorithm has the problems of small capacity and poor robustness. In order to improve the capacity and anti-attack ability of digital watermarking. A color image blind digital watermarking algorithm based on QR code is proposed. The algorithm combines Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT). First, the color image was converted from RGB space to YCbCr space, and the Y component was extracted and the second-level discrete wavelet transform is performed; secondly, the LL2 subband was divided into blocks and carried out discrete cosine transform; finally, used the embedding method to embed the Arnold transform watermark information into the block. The experimental results show that the PSNR of the color image embedded with the QR code is 56.7159 without being attacked. After being attacked, its PSNR is more than 30dB and NC is more than 0.95. It is proved that the algorithm has good robustness and can achieve blind watermark extraction.

KEYWORDS

QR Code, Color Image, Arnold Transform, DWT

1. INTRODUCTION

With the development of the Internet, various digital products have appeared on the Internet. At the same time, with the rapid development of information digitization, there has also been a problem that copyright is not easy to protect [1]. Digital watermarking is a method to solve the problem of copyright protection. According to different embedding methods, it is divided into spatial domain watermark and transform domain watermark. The current digital watermark embedding into the vector transform domain better image can be improved watermark robustness and security [2].

The transform domain watermark technology embeds watermark information into the corresponding frequency coefficients through frequency domain transformation. Common methods include DWT and DCT. The low-frequency component embedding watermark through DCT transformation has strong robustness However, the anti-attack ability is weak[3]. The mixed use of DWT and DCT can balance the robustness and imperceptibility of the watermark image.
Quick response code (quick response code) referred to as QR code, it can store a lot of information. Therefore, using QR code as a watermark can not only improve the robustness of the watermark, but also store more copyright information [4]. Arnold transform has periodicity and is widely used in image scrambling. Using its characteristics to transform QR code information can improve the security of watermarking [5].

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 67-75, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110405

2. RELATED WORK

He et al. [6] proposed a color image watermarking algorithm based on discrete wavelet transform, discrete cosine transform and singular value decomposition (DWT-DCT-SVD). First, convert the carrier image from RGB color space to YUV color space; then, perform a layer of discrete wavelet transform on the brightness component Y, use discrete cosine transform to decompose the low frequency and divide it into blocks, and perform singular value decomposition on each block ; Finally, embed the watermark into the carrier image. However, there is a problem with the DWT-DCT-SCD method. A non-blind watermark image is needed to extract the watermark in the experiment and the watermark non-QR code is used in the experiment. Xu Jiangfeng et al. [7] proposed a digital watermarking scheme combining QR code, chaotic system and DWT-DCT. Carry out DWT operation on the carrier image and perform 4×4 block and DCT operation on the low frequency subbands. Then embed the QR code watermark through the chaotic system into the carrier image. The experimental results show that after the Gaussian noise attack, the PSNR value and NC value are low. The experiment uses gray-scale images, which is less practical. To solve the above problems, this paper proposes a color image blind digital watermarking scheme based on QR code. This scheme selects a color image as the carrier image, converts the RGB color space to the YCbCr color space, embeds the watermark into the luminance component Yand uses the Arnold transformation to encrypt the QR code. While improving the security and robustness of the watermark. It also increased the amount of watermark informatione.

2.1. Discrete Wavelet Transform

In digital image processing, it is necessary to discretize continuous wavelet and wavelet transform. Discretized wavelet and corresponding wavelet transform are called discrete wavelet transform [8]. Discrete wavelet transform is a spatio-temporal scale analysis method of information analysis theory and signal. It has multiple scales in the space and frequency domainand can continuously decompose images from low resolution to high resolution [9,10]. In addition, the DWT algorithm has a wide range of applications in the digital watermarking field. At present, many innovative and efficient joint algorithms related to DWT have been proposed.

In this paper, the RGB color space of the color carrier image is transferred to the YCbCr color space. Perform DWT transformation on Y component to obtain the horizontal and vertical low frequency LL, the horizontal low frequency and the vertical high frequency LH, the horizontal high frequency and the vertical low frequency HL, high-frequency components HH in the horizontal and vertical directions. DWT is performed on the LL subband again to obtain the low frequency component LL2. The low-frequency components represent image features. The high-frequency components represent the edges and details of the image. Embedding the watermark in the low frequency component LL2 can improve the robustness. As shown in Figure 1.

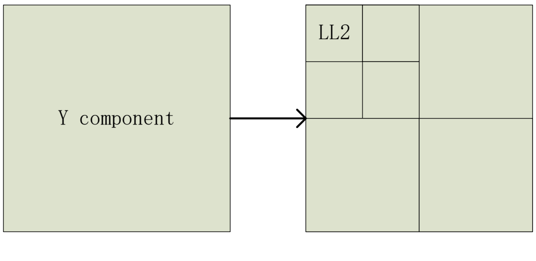


Figure 1. Wavelet decomposition diagram.

2.2. Discrete Cosine Transform

Discrete Cosine Transform can transform the spatial domain signal into the frequency domain signal and has good decorrelation performance [11,12]. The important information of the image after discrete cosine transform is concentrated on the middle and low frequency coefficients. The position is the upper left corner of the DCT matrix, which has the ability to resist attacks. After being attacked, the embedded watermark information can still be extracted [13,14]. In this paper, the LL2 subband is divided into blocks, the blocks are subjected to DCT transformation, then the medium and low frequency coefficients are selected for watermark embedding. the DCT inverse transformation completes the image reconstruction.

2.3. Arnold Transform

The watermark image contains important information. The Arnold transform is used to scramble the image to achieve information encryption. At the same time, the Arnold transform is periodic. The number of scrambling can be used as the watermark key to further enhance its security [15]. The periodicity of Arnold transformation refers to the continuous transformation of the original image, the original image is obtained after t times. The transformation period t is related to the size of the image M×N [16]. The Arnold transformation is defined as follows:

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \left(\mathrm{mod}\, N\right) \qquad (1)$$

Among them, $(\mathrm{x,y})$ represents the pixel position before image scrambling, $(\mathrm{x}^*,\mathrm{y}^*)$ represents the image pixel location after scrambling, N represents the order of the image matrix.

QR code is a type of matrix two-dimensional code. It has many characteristics such as high-speed reading, high capacity, support for error correction processing, wide coding range, low costand easy production. Due to the high capacity of QR codes, using QR codes as digital watermarks can increase the information capacity. Because of its support for error correction processing, the robustness of the watermark can also be improved.

3. Embedding and Extraction of Watermark

This article uses QR code as a watermark which adds more information and improves the security of the watermark. The use of color images as carrier images is more widely used.

3.1. Embedded Watermark

(a) Convert the color carrier image from RGB color space to YCbCr color space according to the algorithm flow and extract the brightness component Y. (b) Perform a two-level DWT transformation on the luminance component to obtain the low frequency subband LL2, then implement 2×2 block division on the LL2 subband to obtain a block matrix. (c) Perform DCT transformation on each block (dct=dct(LL2)) to obtain the transformed DCT matrix. Then extract the first value in the matrix from the DCT transformed block to form a new matrix F. (d) Use Arnold transform algorithm to scramble the original watermark image W to get the scrambled watermark W^* . (e) Embed the watermark W^* into the matrix F using equations 2 and 3 to obtain the matrix F^* . Then replace each value of the matrix F^* with the first value of each block in turn, and perform inverse DCT transformation on each block to obtain $LL2^*$.

$$\lambda_1^* = \begin{cases} \lambda_1 - T + 3a/4 & T \geq a/4 \\ \lambda_1 - T - a/4 & other \end{cases} \qquad W^*(i,j) = 0 \qquad (2)$$

$$\lambda_1^* = \begin{cases} \lambda_1 - T + 5a/4 & T\,^3\; 3a/4 \\ \lambda_1 - T + a/4 & other \end{cases} \qquad W^*(i,j) = 1 \qquad (3)$$

Among them, $T = \lambda_1 \mathrm{moda}$, a is the embedding strength. Used to control the invisibility and robustness of embedded watermarks. λ_1 represents each value in each block matrix. λ_1^* Represents the embedded value.

(f) Implement the second-level inverse DWT transformation on the obtained in the previous step to obtain the component. The brightness component of the embedded QR code is converted from the YCbCr color space to the RGB color space to obtain the color carrier image embedded in the QR code. Figure 2 is a flowchart of watermark embedding.

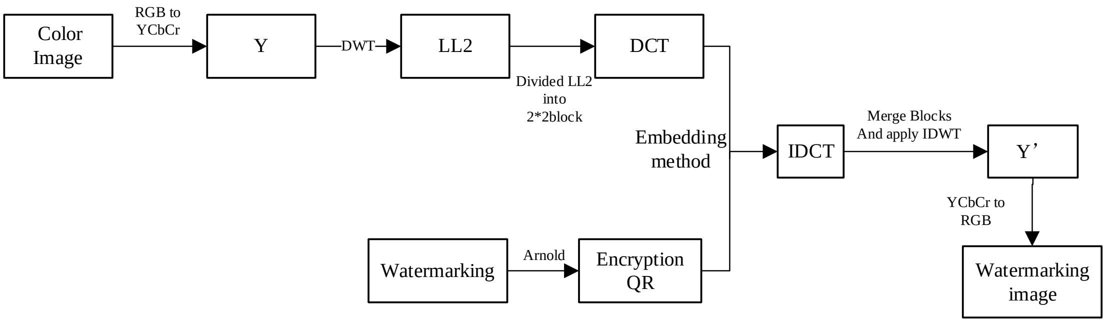


Figure 2. Watermarking embedding flow chart.

3.2. Extract Watermark

(a)First, convert the color carrier image embedded with QR code from RGB color space to YCbCr color space and extract the brightness component Y^*. (b) Perform two-level DWT transformation on the luminance component to obtain the low-frequency subband. Then perform 2×2 block on the subband. (c) Perform DCT transformation on each block to obtain the transformed DCT matrix. dct=dct($LL2^*$). Then extract the first value in the matrix from the block after DCT transformation to form a matrix F^*. (d) The extraction of watermark information is the reverse process of watermark embedding. The watermark is extracted by equations 4.

$$W(i,j) = \begin{cases} 1 & T > a/2 \\ 0 & other \end{cases} \qquad (4)$$

(e) According to the obtained W^* in the previous step. Then perform Arnold transformation on it. Finally the watermark W is extracted. Figure 3 is a flow chart of watermark extraction.

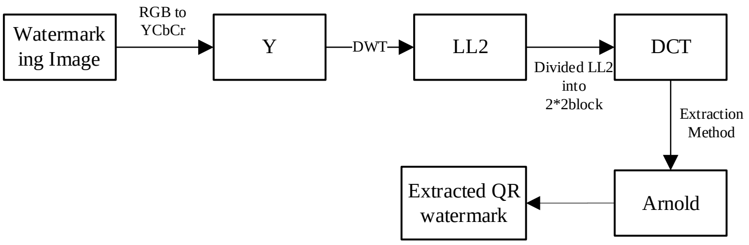
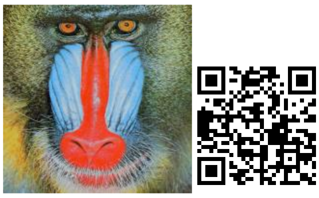


Figure 3. Watermarking extraction flow chart.



(a)Color Image (b)QR Code

Figure 4. Experiment Pictures.

4. THE EXPERIMENTAL RESULTS

The experimental environment is: Intel Core i5-4210M CPU; 2.60GHz frequency; Windows 10 64-bit operating system; Matlab2018a software. Select the color carrier image with 512×512 pixels, the 64×64 QR code is the watermark and the QR code carries the information. Figure 4(a) is a color carrier image, and Figure 4(b) is a QR code watermark image.

4.1. Watermark Evaluation Standard

Experiments usually use peak signal-to-noise ratio (PSNR) to measure the difference between the QR code-embeddsssed image and the unembedded original image. The greater the PSNR, the higher the recognition of the image embedded in the QR code with the original image. The definition of PSNR is:

$$PSNR = 10\log_{10}\frac{255^2 M \times N}{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}(\mathrm{I}^*(x,y) - \mathrm{I}(x,y))^2} \qquad (5)$$

Among them, I^* represents a watermarked color carrier image and I represents a color carrier image. When the PSNR value is greater, the color carrier image is closer to the watermarked color carrier image. The embedded watermark effect is better. PSNR>30dB usually means that the watermark is invisible and the image quality is better.

The normalized correlation coefficient (NC) represents the similarity between the original watermark and the extracted watermark. The NC value ranges from 0 to 1. The closer the NC value is to 1, the higher the similarity between the original watermark and the extracted watermark. The definition of NC is:

$$NC = \frac{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} W(x,y)\mathrm{W}^*(x,y)}{\sqrt{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} W(x,y)^2 \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} W^*(x,y)^2}} \qquad (6)$$

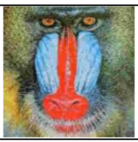
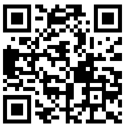
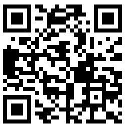
Among them, W represents the original watermark image. W^* represents the extracted watermark image.

4.2. Analysis of Results

In order to test the robustness of the algorithm, the experiment uses JPEG compression, noise filtering, cropping attack, rotation attack and median filtering to attack watermarked images.

a) JPEG compression.

Table 1. JPEG Compression.

			
(a)JPEG(10%)	(b)Extracted(a=1)	(c)JPEG(80%)	(d)Extract(a=1)

In the experiment, the JPEG compression attack was performed on the image. Table 1(a)-(d) are the results of the JPEG compression attack under different factors. It can be measured that the extracted QR code can be recognized by the machine.

According to the data in Table 2, when the JPEG compression factor is 10%, the PSNR value is 48.9158 and the NC value is 1. At this time, the QR code can still be identified, indicating that the algorithm in this paper is robust against JPEG compression attacks and can ensure the integrity of the watermark information.

Table 2. JPEG Compression attack data.

Methods	Compression Ratio					
	10%	20%	30%	40%	60%	80%
PSNR/dB NC	48.9158 1	50.0294 1	50.6425 1	51.0356 1	51.6767 1	52.5922 1

b) Noise attack. The experiment uses salt and pepper noise and Gaussian noise attacks, as shown in Table 3.

Table 3. Noise attack.

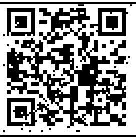
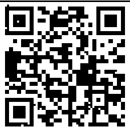
			
(a)SaltandPepper(0.15)	(b)Extracted(a=2)	(c)Gaussian(0.04)	(d)Extract(a=2)

Table 4.Noise attack data.

Method	Salt and Pepper Noise			Gaussian Noise			
	0.01	0.05	0.15	0.01	0.02	0.03	0.04
PSNR/dB	45.3121	41.7544	37.9263	43.0325	42.9243	42.7836	42.5694
NC	1	1	09687	1	0.9997	09995	0.9962

It can be seen from Table 4 that the PSNR value of the color carrier image embedded with the QR code is attacked by salt and pepper noise and Gaussian noise respectively. The PSNR value is also above 30dB. It indicating that the attacked carrier image shows strong robustness and can resist noise attack. From the perspective of the NC value, the extracted watermark maintains a high consistency with the original watermark and the extracted watermark can be identified.

c) Cropping attack. It can be seen from the different cropping areas in Table 6 that the QR code is embedded in the frequency domain of the carrier image.The cropping of different areas can still maintain a high NC value, the PSNR value can also be maintained above 30dB and the QR code can be Identified. Therefore, this algorithm has good resistance to shearing attacks.

Table 5. Cropping attack.

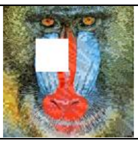
			
(a)Cropping(1/16)	(b)Extracted(a=2)	(c)Cropping(1/128)	(d)Extract(a=2)

Table 6. Cropping attack data.

Method	Cropping Ratio					
	1/8	1/16	1/32	1/64	1/128	1/256
PSNR/dB	41.3916	41.3724	43.0706	45.3883	45.8662	46.0592
NC	0.9600	0.9589	0.9782	0.9868	0.9928	0.9951

d) Rotation attack and Median filtering. It can be concluded from Table 7 that the algorithm can resist rotation attacks and median filtering. The PSNR values are all above 30dB, and the NC values remain above 0.9. It shows that the algorithm has good robustness.

Table 7. Rotation attack and Median filtering data.

Method	Rotation Angle		Median Filtering
	5°	10°	[3×3]
PSNR/dB	42.4715	40.4322	45.2474
NC	0.9693	0.9558	0.9991

e) Comparison results. It can be seen from Table 8 that in the median filter attack, the algorithm in this paper is better than the non-blind QR code watermarking algorithm in the literature [7]. In this paper, the NC values of Gaussian noise and salt and pepper noise are both higher than 0.9. When the JPEG compression factor is 30%, the NC value reaches 1, which is higher than the NC value in the literature [7]. It can be concluded that the algorithm in this paper is more robust than the algorithm in [7].

Table 8. Compare the results of different experiments.

Attacks	NC(Proposed Method)	NC(Reference [7])
JPEG Compression30%	1	0.9875
Cropping 1/16	0.9589	—
Gaussian Noise 0.04	0.9962	0.9300
Salt and Pepper Noise 0.05	1	0.9849
Median Filtering [3×3]	0.9991	0.9873

5. CONCLUSION

In this paper, color images are used as carrier images and QR codes are used as watermark information to increase the watermark information carrying capacity, it also improves the watermark's anti-attack ability. Implementation of Arnold transformation on QR code, then implementation of DWT, DCT and block operations on carrier images. attack experimental data show that the algorithm in this article is resistant to JPEG compression, clipping attack, Gaussian noise, salt and pepper noise, median filtering, and rotation attacks. The PSNR values are all above 30dB. The algorithm proposed in this paper also has shortcomings. In a rotation attack, when the rotation angle exceeds 12°, the extracted QR code will not be recognized and information cannot be obtained through the device. In view of the above problems, the algorithm needs to be continuously improved.

REFERENCES

[1] P. Rasti & S. Samiei, M.et al, (2016) "Robust non-blind color video watermarking using QR decomposition and entropy analysis," J. Vis. Commun. Image Represent, vol. 38, no. 5, pp. 838–847.

[2] J. Wang & Z. Du, (2019)"A method of processing color image watermarking based on the Haar wavelet," J. Vis. Commun. Image Represent., vol. 64, pp. 102627.

[3] S. Kushlev & R. P. Mironov, (2020) "Analysis for Watermark in Medical Image using Watermarking with Wavelet Transform and DCT," 2020 55th Int. Sci. Conf. Information, Commun. Energy Syst. Technol, pp. 185–188, 2020.

[4] Bai TaoTao & Liu Zhen et al, (2014) "Contourlet domain digital watermarking algorithm based on QR code,"Optoelectronics•Laser,vol. 25, no. 4, pp. 769-776.s

[5] Li Guohe & Chen Chen,et al, (2018) "Digital watermark insertion and extraction method for QR code, "Computer Engineering and Applications, vol. 55, no. 10, PP. 103-107, 114.

[6] Y. He & Y. Hu, (2018) "A Proposed Digital Image Watermarking Based on DWT-DCT-SVD," Proc. 2018 2nd IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. 2018, pp. 1214–1218.

[7] Xu Jiangfeng & Zhang Shouqiang, (2018) "DWT-DCT digital watermarking algorithm based on QR code,"Application Research of Computers, vol. 35, no. 5, pp. 1540–1544.

[8] Yuan Zihan & Liu Decheng, et al, (2020) "New image blind watermarking method based on two-dimensional discrete cosine transform," Optik (Stuttg), vol. 204, no. February 2020, pp. 164152.

[9] Zhu Jianzhong & Yao Zhiqiang, (2014). "Color image blind watermarking algorithm based on DWT-SVD and Turbo code,"Journal of Jilin University: Science Edition, vol. 52, no. 4 pp. 773–778.

[10] Y. A. Mekarsari & D. R. I. M. Setiadi, et al, (2018) "Non-blind RGB image watermarking technique using 2-level discrete wavelet transform and singular value decomposition," 2018 Int. Conf. Inf. Commun. Technol, vol. 2018-Janua, pp. 623–627.

[11] M. Ali & C. W. Ahn, et al, (2014) "A robust image watermarking technique using SVD and differential evolution in DCT domain," Optik (Stuttg), vol. 125, no. 1, pp. 428–434.

[12] Sunesh & R.Rama Kishore, (2019) "A Novel and Efficient Blind Image Watermarking in Transform Domain," Procedia Comput. Sci, vol. 167, no. 2019, pp. 1505–1514.

[13] D. O. Munoz-Ramirez & V. Ponomaryov, et al, (2018) "A robust watermarking scheme to JPEG compression for embedding a color watermark into digital images," Proc. 2018 IEEE 9th Int. Conf. Dependable Syst. Serv. Technol, pp. 619–624.

[14] D. O. Muoz Ramirez & V. Ponomaryov, et al, (2019) "Embedding a Color Watermark into DC coefficients of DCT from Digital Images," IEEE Lat. Am. Trans, vol. 17, no. 8, pp. 1326–1334.

[15] Wu Fengbo & Wang Feng, (2016) "Wavelet transform digital image watermarking algorithm based on HVS,"Applied Optics, vol. 35, no. 2, pp. 254–259.

[16] S. Saadi & A. Merrad, (2019) "Novel secured scheme for blind audio/speech norm-space watermarking by Arnold algorithm," Signal Processing, vol. 154, pp. 74–86.

AUTHORS

Xuecheng Gong born in 1994, male, who studies at the Anhui Normal University, researching image processing.

Wanggen Li born in 1973, male, professor at the Anhui Normal University. His current research interests include biological computing and intelligent computing.

AN EXPERIENCE IN ENHANCING MACHINE LEARNING CLASSIFIER AGAINST LOW-ENTROPY PACKED MALWARES

Shang-Wen Chen, Tzu-Hsien Chuang,
Chin-Wei Tien and Chih-Wei Chen

Cybersecurity Technology Institute, Institute for
Information Industry, Taipei, Taiwan R.O.C

ABSTRACT

Both benign applications and malwares would take packing for their different purposes to conceal the real part of the program processes. According to recent research reports, existing machine learning (ML) approach-based malware detection engines are difficult to effectively classify the packed malwares, especially when they are in low entropy packed.

Recently, we counted and found that the ratio of low-entropy packed ransomware is extremely high. This would cause a high error rate of the result on currently used ML approaches. Thus, we propose a new method to extract entropy-related features and use a stack model to build up an ML malware engine to effectively detect low-entropy packed malwares. We evaluate our method by using over 15,000 malware samples collected from VirusTotal and compare the result to related researches. This experience reports our adopted model and features can significantly lower the error rate of low-entropy packed detection from 11% to 1%.

KEYWORDS

Malware detection, low-entropy packing, machine learning classification

1. INTRODUCTION

Machine learning has already been widely used in many fields, such as data analytics, predictive analytics, natural language processing (NLP), sentiment analysis, computer vision, and information security. In the field of information security, malware detection has already been applied to improve detection accuracy.

In the normal process of machine learning, the selected features are extracted from each sample of the training set. These features are used to train a model that can be used to detect malware. If the model learns the features of samples correctly, then it can be used to detect malware, which has similar characteristics to such samples. However, the attackers try to use obfuscated techniques to disguise malware to a normal executable file. Attackers adopt several common obfuscation techniques, such as packing, encryption [12], data confusion. After such processing, the extracted features from the sample could differ from the original features. In addition, these incorrect features may lead to incorrect prediction results by the pre-trained model, and this case was discussed by Aghakhani et al. [4].

The above-mentioned problem is commonly solved by determining whether packing is performed before extracting features from the sample. According to the result of judgment, different

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 77-86, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110406

processing methods are used for packed and non-packed samples. Thus, the correct judgment could be used to extract the correct features. Moreover, the correct features can lead to the correct prediction.

Previously, most researchers almost use entropy [1] [9] [10] to judge whether a new sample is a packed. Entropy is a metric used to measure the uncertainty in a series of numbers or bytes. Moreover, packing is a technique that can hide or disguise the internal behavior of samples. In addition, if an attacker uses a packer to pack a sample, the corresponding bytes between the original and packed samples differ considerably. Thus, the entropy value of a packed sample is assumed to be high. That is, a sample with high entropy is assumed to be the same as a packed sample.

However, in some exceptional cases, the results of known methods or tools show the packed samples to possess low entropy. Such cases have already been mentioned in previous studies [8]; however, these were not considered, because the authors claimed that such a case is extremely rare and can therefore be disregarded.

We collected ransomware samples from the VirusTotal dataset [13], with the time interval between July 2019 and June 2020. We used YARA tool [14] and PEPackerInfo [15] to judge whether these samples are packed and determine their entropy values. The threshold value that indicates whether the sample has low entropy is 7, which was also adopted by previous studies [2]. Table 1 shows that low-entropy packed samples not only exist but also take a large proportion in the dataset of our collected packed samples. Obviously, it is quite different with the claim of Han et al [8]. Thus, the larger percentage of low-entropy packing ransomware is, the higher possibility of detection error rate is. If the case of ransomware is extended to all malware, it is not to ignore the influence of low-entropy packing malware to detection error rate anymore.

Table 2 shows the best detection result of the error rate of the machine learning model with entropy-related features in the research of Mantovani et al. [3]. They attempted to determine the effect of low-entropy packed samples on machine learning. Table 2 shows an error rate of at least 11% by the machine learning model in the detection of packing. This proves that the effect of the low-entropy sample is too large to disregard.

Table 1. Statistics of the amount of different entropies in packed samples collected from VirusTotal

	packed	
Collected Time	**high**	**low**
19/7	1	6
19/8	0	23
19/9	1	12
19/10	0	16
19/11	16	66
19/12	3	60
20/1	328	1785
20/2	16	96
20/3	29	339
20/4	18	1209
20/5	36	1331
20/6	8	640

Table 2. Best result of error-rate detection using machine learning with entropy-related features [3]

Classifier	Train-Testing	$\mathrm{Err}_{\mathrm{notPacked}}(\mathrm{W})$	$\mathrm{Err}_{\mathrm{packed}}(\mathrm{W})$
MLP	75%~25%	6.34%	12.70%
	50%~50%	6.87%	16.14%
	25%~75%	6.89%	11.91%

We have had considered an assumption that if the threshold value is changed from 7 to another nearby values, the error rate of machine learning may be decreased. However, this assumption has been proved that it is almost not effective for reducing the threshold value. On the contrary, the error rate could be increased. Figure 1 shows 341 distinct entropy values in each range in our collected low-entropy dataset and the sum of the amount of the different entropies within the ranges 5-6 and range 6-7 is 93.5% of the total amount. Because the values are centralized near 7, if this threshold is decreased, the error rate could increase. Therefore, we tried to find a method to lower error rate without changing threshold value. That is, the objective of this study is to decrease the detection error rate of machine learning in low-entropy packing malware so that a new model could detect all low-entropy packing malware correctly.

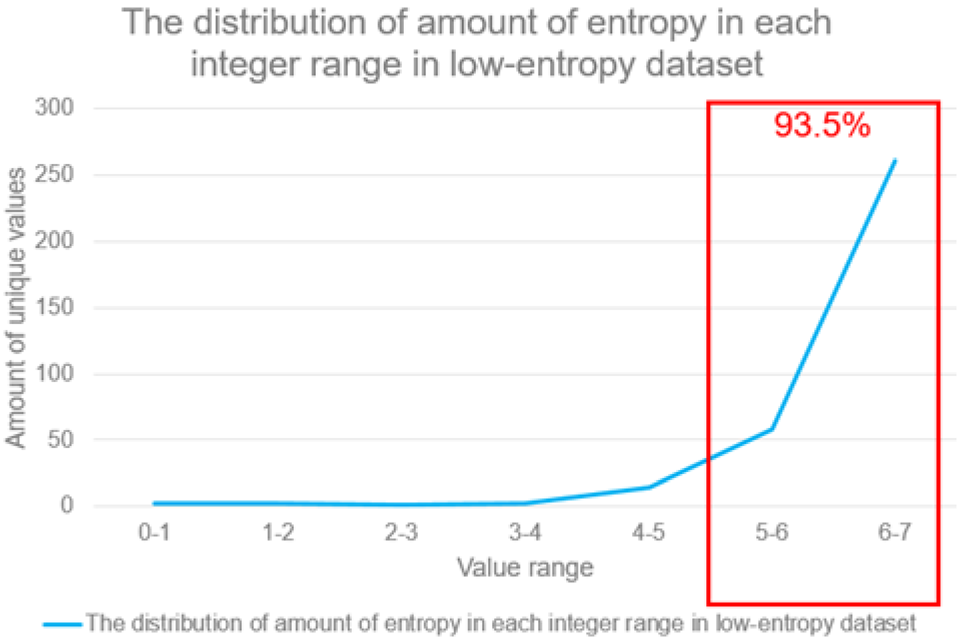


Figure 1. Distribution of the amount of distinct entropy values in each integer range in the low-entropy dataset

The remainder of this paper is organized as follows: Section 2 provides an overview of literature on entropy packing detection and describes the central concept of our adopted model. Section 3 introduces and details our proposed model. Section 4 details the results of experiments with different datasets. Section 5 presents the conclusions of this study.

2. Related Works

This section first reviews the literature on entropy packing detection in recent years, and then we detail as to why we adopted our model as the solution.

As described in section 1, entropy has become an indicator of packing in research. Thus, many studies [7] [8] have adopted it as main feature or use some of its features in packing.

In 2008, Perdisci et al. [11] proposed features that captured specific anomalies introduced by packers in the portable executable (PE) [29] file format. The authors applied pattern-recognition techniques for fast detection of packed executables so that only executables detected as packed are sent to an universal unpacker such as UPX [16], PackerID [17], and NFD [18]. However, the limitation of such a method is that unknown packed files cannot be unpacked because of the use of universal unpacking tools. In contrast, we aimed to detect all packing samples that are not limited to only those packed by universal packers.

In 2012, Ugarte-Pedrero et al. [5] selected entropy as the main unpacking feature, and conducted several experiments by using some samples of the Zeus botnet. Zeus is one of the first bot families to adopt a low-entropy packing scheme. However, their proposed method was customized to a single specific case. Thus, it fails to consider the samples that use other common low-entropy techniques.

In 2016, Raphel et al. [6] attempted to refine the use of entropy to recognize samples adopting an XOR-based scheme. XOR encryption is recognized as a form of obfuscation that is mainly used to encrypt small parts of a code such as shellcodes. However, their solution aims to a specific problem, and the method is therefore not applicable to common packing detections. In contrast, our solution can be used for common packing detection.

These researches attempted to solve the entropy-packing detection problem, but their methods could not either be a general solution or handle samples which are packed with unusual packers. As described in section 1, low-entropy packing problem displays that entropy does not be a directly indicator of packing anymore. However, we believe that entropy is still an effective factor to judge packing problem. Thus, we proposed a new usage of entropy-related features and used these features in our proposed model. Several different algorithms have been proposed for packing detection in recent years, for example, random forest [19], deep neural network [20]. Although each algorithm has its own advantages, we combined these advantages for building our model. Thus, we adopted a stacked model that can combine the results of different models to output a balanced result. As a result, the error rate of the model that uses new entropy-related features decreases effectively.

The contribution of this study to literature is two-fold. First, we propose a new method for extracting entropy-related features. Second, our adopted stacked model effectively decreases the error rate by verifying 15,000 samples from two datasets with 0.25% and 0.46% error rates.

3. PROPOSED METHOD

3.1. Detailed Description for Our Adopted Model

In this subsection, we discuss in detail why we chose to combine the advantages of algorithms. The first reason is the stability of performance of different algorithms. As described earlier, different algorithms have their own advantages because of their own characteristics. For example, the support vector machine (SVM) [21] is good at bisection question [33]. However, SVM does not perform well in other aspects. Thus, the total performance of a single algorithm is considered unstable, and therefore we disregarded the use of only one algorithm to build a model. The second reason is the overfitting problem. Each model has the possibility to overfit a specific dataset under different conditions. However, the possibility of overfitting multiple models is lower than the possibility of overfitting of a single model. Thus, we adopt a stacked model for machine learning. The model comprises three characteristics: the use of multiple models, merging the results of multiple models, and lowering the error rate.

3.2. Model Framework Description

Figure 2 depicts the structure of stacked model the inputs of which are the features extracted from samples. The model outputs the percentage of unpacking and packing. The process of the stacked model is divided into two stages.

The first stage portrays multiple models, the inputs of which are the same as those of the stacked model. Although we adopted the four models displayed in the top part of Figure 2 in the first stage, they can be replaced by the other models when needed. The outputs of each model display the corresponding percentage value of unpacking and packing.

In the second stage, the inputs of the stack model comprise eight output values of the models in the first stage. These values are processed by the stacked model, and the balanced values obtained after processing are the outputs of the stacked model.

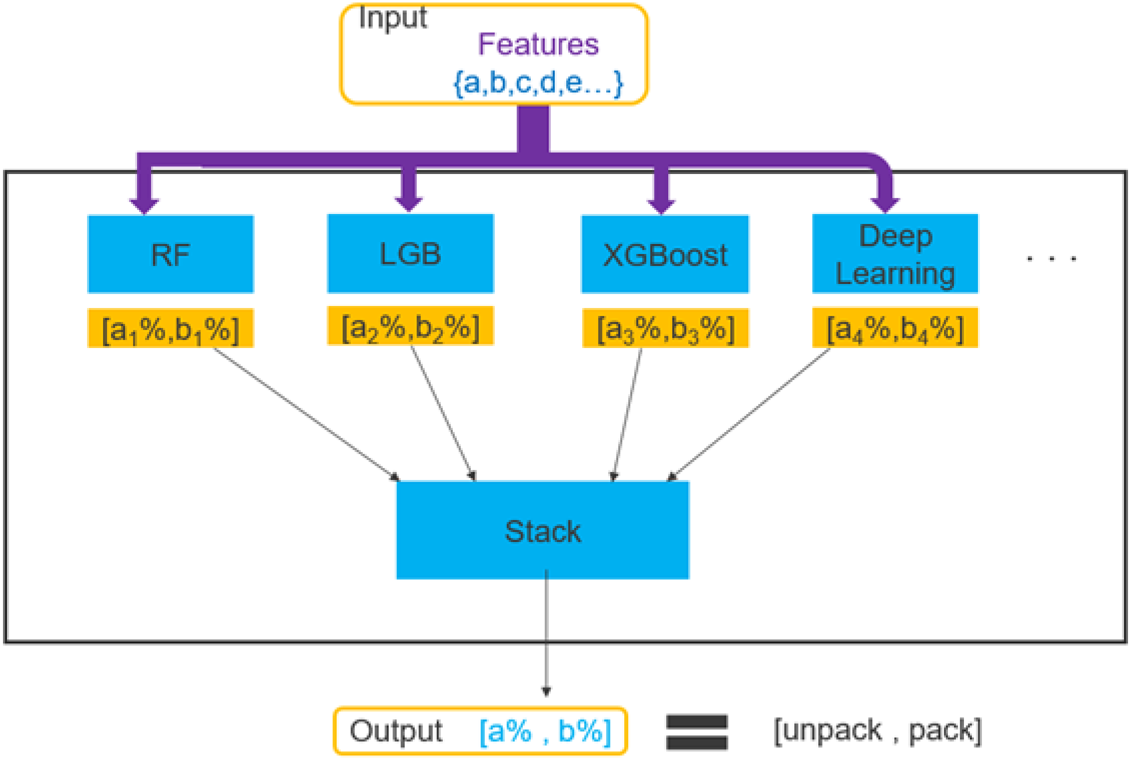


Figure 2. Structure of the proposed stacked model

3.3. Feature Selection

In this study, we adopted 7,068 features in our model, and the detailed composition is described as follows. Our features are divided into three categories: assembly, byte and keyword features. Moreover, we used IDA pro [22] and xxd [23], which is a Linux command, to generate the required assembly and byte files.

For assembly features, we adopted opcode, registers, and metadata, for which we selected 26 common registers in the x86 [24] architecture. Similarly, we also chose 93 common opcodes in the x86. Finally, we selected 26 metadata of a sample, comprising data such as file size and total lines of assembly file.

For byte features, we adopted 1-gram, metadata, byte string lengths, image, and entropy as bytes features. The image features were acquired using mahotas [25], which is a Python package, to calculate haralick [26] features of a byte image. Further, we adopted 15 entropy-related features in this study. Figure 3 shows the distribution of the number of sections in each sample in the collected dataset. As shown, the most common number of sections in a sample is five. However, 15 is the largest number of sections in a sample. To adapt the model itself to most sample cases, we decided to adopt 15 entropy values from sample sections. Assuming that the total number of sections in a sample is k; if k < 15, the remaining 15-k parameters will be assigned as zero. On the contrary, if k ≥ 15, only the first 15 entropy values will be used, while the others are disregarded by the system.

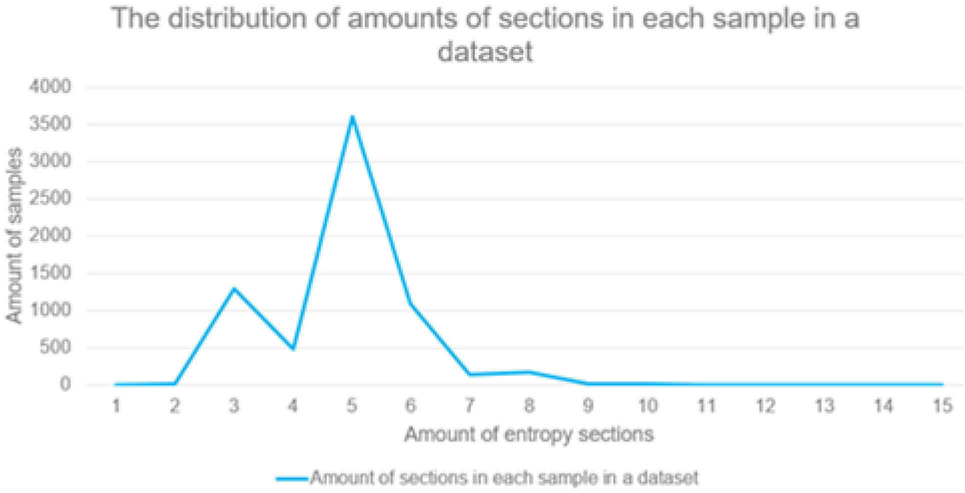


Figure 3. Distribution of amounts of sections in each sample in collected dataset

For keyword features, we counted the number of occurrences of 6,482 keywords. The keywords were decided by the following process. First, we filtered the keywords that occur over 100 assembly files. Then, these keywords were filtered twice using the feature_importance function of the XGBoost [32] model per 10,000 keywords. Next, we acquired 8,580 keywords from the filtered processing, and then erased unnecessary keywords, such as memory locations and uncompiled data. Finally, we obtained a total of 6,482 keywords.

3.4. Model-Training Method

To prove the effectiveness of the new usage of entropy-related features, we generate two feature sets: one containing all of the extracted features from the sample and the other containing the extracted features without the 15 entropy-related features. Then, these two feature sets were used to train their models separately.

4. EXPERIMENTS

4.1. Dataset

We used two datasets in our experiments: one was acquired from [3], and the other was collected from VirusTotal; these are termed as dataset1 and dataset2, respectively. The samples collected from VirusTotal are all ransomware, and the collecting period is between May 29 and June 30, 2020. We selected 15,000 and 5,000 samples from dataset1 for training and testing, respectively. Similarly, for dataset2, we selected 10,000 samples each as the training and testing sets.

4.2. Data Labelling

For dataset1, we adopted the labels used by Mantovani et al. [3]. The training set of dataset1 comprises 7,500 packed and unpacked samples each. In addition, the testing set of dataset1 comprises 2,500 samples each as packed and unpacked. For dataset2, the labels were decided using our proposed processing procedure, which is described in the following text. As a result, we obtained 5,000 samples each as packed and unpacked samples in the training set as well as in the testing.

The proposed processing procedure has two stages. In the first stage, the input was a new unknown sample, which was filtered by known tools first. The tools that we used in this stage were PackerID, NFD, DIE [27] and ExeScan [28]. If the output of any tool indicates that the unknown sample is packed, then this sample will be labelled as packed. However, if all of the outputs of the four tools indicate that the unknown sample is unpacked, it will enter the second stage.

The second stage uses five static features to distinguish whether the sample is packed. The names and detailed descriptions are listed in Table 3. We also tested the performance of several other static features, such as the amounts of executable sections, entropy of Portable Executable (PE) [29] header, and entropy of whole file. However, the most distinguishable features are the five features that we adopted.

Table 3. Description of static features

Features name	Descriptions
rwx section number (rwx)	The amounts of sections that can read, write, and execute simultaneously
Non-standard section number (nss)	The amounts of sections, the name of which do not exist in the Microsoft list
Execution only section number (exe)	The amounts of sections that can only perform the execute task
[.text section] virtual size > rawdata size (.text)	In general, the size of a virtual address is greater than the size of rawdata in the .text section
Import number of address table & .dll (iat & dll)	The amount of import address tables < 50 and dll < 4

Figure 4 shows the processing flow chart of the second stage. The sample was judged according to a specific order of static features sequentially. This process involves the following five steps, the results of which is one of True or False. Step 1 analyzes the sample to check whether there exists at least a section with the access authorization of read, write and executable simultaneously. In step 2, the sample is examined to check whether there exists at least a section name that is not listed in the standard section name list of Microsoft [30]. In step 3, the sample is examined to check whether the virtual size of .text section is greater than the rawdata size of .text section. In step 4, the sample is analyzed to check whether there exists at least one section that owns only the accessibility of executable. In step 5, the sample will be checked to determine whether there exists at least one section with <50 import address table and <4 dynamic link library (dll) [31].

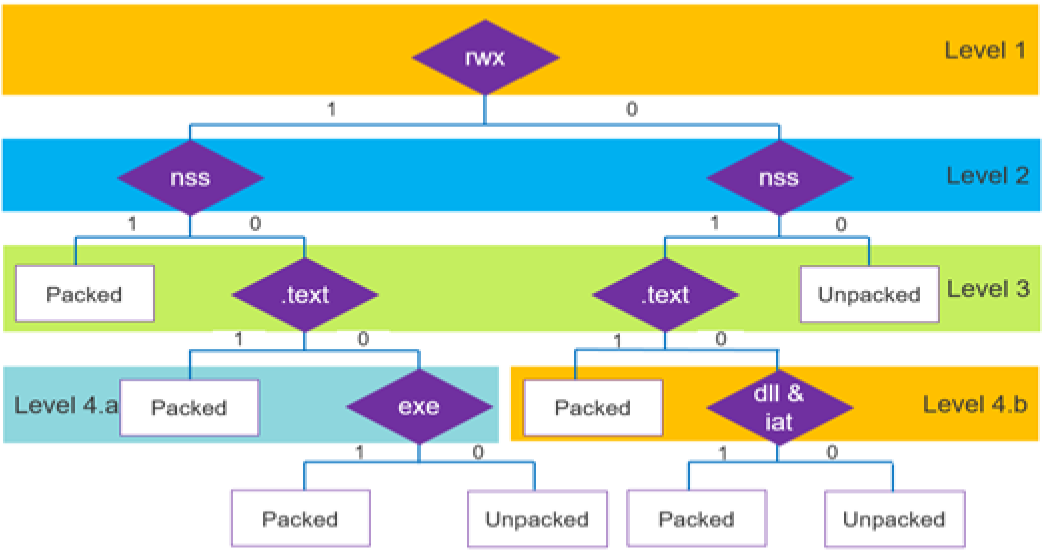


Figure 4. Diagram of processing flow chart of second stage in data labelling

There are five levels in Figure 4, and the concepts of labelling process in each level described as follows:

Level 1: judge if there exists any section that owns access authority of read, write and execute at the same time
Level 2: judge if there exists any non-standard section in sample
Level 3: judge if there exists the case that the size of virtual address > the size of rawdata in the text section
Level 4.a: judge if there exists any section which only owns execution authority
Level 4.b: judge if there exists any section whose amount of import address tables < 50 and dll < 4

4.3. Experimental Result

We conducted two experiments, one using dataset1 and the other using dataset2. In each experiment, we trained two models with different features. One uses all extracted features from the dataset, while the other uses all extracted features except the entropy-related features from the same dataset.

Table 4 shows the result of the error rate of dataset1. We also compared our results with the results of Mantovani et al. [3]. Parameter w indicates the vectors of all features, and parameter w' indicates the vectors of all features except the entropy-related features. We observed that the error rate of the proposed model with all features of packed samples is better than that of the model in [3]. Moreover, the error rate of the packed samples with all features in our model was only 0.25%.

Table 4. Performance of error rate of dataset$_1$

	$\mathbf{Err_{unpack}(w)}$	$\mathbf{Err_{packed}(w)}$	$\mathbf{Err_{unpack}(w')}$	$\mathbf{Err_{packed}(w')}$
Dataset$_1$	0.76%	0.25%	0.8%	0.29%
Mantovani et al. [3]	6.89%	11.91%	6.33%	12.93%

Table 5 displays the results of the error rate of dataset2 compared with the results of Mantovani et al. [3]. As shown, the error rate of packed samples using the proposed model with all features is still better that of the previous model [3]. In addition, the error rate of our model to dataset2 is 0.46%.

Table 5. Performance of error rate of dataset$_2$

	$\mathbf{Err_{unpack}(w)}$	$\mathbf{Err_{packed}(w)}$	$\mathbf{Err_{unpack}(w')}$	$\mathbf{Err_{packed}(w')}$
Dataset$_2$	1%	0.46%	0.88%	0.48%
Mantovani et al. [3]	6.89%	11.91%	6.33%	12.93%

Moreover, we observed that the error rate of packed samples with all features is better than the error rate of packed samples without entropy-related features, as shown in Tables 4 and Table 5; therefore, the effectiveness of the proposed usage of entropy-related features is proved. In summary, our proposed model can effectively lower the error rate of detection of packed samples, and our new usage of entropy-related features helps to reduce the error rate of the model.

5. Conclusion and Future Work

In this paper, we proposed a new method for extracting entropy-related features. We combined these features and other common features and applied them to our adopted stacked model. Moreover, our stacked model effectively decreases the error rate. We verified the performance of our model using 15,000 samples from two different datasets, the error rates of which were obtained as 0.25% and 0.46%.

The training samples of dataset2 that we used in the experiments were collected from 29 May to 30 June, 2020. Moreover, we continuously collected ransomware samples from VirusTotal. We hope to train new models with different lengths of time intervals and evaluate their performances. Then, we can decide the best time interval to retrain the model to keep a better performance.

References

[1] G. Jacob, P. M. Comparetti, M. Neugschwandtner, C. Kruegel, G. Vigna, "A Static, Packer-agnostic Filter to Detect Similar Malware Samples," in Proc. of the Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), 2013

[2] X. Ugarte-Pedrero, D. Balzarotti, I. Santos, P. G Bringas. "Sok: Deep packer inspection: A longitudinal study of the complexity of run-time packers." in 2015 IEEE Symposium on Security and Privacy (SP) , 2015, pp. 659–673.

[3] A. Mantovani, S. Aonzo, X. Ugarte-Pedrero, A. Merlo, D. Balzarotti, "Prevalence and Impact of Low-Entropy Packing Schemes in the Malware Ecosystem," in: Network and Distributed System Security (NDSS) Symposium, NDSS 20, 2020.

[4] H. Aghakhani, F. Gritti, F. Mecca, M. Lindorfer, S. Ortolani, D. Balzarotti, G. Vigna, C. Kruegel, "When Malware is Packin' Heat; Limits of Machine Learning Classifiers Based on Static Analysis Features," in Network and Distributed System Security (NDSS) Symposium, NDSS 20, 2020

[5] X. Ugarte-Pedrero, I. Santos, B. Sanz, C. Laorden, P. G. Bringas. “Countering entropy measure attacks on packed software detection,” in Consumer Communications and Networking Conference (CCNC), 2012, pp. 164–168.

[6] J. Raphel, P. Vinod. “Information theoretic method for classification of packed and encoded files,” in Proceedings of the 8th International Conference on Security of Information and Networks, SIN ’15, ACM, New York, NY, USA, 2015, pp. 296–303.

[7] R. Lyda, J. “Hamrock. Using entropy analysis to find encrypted and packed malware,” IEEE Security & Privacy, vol. 5, no. 2, 2007

[8] S.-W. Han, S.-J. Lee. “Packed pe file detection for malware forensics,” KIPS Transac.: PartC, vol. 16, no. 5, pp. 555–562, 2009.

[9] M. Z. Shafiq, S. Tabish, M. Farooq, “PE-Probe: Leveraging Packer Detection and Structural Information to Detect Malicious Portable Executables,” in Proc. of the Virus Bulletin Conference (VB), 2009

[10] R. Arora, A. Singh, H. Pareek, U. R. Edara, “A Heuristicsbased Static Analysis Approach for Detecting Packed PE Binaries,” International Journal of Security and Its Applications, 2013

[11] R. Perdisci, A. Lanzi, W. Lee. “Classification of packed executables for accurate computer virus detection,” Pattern Recog. Lett., vol. 29, no. 14, pp. 1941–1946, 2008.

[12] J. A. Clark, “Invited Paper. Nature-Inspired Cryptography: Past, Present and Future,” Citeseer, pp. 1647-1654, 2003.

[13] VirusTotal. https://www.virustotal.com

[14] GitHub–Yara-Rules/rules: Repository of yara rules. https://github.com/Yara-Rules/rules

[15] PEPackerInfo. https://sites.google.com/site/robertoperdisci/projects/cpexe

[16] UPX. https://upx.github.io/

[17] PackerID. https://github.com/sooshie/packerid

[18] NFD. https://github.com/horsicq/Nauz-File-Detector/releases/

[19] L. Breiman. Random forests. Machine Learning, 45(1): 5–32, 2001.

[20] I. Arel, D. Rose, R. Coop, "DeSTIN: A Scalable Deep Learning Architecture with Applicationto High-Dimensional Robust Pattern Recognition," Proc. of the AAAI 2009 Fall Symposiumon Biologically Inspired Cognitive Architectures (BICA), November, 2009

[21] Osuna E, Freund R, Girosi F (1997) Support vector machines: Training and applications

[22] IDA Pro – Hex Rays. https://www.hex-rays.com/products/ida/

[23] xxd linux command man page. https://www.commandlinux.com/man-page/man1/xxd.1.html

[24] coder32 edition | X86 Opcode and Instruction Reference 1.12. http://ref.x86asm.net/coder32.html

[25] mahotas. https://pypi.org/project/mahotas/

[26] Mahotas – Haralick features. https://www.geeksforgeeks.org/mahotas-haralick-features/

[27] DIE. https://github.com/horsicq/DIE-engine/releases

[28] ExeScan. https://github.com/cysinfo/Execscan

[29] M. Pietrek, “Peering Inside the PE: A Tour of the Win32 Portable Executable File Format,” Microsoft Systems Journal, vol. 9, no. 3, 1994, pp. 15-34.

[30] PE Format. https://docs.microsoft.com/en-us/windows/win32/debug/pe-format#special-sections

[31] Dynamic-link library. https://www.computerworld.com/article/2585730/dynamic-link-libraries.html

[32] XGBoost Documentation. https://xgboost.readthedocs.io/en/latest/

[33] Bisection method. https://www.sciencedirect.com/topics/engineering/bisection

Towards Adversarial Genetic Text Generation

Deniz Kavi

Text generation is the task of generating natural language, and producing outputs similar to or better than human texts. Due to deep learning's recent success in the field of natural language processing, computer generated text has come closer to becoming indistinguishable to human writing. Genetic Algorithms have not been as popular in the field of text generation. We propose a genetic algorithm combined with text classification and clustering models which automatically grade the texts generated by the genetic algorithm. The genetic algorithm is given poorly generated texts from a Markov chain, these texts are then graded by a text classifier and a text clustering model. We then apply crossover to pairs of texts, with emphasis on those that received higher grades. The approach described in this paper was designed to be as modular as possible and as such, changes to the grading system and further improvements to the genetic algorithm are to be the focus of future research.

1 Introduction

Text generation can be described as a "next word prediction" problem. This method of approaching text generating can be explained as, given a string of words, predict what the next word will be. Originally, text generation algorithms used a small part of the input text. For example, an algorithm might attempt to predict a word using just the previous two words of the sentence. However, since the information the algorithm has access to is limited, its ability to generate a coherent text is also limited.

Models that use "attention" to determine what parts of the text are relevant to what will follow. When trained on a large corpus of natural language, these models are currently the state of the art in natural language processing and text generation [8][1][12]. Attention allows neural network models to "pay attention" to only the relevant parts of the previous information. Whereas a Markov chain or a frequency based model would only have knowledge of some part of a sentence, neural networks using attention are able to pick required and relevant information from previous sentences. That being the case, models using attention can "remember" much more information than their predecessors as relevant words are more important than randomly picked words.

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 87-95, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110407

Although researchers have used genetic algorithms for text generation [6][5], genetic algorithms have received little attention relative to deep learning approaches. In this paper, we propose an adversarial approach to the problem of text generation with genetic algorithms.

2 Related Works

In this section, we will evaluate previous approaches to the two components of our approach: text grading and generation. More specifically, we will be examining how effective they are at the target task and whether they could be applied to our own research.

2.1 Markov Chain for Text Generation

Markov chains can be represented as a sequence of states, where certain states come after each other. The order of the respective states is determined by the probability of of how these states are ordered in the input text dataset. And although any type of data, ranging from financial data to weather data can be represented in the form of states, we've represented text as a series of states, each word being a state. After the input text, in our case the ASAP dataset, is processed via the Markov chain, the model will be able to predict and generate the most probable word that will follow the it was given. We used Markov chains for our initial population, we used Markov chains because they performed less well compared to their deep learning counterparts. The reason we wanted it to perform worse was so that most of the task of text generation could be left to the genetic algorithm without causing problems in the grading system. As the grading algorithms hadn't seen samples of entirely randomly generated texts their ability to grade them would be low, so we used Markov chains that the grader could understand without lowering the genetic algorithm's contribution.

2.2 Transformer Language Models

Transformers are unsupervised machine learning models trained on a large corpus of the target language to predict the word or token which will follow the input text. Unlike Markov chains, transformers aren't constrained by the number of states they can have in memory, thanks to a parameter called attention. With emphasis on attention, a transformer model is able to remember relevant details of the previous text, which would be part of the text data that it pays attention to, allowing for it to recall relevant details without the need to analyze massive amounts of information.

2.3 Adversarially Learned Neural Outlines

In Subramanian et al. [11] the authors propose the usage of a generator, first adversarially producing a sentence outline and then generating words sequentially conditioned by both the outline and previous outputs. This is inspired by GANs [4] and Autoencoder [10] models in that there are generator and discriminator neural networks. They fit "a non-parametric kernel density estimator(KDE) on the samples produced by a GAN and then evaluating the likelihood of real examples under this KDE."

2.4 Previous Genetic Algorithm Approaches to Text Generation

In Manzoni et al. [6], the authors describe a process by which they use word embeddings instead of the words themselves as input to the genetic algorithm. They first mapped every word of sample sentences of k length to a wor2vec vector, applying mathematical operations to the vectors and decoding the modified vectors, finally interpreting them as words. The mutation and crossover parts of a genetic algorithm would be done through linear algebra operations as the words are encoded as vectors.

2.5 Automated Essay Grading

An earlier system, the Intelligent Essay Assesor(IEA) [3], uses Latent Semantic Analysis [2] to grade essays. It measures and takes the sum of individual words' "meanings" to evaluate the whole passage's meaning. IAE compares the input essay to other essays in terms of the quality of its content and its form. The drawbacks with this approach is that the grader will be entirely unable to compare and thus, grade essays that it hasn't seen examples of.

A more artificial intelligence oriented system called IntelliMetric [9] uses manually determined syntactic and semantic features to feed into machine learning algorithms. This approach is very similar to ours in that the problem is essentially framed as a text classification task, but likely with a different dataset.

3 Methods

The primary difference of our approach to the problem of text generation is the use of a "grader", which is a supervised machine learning model trained on a dataset [1] of essays and their human graded scores. We used the training set with 12977 sample essay with labels(grades). The grades above 50(which there were 2 of) were changed to 50 to fit the way the classifier processed data.

We used an XLnet [12] based classifier, though any text classification or regression method would be usable for the dataset and the task of essay grading. XLnet is originally a language model trained on a large English corpus, we replaced the last layer of its architecture to a softmax layer to fit the task of text classification, so that it returns a grade when given a text as input.

[1] https://www.kaggle.com/c/asap-aes/data

To determine if the essays had topical consistency, we implemented a text clustering algorithm, which, without seeing the texts' labels would seperate samples into "clusters" based on similarities to other texts. The algorithm used was Scikit-learn's [7] "MiniBatchKMeans" model for clustering. The inputs to the clustering model were TF-IDF vectors for The Automated Student Assessment Prize(ASAP) dataset, with the stopwords removed. Instead of predicting what topic(cluster) each text belongs to, we only need to make sure that it belongs to any topic, as the model having a high confidence in the text belonging to a topic means that the text has a consistent topic, an attribute which should be rewarded. To calculate how closely a model follows its topic we take the reciprocal of the distance between the model's prediction and the cluster center, where predictions closer to the cluster center means that the model is more confident that the text belongs to a particular topic. The sum of the grade given to the essay by the text classifier and the text clustering model was used as the fitness function of the genetic algorithm, where those with higher fitness values will have a higher probability of passing their genes on to future populations.

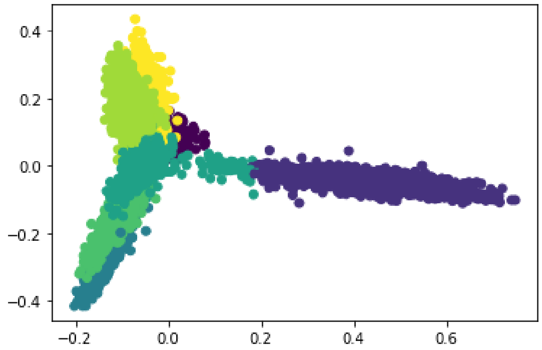


Figure 1: Clusters generated by the KMeans model, visualized in 2D

The Genetic algorithm first starts with a population of essays generated using Markov chains, which is a simpler and less successful approach simply working with the last word and picking the most probable word that would follow from its vocabulary. We used the open source markovify [2] library to implement this technique. We picked a weaker algorithm because our approach to grading these texts would not give reliable results on entirely random sequences of words or strings. Markov chains also increase the speed of the process as the children of the population are more likely to be readable if they are formed from reasonably readable parents.

This first population is then moved to the mating pool in relation to the scores they receive, for example a text that scored an 8 would be copied to the mating pool 8 times and one with a score of 1 would be copied only a single time.

[2] github.com/jsvine/markovify, accessed in August 2020

Doing so allows for a higher probability of texts with higher scores “breeding” and exchanging attributes or genes with other texts. Crossover is then applied to the population in the mating pool breed better future essays. Sentences of the two partners going through crossover are passed down to the child where the child would become a mix of the two parents. See Figure 2 as an example. In Figure 2, there is a 50 percent chance of a sentence from the first text and a 50 percent chance that the sentence is selected from the second text, in this specific example, by pure luck, sentences from the first text were selected more frequently. There also is a small probability(1/10) that individuals words in the sentences might be changed to those of the text’s partner. Crossover is further explained in Algorithm 1. The actual program code was written in python.

Algorithm 1: Crossover

```
Data: array $T_1$ of sentences; array $T_2$ of sentences; float $m$, mutation
      rate
Result: array child, a combination of the two input texts
empty array child;
prob ← random(0,1);
if $prob < 0.5$ then
    add sentence from $T_1$ to child;
end
else
    add sentence from $T_2$ to child;
end
if $m > random(0, 1)$ then
    replace a word in n sentence of child with corresponding word from
      $T_1$ or $T_2$
end
```

Algorithm 2: Text Generation with Genetic Algorithm

```
Data: integer $f$, minimum fitness required for the algorithm to stop;
      integer $maxGen$, maximum number of generation for the
      algorithm to stop
Result: Population of computer generated texts
generation $\leftarrow 0$;
create initial population of texts generated by Markov chains;
get fitness for each individual text in population;
while $Fitness < f$ and $generation < maxGen$ do
    add text as many times as its score to the mating pool;
    perform crossover;                    ▷ as described in algorithm 1
    empty population;
    add children to population;
    increase generation by 1;
end
```

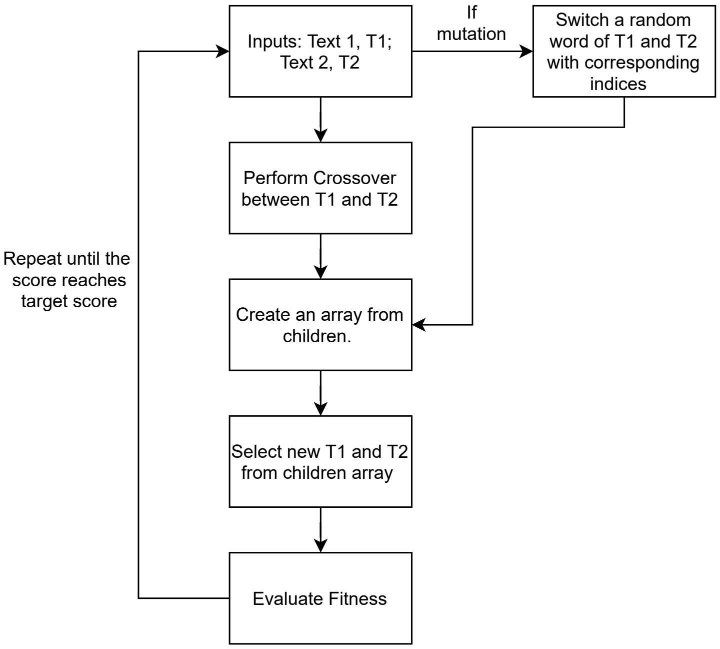
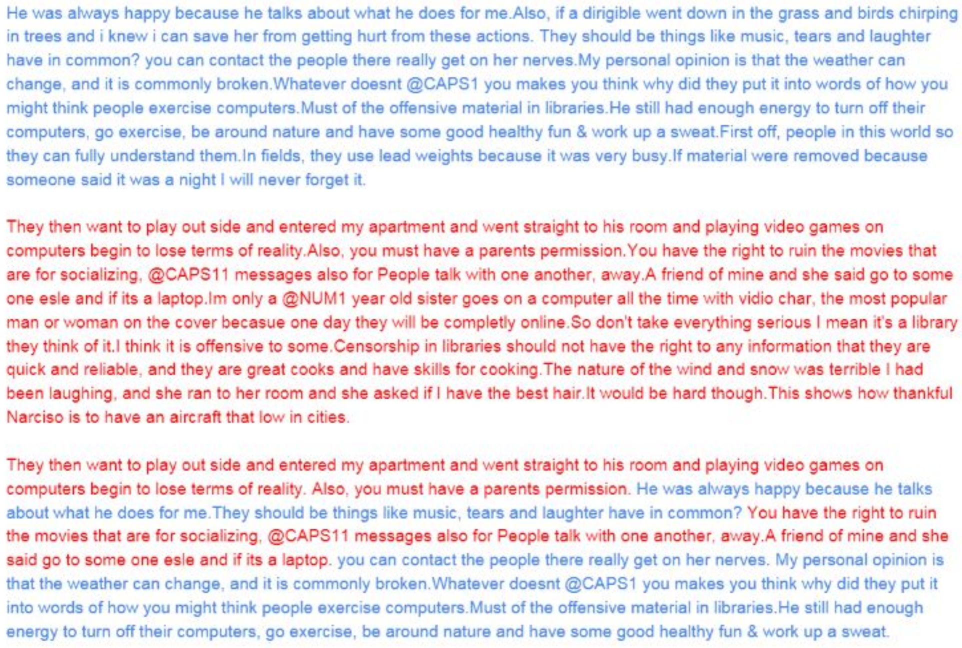

Inputs: Text 1, T1; Text 2, T2
If mutation
Switch a random word of T1 and T2 with corresponding indices
Perform Crossover between T1 and T2
Repeat until the score reaches target score
Create an array from children.
Select new T1 and T2 from children array
Evaluate Fitness

Figure 2: Flowchart of the Text Generation Process



He was always happy because he talks about what he does for me.Also, if a dirigible went down in the grass and birds chirping in trees and i knew i can save her from getting hurt from these actions. They should be things like music, tears and laughter have in common? you can contact the people there really get on her nerves.My personal opinion is that the weather can change, and it is commonly broken.Whatever doesnt @CAPS1 you makes you think why did they put it into words of how you might think people exercise computers.Must of the offensive material in libraries.He still had enough energy to turn off their computers, go exercise, be around nature and have some good healthy fun & work up a sweat.First off, people in this world so they can fully understand them.In fields, they use lead weights because it was very busy.If material were removed because someone said it was a night I will never forget it.

They then want to play out side and entered my apartment and went straight to his room and playing video games on computers begin to lose terms of reality.Also, you must have a parents permission.You have the right to ruin the movies that are for socializing, @CAPS11 messages also for People talk with one another, away.A friend of mine and she said go to some one esle and if its a laptop.Im only a @NUM1 year old sister goes on a computer all the time with vidio char, the most popular man or woman on the cover becasue one day they will be completly online.So don't take everything serious I mean it's a library they think of it.I think it is offensive to some.Censorship in libraries should not have the right to any information that they are quick and reliable, and they are great cooks and have skills for cooking.The nature of the wind and snow was terrible I had been laughing, and she ran to her room and she asked if I have the best hair.It would be hard though.This shows how thankful Narciso is to have an aircraft that low in cities.

They then want to play out side and entered my apartment and went straight to his room and playing video games on computers begin to lose terms of reality. Also, you must have a parents permission. He was always happy because he talks about what he does for me.They should be things like music, tears and laughter have in common? You have the right to ruin the movies that are for socializing, @CAPS11 messages also for People talk with one another, away.A friend of mine and she said go to some one esle and if its a laptop. you can contact the people there really get on her nerves. My personal opinion is that the weather can change, and it is commonly broken.Whatever doesnt @CAPS1 you makes you think why did they put it into words of how you might think people exercise computers.Must of the offensive material in libraries.He still had enough energy to turn off their computers, go exercise, be around nature and have some good healthy fun & work up a sweat.

Figure 3: Simplified Crossover Example

4 Discussion and Future Work

If the basic structure of our approach is kept, then there would be an initial text population, scored with an automated system and crossed-over with the other individuals within the population. The first attribute of the algorithm, the initial population, could be changed from Markov chains to entirely random sequences of words if the grader is able to work with random texts. Or any other population of texts that could be combined to generate more meaningful texts. The grading system could also be replaced with any system that would be able to grade medium length texts automatically at a reasonable speed. The implementation of crossover could also be changed, possibly with taking into account the fact that words can be represented as vectors with word embeddings. When words represented as vectors, arithmetic operations can also be applied to words to change their meanings or as a way to perform the steps of the genetic algorithm. More types of mutations could also be added to improve variety in word choice and order. In summary, the algorithm can be largely modified while most of its core properties can be kept. This should allow for further experimentation for text generation with genetic algorithms.

5 Conclusion

We demonstrate a proof-of-concept for a genetic algorithm for text generation using automated essay grading as the fitness function. The algorithm uses sample texts generated by Markov chains trained on the ASAP dataset and merges those samples with each other to produce texts with higher grades. Our system for grading is a combination of a text classifier and text clustering algorithm trained on the aforementioned dataset. The highest score achieved by the text generated by the genetic algorithm was 54/58.

Most of the components of the approach described may be changed based on differing needs. The process used to generate the initial text population may be changed from Markov chains to any process that outputs text. Future work may even create an entirely random initial population. We chose to use Markov Chains as they provided somewhat but not completely meaningful text. Experimentation with mutation rates should also provide an increase in the quality of the texts generated. The method used for evaluating the quality of the text is also replaceable as any process that gives a quantitative assessment of texts could be used as a fitness function. Other approaches to text evaluation could be added to the algorithms described in this paper, which likely would increase performance.

6 Acknowledgments

We thank Dr. David Perkins(Hamilton College) for his guidance and help throughout the development of this project.

References

[1] Tom B. Brown et al. *Language Models are Few-Shot Learners.* 2020. arXiv: 2005.14165 [cs.CL].

[2] Peter Foltz. "Latent Semantic Analysis for Text-Based Research". In: *Behavior Research Methods* 28 (Feb. 1996), pp. 197–202. DOI: 10.3758/BF03204765.

[3] Peter Foltz, Darrell Laham, and T. Landauer. "The intelligent essay assessor: Applications to educational technology". In: *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* (Apr. 1999).

[4] Ian J. Goodfellow et al. *Generative Adversarial Networks.* 2014. arXiv: 1406.2661 [stat.ML].

[5] Ruli Manurung, Graeme Ritchie, and Henry Thompson. "Using genetic algorithms to create meaningful poetic text". In: *J. Exp. Theor. Artif. Intell.* 24 (Mar. 2012), pp. 43–64. DOI: 10.1080/0952813X.2010.539029.

[6] Luca Manzoni et al. *Towards an evolutionary-based approach for natural language processing.* 2020. arXiv: 2004.13832 [cs.CL].

[7] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[8] A. Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019.

[9] Lawrence Rudner, Veronica Garcia, and Catherine Welch. "An Evaluation of IntelliMetric$^{\mathrm{TM}}$ Essay Scoring System". In: *Journal of Technology, Learning, and Assessment* 4 (Jan. 2006).

[10] Juergen Schmidhuber. *Deep Learning in Neural Networks: An Overview.* 2014. arXiv: `1404.7828 [cs.NE]`.

[11] Sandeep Subramanian et al. "Towards Text Generation with Adversarially Learned Neural Outlines". In: *NeurIPS 2018.* Dec. 2018. URL: `https://www.microsoft.com/en-us/research/publication/towards-text-generation-with-adversarially-learned-neural-outlines/`.

[12] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* 2019. arXiv: `1906.08237 [cs.CL]`.

ADGRAPH: ACCURATE, DISTRIBUTED TRAINING ON LARGE GRAPHS

Lizhi Zhang, Zhiquan Lai, Feng Liu and Zhejiang Ran

Parallel and Distributed Key Laboratory of National Defence Technology,
National University of Defence Technology, Changsha, China

ABSTRACT

Graph neural networks (GNNs) have been emerging as powerful learning tools for recommendation systems, social networks and knowledge graphs. In these domains, the scale of graph data is immense, so that distributed graph learning is required for efficient GNNs training. Graph partition-based methods are widely adopted to scale the graph training. However, most of the previous works focus on scalability other than the accuracy and are not thoroughly evaluated on large-scale graphs. In this paper, we introduce ADGraph (accurate and distributed training on large graphs), exploring how to improve accuracy while keeping large-scale graph training scalability. Firstly, to maintain complete neighbourhood information of the training nodes after graph partitioning, we assign l-hop neighbours of the training nodes to the same partition. We also analyse the accuracy and runtime performance of graph training, with different l-hop settings. Secondly, multi-layer neighbourhood sampling is performed on each partition, so that the mini-batch generated can accurately train target nodes. We study the relationship between convergence accuracy and the sampled layers. We also find that partial neighbourhood sampling can achieve better performance than full neighbourhood sampling. Thirdly, to further overcome the generalization error caused by large-batch training, we choose to reduce batchsize after graph partitioned and apply the linear scaling rule in distributed optimization. We evaluate ADGraph using GraphSage and GAT models with ogbn-products and Reddit datasets on 32 GPUs. Experimental results show that ADGraph achieves better performance than the benchmark accuracy of GraphSage and GAT, while getting 24-29 times speedup on 32 GPUs.

KEYWORDS

Graph neural networks; Distributed training; Multi-GPU; Deep learning; Parameter Server.

1. INTRODUCTION

Graph neural networks (GNNs) are becoming more and more influential in solving various challenges in many practical applications, such as social networks [1], paper citations [2], biological networks [3, 4], product customer relationships [5], recommendation systems [1], and knowledge graphs [6], which data can be naturally represented as graph structures. The graph data structure is widely used to model data with complex connections between elements because of good expressive ability. The powerful function of GNNs in modelling the dependency relationship between graph nodes has made a great breakthrough in the research field related to graph analysis, which is an emerging field in deep learning [7, 8].

Simultaneously, the scale of graphs in industry domains has developed rapidly [9]. For example, the social network maintained by Facebook has nearly 2 billion users, and Amazon's customer shopping network has hundreds of millions of nodes. Larger datasets and network structures can improve the accuracy of tasks. Technologies that can effectively analyse and process large-scale

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 97-115, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110408

graph data have gradually become one of the research hotspots in academia and industry currently [10]. However, compared with real-world graphs, many optimizations on graph datasets mainly focus on small datasets. For example, Cora [11, 12], Citeseer [13, 14], Pubmed [15] and Blog [16]. Their specific parameters are shown in Table 1. Most of the evaluations are carried out for small graphs on a single machine, and there are only 2700 to 20000 nodes in the tasks of node classification. There are a few kinds of research on distributed training for large-scale graphs. Even though some authors aim at distributed graph learning, these small datasets are mainly used for training [17-19]. Because models are widely developed on these small datasets, most models cannot be extended to larger graphs. GraphSage [20] and Cluster-GCN [5] provide a method to perform random mini-batch training does not need to read graph features of all nodes into GPU or CPU memory. However, these two mini-batch training methods are still limited in accelerating the training of large-scale graph datasets in a single machine.

A few works have been developed to scale GNNs training on large graph data in the distributed clusters. However, they focus on the scalability other than the accuracy, such as NeuGraph [21] and PCGCN [22] aim to speedup GNNs training. However, there is no discussion about the changes in the accuracy of graph training [23]. Some GNNs frameworks [19, 24] built-in industrial scene adopt distributed mini-batch training. Nevertheless, none of these frameworks uses appropriate graph partitioning to maximize the accuracy of GNNs training. Moreover, for distributed deep learning training on multiple GPUs, it remains a problem that as the GPU number increases, the training accuracy decreases [25].

Table 1. Small datasets used in graph neural networks.

Datasets	**vertex**	**edge**	**feature**	**label**
Core	2708	5429	1433	7
Citeseer	3327	4732	3703	6
Pubmed	19417	44338	500	3
Blog	10400	678300	128	32

We introduce ADGraph, which uses neighbourhood-contained graph partition, multi-layers neighbourhood sampling and overcoming generalization error method in distributed GNNs training. On the one hand, the training of graph models on large-scale datasets can be accelerated through multi-GPU training. On the other hand, the distributed graph learning can still maintain high training accuracy through appropriate graph partition, neighbour sampling and distributed optimization methods. In order to verify our proposed methods, we train GraphSage and GAT model with two large-scale graph datasets (ogbn-products and Reddit) on GPU clusters, which significantly reduces training time. We also use a graph partition that includes neighbourhoods and a multi-layer neighbourhood sampling strategy. Even in a distributed environment, it can achieve the same accuracy as single-GPU training. In summary, our contributions are as follows:

- We use neighbourhood contained graph partition to ensure the completeness of training nodes information in each partition. Then, the mini-batch generated from multi-layers neighbourhood sampling can train nodes accurately.
- We combine a distributed optimization method with graph learning. We use the data synchronization method and linear scaling rule in the distributed training. This further improves the training accuracy of the GNN models.
- We test the GraphSage and GAT model on the ogbn-products and Reddit datasets. Experimental results show that the test accuracy on 32 GPUs is better than the benchmark accuracy using GraphSage and GAT models. The running time is accelerated 24-29 times on different datasets.

The remainder of this paper is organized as follows. Section 2 introduces the background of distributed GNNs training. We present ADGraph training methods in Section 3. Section 4 evaluates and analyses the key technologies to fulfil accurate and large-scale GNNs training. Finally, we conclude the paper in Section 5.

2. Background

2.1. Graph Neural Networks

Graph neural networks (GNNs) are representative work in deep learning. Training neural network on graph data has been widely used because its model accuracy is much higher than that of traditional multi-layer perceptron [26]. GNNs layers generate intermediate embedding by aggregating the information from the in-edge neighbours of the target nodes. After superimposing several GNNs layers, the final embedding is obtained, which integrates the whole receptive field of the target node. Specifically, the graph neural networks iteratively update the node representation according to:

$$h_i^{l+1} = \sigma(W^l \frac{1}{|N_i|} \sum h_j^i) \tag{1}$$

Where h_i^{l+1} is the embedding of node i in the *(l+1)-th* layer. N_i is the node set connect with node i. N_i represents the in-edge number of node i. σ() represents the nonlinear activation function. W^l is the learnable parameters of layer l. GNNs first aggregate all values from the in-edge neighbours of each node to obtain new values for these nodes. After that, GNNs propagate this new value to target nodes throughout-edges. After l times of such aggregation and propagation, the calculation of GNNs is completed.

GraphSage [20] only needs to aggregate data sampled from the graph, without considering other nodes. GraphSage provides different ways to aggregate information of adjacent nodes. In the average version of GraphSage, the update formula of node embedding is Equation (2). GraphSage can form a mini-batch by sampling a specific size of neighbour nodes and does not need to get the adjacency matrix of the whole graph. This is very useful in training large-scale datasets.

$$h_i^{l+1} = \sigma(W^l Concat(h_i^l, \frac{1}{|N_i|} \sum_{j \in N_j} h_j^l)) \tag{2}$$

Graph Attention Network (GAT) [27] aggregates features of neighbouring vertices to the central node and learns new nodes features by using local stationary on Graph. GAT makes use of the attention coefficient and introduces anisotropy into the neighbourhood aggregation function. This network adopts a multi-head structure to increase the learning ability. Equation (3) is the updated formula of GAT, where W_l^k are k linear projections heads, and e is the attention coefficient of each head. GAT has stronger learning ability because the model can better capture correlation between node features [28].

$$h_i^{l+1} = Concat_{k=1}^{K}(\sigma(\sum_{j \in N_i} e_{ij}^{k,l} W^{k,l} h_j^l)) \tag{3}$$

2.2. Parameter Sever architecture and data parallelism

Parameter server (PS) is one of the commonly used frameworks for distributed deep learning [29]. PS aims to improve the training efficiency of big data and large models while maintaining accuracy. There are two parts in Parameter Server architecture: parameter server (PS) and worker. As shown in Figure 1, PS maintains a global shared parameter and updates all parameters together. Each worker is responsible for handling local training tasks, obtaining the latest model parameters from PS nodes, and sending model gradients generated by the local worker to PS [30].

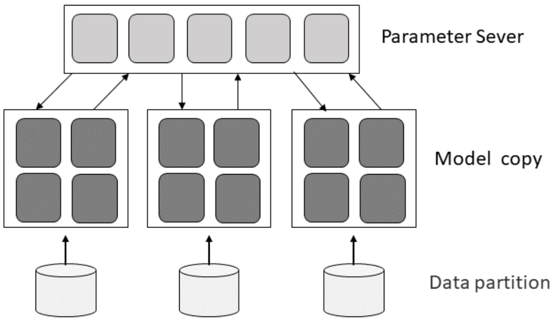


Figure 1. The architecture of the parameter server.

The mainstream method of distributed deep learning is data parallelism, which has higher training efficiency [31]. As shown in the Figure 2, In data parallel method, the whole dataset is divided into multiple machines. Each machine has a local copy of the model and updates the local model with the assigned data [32]. In the synchronous update, gradients of different batches are calculated at each worker. Gradients are averaged across machines to apply consistent updates to model copies in each worker [33]. This synchronization method is widely used in large-scale systems.

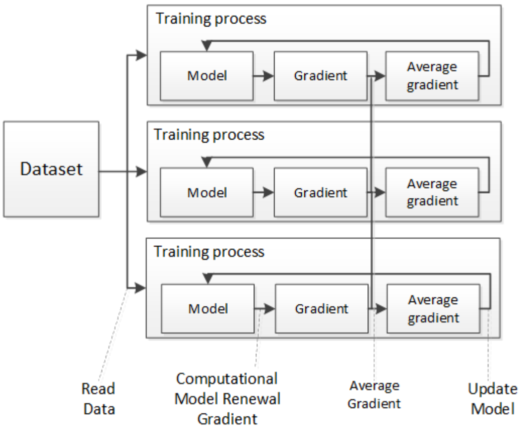


Figure 2. The process of data parallelism.

2.3. Large mini-batch Stochastic Gradient Descent

When mini-batch is used for GNNs training on multi-GPUs, linear scaling rule can reduce the training error caused by large mini-batch. After the mini-batch training is completed, stochastic gradient descent (SGD) [34] performs the following update:

$$w_{t+1} = w_t - \eta \frac{1}{n} \sum_{x \in \beta} \nabla l(x, w_t) \tag{4}$$

Here β is a mini-batch sampled from the neighbourhood of target nodes in partitioned graphs. t is the update times and η is the learning rate. $n=|\beta|$ is the mini-batch size. According to Equation (4), when learning rate η and batchsize are n, after k iterations of SGD, Equation (5) can be obtained:

$$w_{t+k} = w_t - \eta \frac{1}{n} \sum_{j<k} \sum_{x \in \beta_j} \nabla l(x, w_{t+j}) \tag{5}$$

On the other hand, using mini-batch of size kn and learning rate $\hat{\eta}$ to update once can get:

$$\hat{w}_{t+1} = w_t - \hat{\eta} \frac{1}{kn} \sum_{x \in \beta_j} \nabla l(x, w_t) \tag{6}$$

According to Equations (5), (6), it can be seen that the results of updating k times with a small-batch and updating once with a large-batch are different. Therefore, in order to keep the weights unchanged in both cases after SGD update. Set learning rate $\hat{\eta}=k\eta$ when updating large batches, and the updated results w_{t+k} and $\hat{w}_{t+1}$ will be approximately the same [35]. That is, linear scaling rule can reduce the distributed GNNs training error on Multiple GPUs.

3. ADGRAPH TRAINING METHODS

The difference between distributed training of graph neural networks and traditional distributed training lies in graph partition and mini-batch sampling. The overall structure of distributed training with four machines is shown in Figure 3. Specifically, there are three processes in graph distributed training: Graph Server, Sampler and Trainer. The Graph Server process needs to run on each machine to store the graph partitions (including graph structure, nodes features and nodes labels). Sampler process samples nodes from Graph Server and generate mini-batch required by trainer process. It can be noticed that a sampler process can obtain data from multiple Graph Servers. Trainer process can only obtain mini-batch from the sampler on its local machine. Then, the trainer calls the all-reduce primitive to update model parameters.

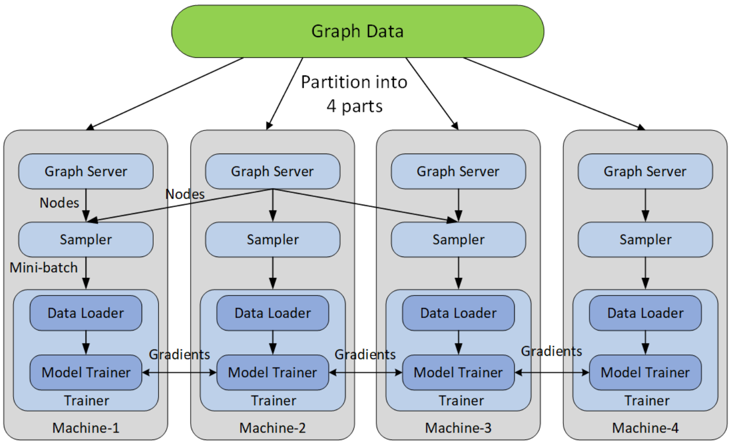


Figure 3 . Distributed graph neural networks training process.

The training process starts with the partition of graph data and then carries out mini-batch training. The steps performed in each mini-batch training process include: (1) Sampling the neighbourhood of target nodes from a local partition to generate mini-batch. (2) Obtaining features and labels involved in mini-batch from the global graph data. (3) Performing forward and backward propagation on the features to calculate the gradients of each layer. (4) Trainer process uses all-reduce to accumulate gradients. Then the trainer applies averaged gradients to update parameters of the model.

3.1. Graph partition containing neighbourhood information

Graph partitioning is the first step in distributed graph learning. Firstly, nodes are assigned to partitions using METIS [36] or random graph partitioning algorithm. Then the partitioned graph structure is constructed according to the result of node allocation. Finally, node features are segmented according to the partition results. We partition the graph structure, node features and labels, and distribute them on cluster machines in distributed training. There are two potential problems after graph partition: (1) Deleting some edges between nodes may affect performance. (2) Graph clustering algorithm (METIS partition) tends to cluster similar nodes together, resulting in the distribution of node categories different from the original dataset. Therefore, estimation of the gradient is biased when performing SGD updates.

We solve these two problems by two methods, namely partition graph with the intact neighbourhood of target nodes and increasing batch label entropy.

3.1.1. Keeping the neighbourhood information of the target nodes intact

According to Equation 1, nodes in the *l-hop* neighbourhood of the target nodes contain enough and necessary information for training the *l-layer* GNNs model. Therefore, in *l-layers* GNNs model, the embedding of target node only depends on its *l-hop* neighbourhood, rather than the entire graph.

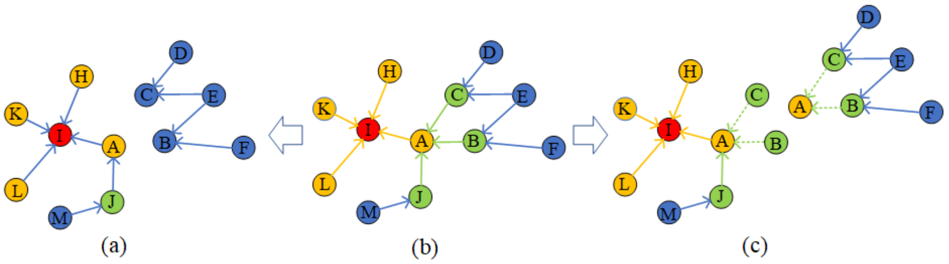


Figure 4. Example of graph partition(b) is the original graph, (a) is the direct partition of the graph without keeping the neighbourhood of I, the partition in (c) repeatedly stores the *2-hop* neighbour B and C.

After graph partition, nodes at the subgraph boundary are stored in two adjacent subgraphs. For completeness and efficiency, each partition contains not only nodes and edges belonging to itself, but also *l-hop* neighbourhoods of nodes in other partitions. As shown in Figure 4, (b) is the original graph. It is assumed that the GNNs model has two layers, thus, generating the embedding of node *I* needs the complete information of *1-hop* neighbour *HKL* and *2-hop* neighbour *BCJ*. Graph (a) is the result of partitioning Graph (b). Due to the lack of connection between *AB* and *AC*, the embedding of *I* on the partitioned subgraph will lack the information of *B* and *C*, resulting in inaccurate generating the embedding of *I*. By contraries, Graph (c) keeps the neighbourhood of *I*. Although subgraph becomes smaller after graph partition, the information to generate the embedding of *I* is intact. Therefore, this partition method ensures the accuracy of model training.

The *l-hop* neighbourhood of target nodes can provide enough information for the target nodes in the partition which avoids missing connection information after partition. Even after graph partition, the data in each partition is intact. This ensures the convergence accuracy when training *l-layer* GNNs, and it can achieve the same performance as that without partition.

3.1.2. Increase batch label entropy

Distributed graph learning is the training of graph neural networks by using the graph data to predict and simulate unknown large-scale graph data. Therefore, each mini-batch should be generally representative. We should generalize rules from existing graph data to make decisions on unknown graph data. If the training data is not representative, the rules will be poorly summarized, and significant deviations will be generated in the inference process on unknown graph data.

Vanilla Cluster-GCN [5] shows that METIS method partitions the graph into a large number of partitions, and nodes in the partitions tend to specific categories. Generating mini-batch from these partitions may lead to lack of representativeness of mini-batches. In this case, label entropy of most mini-batches is smaller than that of random partition. This indicates that the label distribution of mini-batches is biased towards some specific labels. This will increase the variance between different batch and may affect the convergence of SGD.

In order to avoid label deviation, we do not partition the graph into a large number subgraph. We set the number of partitions consistent with the number of machines. This can reduce network communication when the neighbourhood expands and make each batch have various labels that will not bias towards specific labels. In Figure 5, ogbn-products dataset is divided into eight partitions by METIS and random methods to show the example of label distribution. We calculate entropy according to the label distribution of each batch. It can be seen that the label entropy of the mini-batches from METIS partition and random partition is similar. Therefore, the convergence of SGD will not be affected by METIS partition.

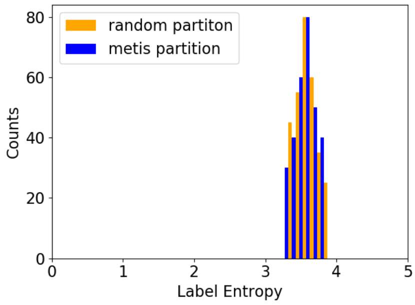


Figure 5. The label entropy of each batch when ogbn-products is partition into 8 parts.

3.2. Multi-layer neighbourhood sampling

Single-layer subgraph can only perform forward propagation once. Therefore, multi-layer neighbourhood sampling is required to generate the mini-batch when compute multi-layer GNNs. However, the degrees of nodes in a large graph are generally vast, multi-layer subgraph sampling will cause exponential expansion of neighbour nodes which consumes lots of memory. We use partial neighbourhood sampling for multi-layer GNN. For different *hop* neighbourhoods of the target nodes, a fixed number of neighbours are sampled. Due to the neighbours of the target nodes are randomly sampled, all neighbours will participate in training after multiple epochs. The clustering coefficient is a popular measure of how clustered a node's local neighbourhood is. GraphSage [20] has proved that partial neighbourhoods sampling is capable of approximating clustering coefficients to an arbitrary degree of precision, even when the node feature inputs are sampled from an absolutely continuous random distribution.

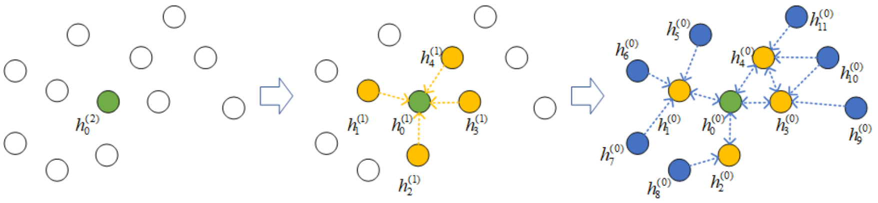


Figure 6. Multi-layer neighbourhood sampling to generate a mini-batch.

Figure 6 shows the process of multi-layer neighbourhood sampling: 1) For each gradient descent step, we select some target nodes to calculate their final representations at *l-th* layer. 2) Then, obtain part of *1-hop* neighbours of the target nodes at *l-1* layer, and obtain part of the *2-hop* neighbours of the target nodes at *l-2* layer. 3) This process continues till the input layer. The iteratively constructing dependency graph of multi-layer neighbourhood sampling generates a mini-batch. The forward calculation process is the opposite, which is calculated from (c) to (a).

Since the sampled mini-batch contains *l-hop* neighbourhood of the target nodes *v*, the information is intact when perform *l-layers* GNNs to calculate the embedding of target nodes. The embedding of the target nodes in *l-th* layer is calculated from the equation as follows:

$$h^{(l)} = A_{l-1}\sigma(\ldots A_1\sigma(A_0 X W^0)W^1\ldots)\, W^{l-1} \qquad (7)$$

Where A is the subgraph represented in adjacency matrix in each mini-batch, X is the feature matrix, W is the weight parameter of each layer, and its loss function can be expressed as:

$$L = \frac{1}{|\mathrm{v}_A|}\sum loss(y_i, h_i^l) \tag{8}$$

In each step, we first sample the mini-batch of target nodes v, then perform SGD to update the parameters based on the gradient L. The gradient calculation and update only require the adjacency matrix A and node features X of the current mini-batch, ensuring the accuracy of the embedding in the last layer. Since the *k-hop* neighbourhood contains sufficient and necessary information for training GNNs model, trainers become independent of each other. They train mini-batch without additional communication with other trainers. Therefore, training of the GNNs model is similar to that of conventional deep learning.

3.3. Overcoming generalization errors of distributed GNNs training

In graph deep learning, it is stated that model trained with large batchsize is often inferior to that with small batchsize [37]. [38] found that when the training accuracy is consistent, the generalization performance of a model trained with large batchsize will be significantly lower than model trained with small batchsize.

Unlike single GPU training, data is sampled from the same dataset in each step. The batchsize is small compared to the entire data set, which can ensure sufficient parameter updates. However, in distributed GNNs training, the subset on each machine becomes smaller after graph partition. Each trainer samples mini-batch from the local partition. A larger batchsize will cause the parameter update insufficient, resulting in larger generalization errors. After graph partition, the batchsize should be reduced, which can increase the amount of model parameter updates. In this way, an accurate model can be trained efficiently under distributed training.

Although the generalization performance of the model can be guaranteed by increasing the amounts of updates, this will affect the benefits of distributed training. In order to maintain the accuracy of training and generalization while training on multiple machines, the linear learning rate scaling rule has a particularly important role in distributed learning. Because this allows data parallelism to be extended to more GPUs, it also improves the distributed training accuracy. Facebook large-scale training [25] has proved that small batch and large batch SGD updates not only get the same final precision model, but also match the training curves very well. We prove that the linear scaling rule is effective in the large-scale real-world graphs through experiments.

4. Evaluation and Analysis

4.1. Experiment Setup

In this section, focusing on the task of node classification, we will evaluate the efficiency and performance of the proposed methods through experiments on several GNNs models and datasets.

GNNs models. We use the following two representative GNNs models in the experiment: GraphSage and GAT. Related concepts have been introduced in the background. Because they adopt different aggregation method of neighbours in the graph, we choose these models in experiment. Table 2 shows the default settings of parameters when training the two models.

Table 2. Parameter settings for GNN models in the experiment.

Models	layers	hiddens	Sampled neighbours	batchsize	Epochs
GraphSage	3	512	(10,10,10)	250	30
GAT	3	128	(10,10,10)	250	30

Table 3. Large-scale graph data statistics.

Datasets	vertex	edge	feature	label	Avg degree
ogbn-products	2,449,029	61,859,140	100	47	50
Reddit	232,965	1,606,919	602	41	100

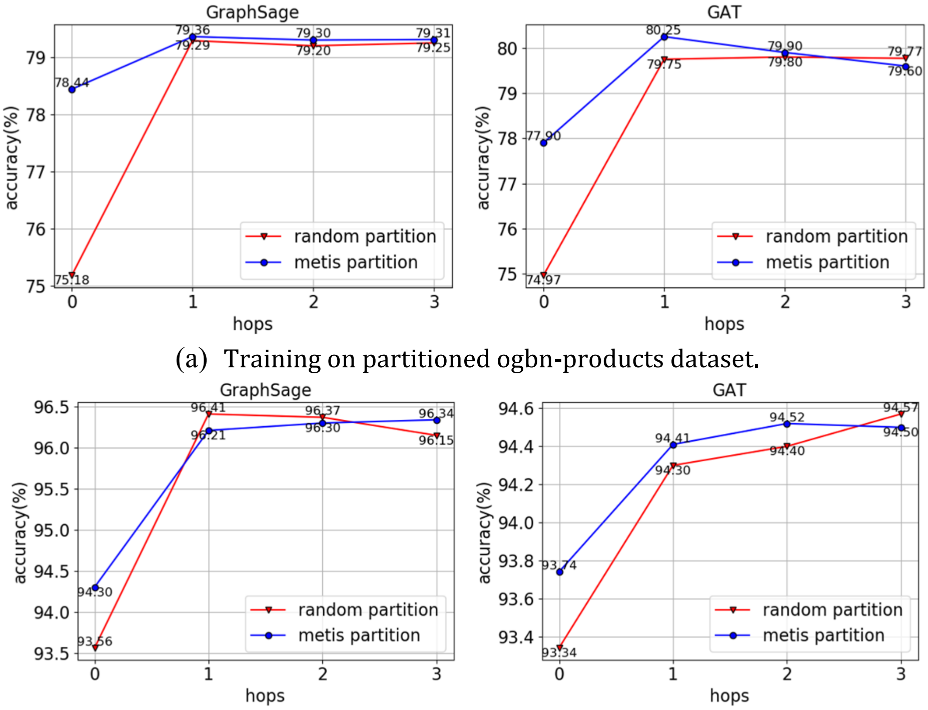
Datasets. The real-world graph datasets used in the experiment are listed in Table 1. The Reddit [20] dataset is formed by Reddit online discussion forum, and ogbn-product [9] comes from Amazon product co-purchasing network. The *feature* column in table 1 represents the feature dimensions of each node, and the *label* column indicates the number of label categories.

Experimental environment. We evaluated experimental results on GPU cluster and used up to 8 machines on the cluster for training. Each machine has four Nvidia Tesla GPU and two Intel Xeon CPU, and machines are connected through InfiniBand ConnectX FDR 56GB/s internet. Operating system version used is Redhat4.8, and libraries of CUDA10.0 are used. Our experiments are carried out on Pytorch, a deep learning framework, and Deep graph Library (DGL) [39]. DGL is a Python package that interfaces between tensor-oriented frameworks (such as Pytorch and MXNet) and graph structure data, which makes it easy to implement GNNs.

4.2. Comparison of Partition Methods

We use random and METIS [36] methods to partition the datasets, and compare convergence accuracy and running time of the two methods. By increasing neighbour layers around the target nodes in each partition (hop=k means expanding the neighbourhood of the target nodes to the k-th layers), the effect of neighbourhood partition with different layers is verified.

Accuracy. The convergence accuracy of the experiment is shown in Figure 7. It can be seen that when hop=0, the performance gap between these two methods is the largest. Figure 9 shows the number of edges cut by the random and METIS partition when the hop=0. It can be noticed that the edges cut by METIS partition is much less than random partition. This is due to random partition randomly assigned nodes into partitions, which will increase the randomness of edge segmentation and make many neighbours of nodes disappear. However, METIS divides into clusters, and nodes with many connections will be partitioned into the same partition. The segmentation is mainly between clusters, which will significantly reduce cut edges. Therefore, under the METIS partition, similar nodes can be clustered together to capture the clustering and graph structure better. METIS partition preserves the graph structure better, and results obtained during information aggregation are more accurate than random partition.



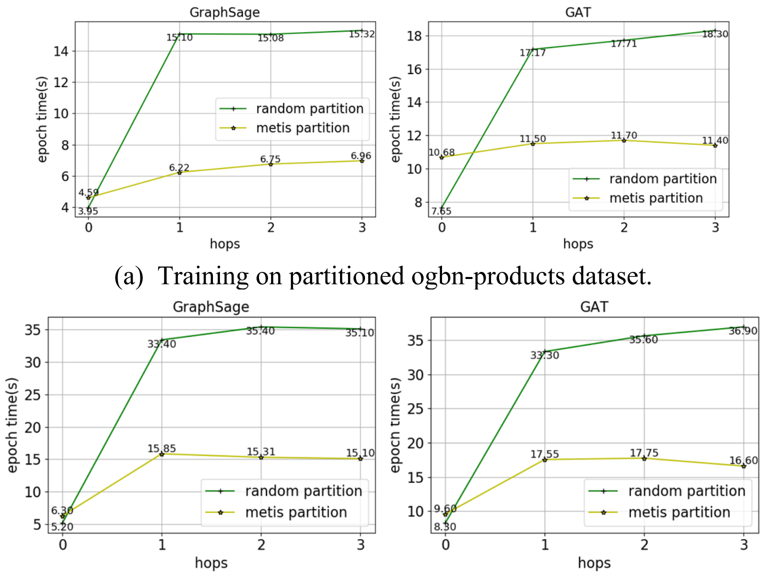
(a) Training on partitioned ogbn-products dataset.

(b) Training on partitioned Reddit dataset.

Figure 7. Convergence accuracy after partition datasets with different *hops*.

It can be seen from Figure 7 that when hop goes from 0 to 1, the convergence accuracy will increase significantly. The reason can be explained by Figure 10. Because of the strong connectivity among nodes in the graph, it will almost extend to the whole graph when expanding the 1-hop neighbourhood of the target nodes. When hops increase from 1 to 3, the convergence accuracy remains unchanged. Because the number of nodes in the extended neighbours is only slightly increased. Furthermore, when hops are greater than or equal to 1, the convergence accuracy of random partition and METIS partition is not much different. Because in these two methods, the number of nodes in the partitions is very similar, both can effectively expand the structure to the whole graph.

Run time. Figure 8 shows that the epoch time of the METIS partition is similar to random partition when *hop=0*. However, the epoch time of METIS is much shorter than that of random partition when *hops>0*. Although random and METIS partition can be extended to the full graph, nodes connectivity in local machines are different. METIS partition can generate more closely connected clusters, while nodes in the partitions generated by a random partition are randomly connected. Therefore, the graph data read during training is different. Random partition requires more communication with other machines for nodes data. Cross-machine communication takes more time, which leads to longer epoch time for random partition. This shows that graph clustering is significant, and the partitions of graph should not be randomly generated.



(a) Training on partitioned ogbn-products dataset.

(a) Training on partitioned Reddit dataset.

Figure 8. The epoch time after partition datasets with different *hops*.

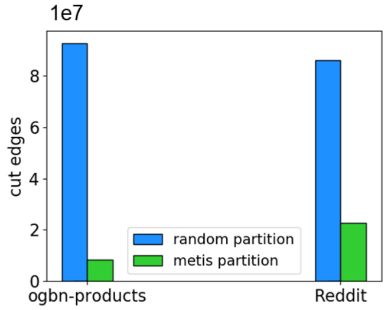


Figure 9. The number of cut edges after random and METIS partition.

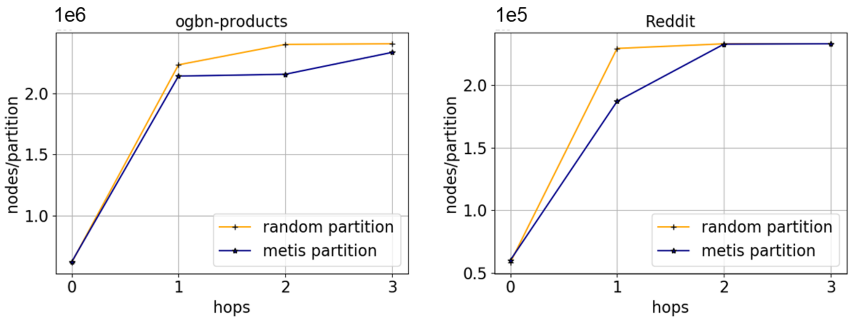


Figure 10. The average number of nodes in each partition as *hops* increase.

4.3. Multi-Layer Neighbourhood Sampling

When verifying the multi-layer sampling method, we fix the sampled neighbours for each node. We test convergence accuracy by increasing the sampled layers. Figure 11 shows the curves of training accuracy on 2 datasets, from which we can observe that the increase of layers can improve the training accuracy, but the improvement in accuracy after the third layer is not obvious. It is worth noting that the training accuracy of GAT failed to converge within 30 epochs and get a dramatic loss of accuracy when 4 layers are used. A possible reason is that the structure of deeper GAT is more complex and has more parameters, resulting in the optimization for deeper GAT becomes more difficult.

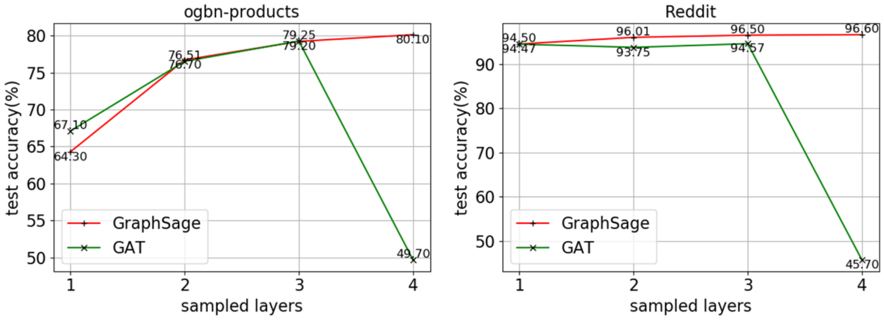


Figure 11: The model accuracy changes with the increasing sampled layers.

We test convergence accuracy and run time by changing the sampled neighbours from 2-32 (*sampled neighbours* = *k* means sampling *k* neighbour nodes of the current nodes). It can be seen from Figure 12 that the sampled neighbours have a great influence on the training performance (For ogbn-products training GAT, the memory will be exceeded when *sampled neighbours=32*, so we set sampled neighbours to 24). When the number of sampled neighbours increases from 2 to 16, the convergence accuracy can be greatly improved. However, when the number of sampled neighbours exceeds 16, there is almost no change in accuracy. This is because the neighbours are randomly sampled, and all nodes in the neighbourhood will participate in the training after several epochs. When *sampled neighbours=16*, the neighbourhood of target nodes can be captured effectively. In this situation, the training accuracy is the highest, and the accuracy is no more improved by increasing sampled neighbours.

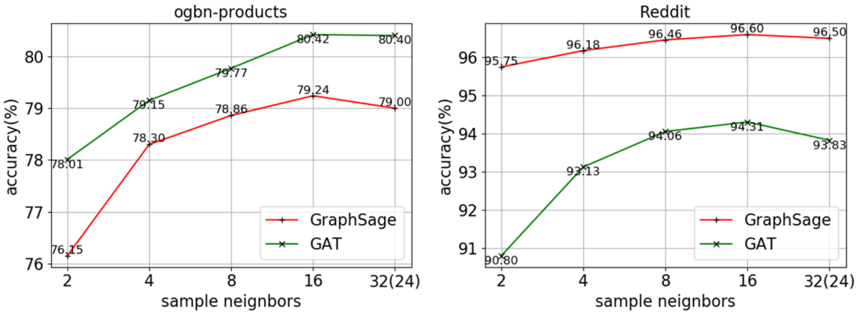


Figure 12. The model accuracy changes with the increasing sampled neighbours.

Figure 13 shows the training epoch time changes as the sampled neighbours increasing. We can notice that when the sampled neighbour doubles, the epoch time increases significantly. This is due to the increment of the sampling time and the model computation time caused by the adding of sampled neighbours. Balancing the performance gains and time consumption, the appropriate sampled neighbours can be chosen. In this way, ADGraph can optimize the convergence accuracy and guarantee training efficiency.

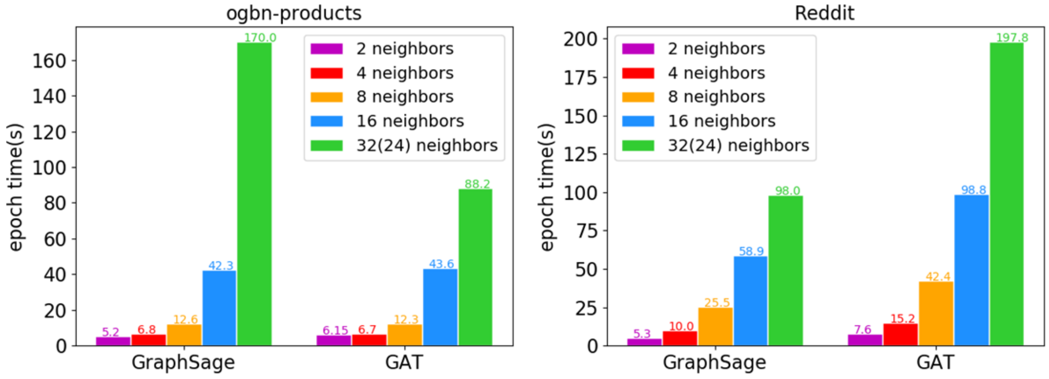


Figure 13. The increment of the run time with the adding sampled neighbours.

4.4. Change Batch Size To Overcome Generalization Error

In distributed GNNs training, the choice of batchsize is significant. In this experiment, we train the GraphSage model on ogbn-products and Reddit datasets with 1 to 32 GPUs. The batchsize on each GPU varies from 10 to 1000. Table 4 and Table 5 list the results. When there are few GPUs, and the batchsize is 10, convergence accuracy on the ogbn-products and Reddit datasets stops at 26.94% and 14.83%, respectively. It can be noticed that small batchsize in the distributed graph model training cannot converge. When there are 1 and 2 GPUs, higher model accuracy can be obtained with 500 and 1000 batchsize, because the large number of training nodes can guarantee enough updates. As the number of machines increases, the number of training nodes on each machine decreases after the graph partition. If larger batchsize is applied, parameter updates on each machine will be reduced, resulting in poor convergence accuracy. Therefore, after graph partitioned, ADGraph can efficiently obtain an accurate model on 8 machines by reducing batchsize (*batchsize = 250*) and increasing the parameters updates of the model.

Table 4. The test accuracy on ogbn-products with the increment of batchsize and GPUs.

Batchsize/GPU	**1**	**2**	**4**	**8**	**16**	**32**
10	26.94	26.94	26.98	30.84	55.83	56.74
50	73.04	76.66	78.11	78.31	76.94	76.36
100	75.84	78.51	79.31	78.19	78.87	78.68
250	79.2	79.34	79.45	79.36	79.36	79.13
500	79.4	79.35	79.02	78.61	77.86	74.52
1000	79.3	79.25	78.61	77.48	76.65	72.8

Table 5. The test accuracy on Reddit with the increment of batchsize and GPUs.

Batchsize/GPU	**1**	**2**	**4**	**8**	**16**	**32**
10	14.83	14.83	14.84	15.04	15.2	15.22
50	14.83	92.6	95.61	95.12	95.23	96.26
100	94.96	95.79	96.4	96.27	96.12	96.32
250	96.47	96.65	96.6	96.56	96.32	96.55
500	96.69	96.68	96.64	96.44	96.14	93.34
1000	96.67	96.54	96.36	96.13	95.45	92.39

4.5. Apply Learning Rate Scaling Rule

Table 6 and Table 7 show the effect of using linear learning rate rule in distributed graph training. GPU=1 is the result of model training with unpartitioned datasets on DGL [39]. When *GPU=32*, ADGraph use the linear scaled learning rate to compare with the fixed learning rate. It can be seen that when applying a fixed learning rate, the model convergence accuracy is poor on multiple GPUs. When training on 32 GPUs, it can achieve similar accuracy to that of a single GPU training with the linear scaled learning rate. Experiments prove that the linear learning rate is effective in real-world graphs.

It is worth noting that when training the GAT model on the ogbn-products dataset and the GraphSage model on the Reddit dataset, the convergence accuracy on 32 GPUs is even higher than that of a single GPU. This proves that distributed graph training of ADGraph has the same convergence performance as single GPU training on DGL.

Table 6. The performance gains of linear scaling rule on ogbn-products.

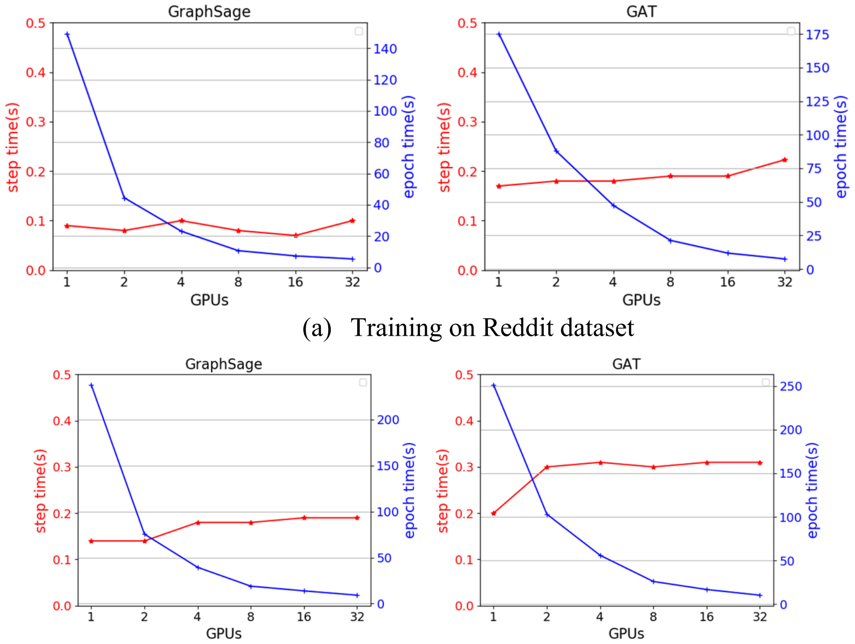
Models	Batchsize*GPUs	Learning rate	Accuracy (%)
GraphSage	250*1	0.003	79.2
	250*32	0.003	74.58
	250*32	0.096	79.13
GAT	250*1	0.0005	79.25
	250*32	0.0005	77.13
	250*32	0.016	80.18

Table 7. The performance gains of linear scaling rule on Reddit.

Models	Batchsize*GPUs	Learning rate	Accuracy (%)
GraphSage	250*1	0.0015	96.47
	250*32	0.0015	94.95
	250*32	0.048	96.55
GAT	250*1	0.0005	94.57
	250*32	0.0005	91.5
	250*32	0.016	94.41

4.6. Run Time

Figure 14 shows the curve of epoch time and step time as the GPU increases. The red curve is step time when GPUs changes from 1 to 32 (mini-batch size varies from 250 to 8000). The curve is relatively stable, and the increase in the number of GPUs did not significantly increase step time. The blue curve shows the reduction of each epoch time as the GPUs increases. The overall epoch time is continuously decreasing. In general, the epoch time of 32 GPUs is 24-29 times faster than the epoch time of a single GPU, which can significantly improve distributed graph training efficiency. The maximum scalability efficiency of ADGraph can reach 91%, which is higher than 83% of DistDGL [10] (The result was shown in the experiment of DistDGL).



(a) Training on Reddit dataset

(b) Training on ogbn-products dataset

Figure 14. The curve of epoch time and step time as the GPUs increase.

5. Conclusion

In this paper, we present ADGraph for accurate and distributed GNNs training on large graphs. We first used a graph partition method that contains the neighbourhood of the training nodes, and investigated the accuracy and time efficiency of the model training as the number of neighbourhood *hops* increases. Then the complete neighbourhood information of target nodes is obtained through multi-layer neighbourhood sampling. We also analyse the accuracy and runtime performance of graph training, with different l-hop settings. We found that the training time will increase dramatically as the increment of sampled neighbours and sampled layers, but the accuracy of the model not always increases. Then, we explored the influence of batchsize and the number of GPUs on the training of distributed GNNs. The results show that training on the graph partitions needs to reduce the batchsize appropriately. Finally, the linear scaling rule is applied to further improve the training accuracy. The distributed training accuracy can exceed the benchmark accuracy of GraphSage and GAT on DGL. The accuracy of training on 32 GPUs is the same as that of single GPU training, and there is a speedup of 24-29 times.

We have also noticed that there are some shortcomings in the proposed methods. Although the graph partition maintains the integrity of the neighbourhood information of the target nodes, the expanded neighbourhood range is too large, causing the graph partition to lose its meaning. We also found that due to the dependencies between the partitions, the sampling process may communicate with other machines, which will affect the sampling speed. Later, we will study more accurate graph clustering in order to achieve a more reasonable graph partition. Another problem is that the data transfer time from the sampler process to the trainer process often takes up most of the time, resulting in low utilization of computing resources. We are trying data prefetching and caching technology to speed up training and improve the scalability efficiency of distributed training.

ACKNOWLEDGEMENTS

We sincerely thank the anonymous reviewers for their insightful comments and suggestions. This work was supported in part by National Nature Science Foundation of China.

REFERENCES

[1] Z. Guo and H. Wang, "A deep graph neural network-based mechanism for social recommendations," *IEEE Transactions on Industrial Informatics,* vol. 17, no. 4, pp. 2776-2783, 2020.

[2] Z. Xinyi and L. Chen, "Capsule graph neural network," in *International conference on learning representations*, 2018.

[3] D. Szklarczyk *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research,* vol. 47, no. D1, pp. D607-D613, 2019.

[4] W. Jin, R. Barzilay, and T. Jaakkola, "Junction Tree Variational Autoencoder for Molecular Graph Generation," 2018.

[5] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 257-266.

[6] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D Knowledge Graph Embeddings," 2017.

[7] J. Zhou *et al.*, "Graph Neural Networks: A Review of Methods and Applications," 2018.

[8] Z. Jia, Y. Kwon, G. Shipman, P. Mccormick, and A. Aiken, "A distributed multi-GPU system for fast graph processing," *Proceedings of the VLDB Endowment,* vol. 11, no. 3, pp. 297-310, 2017.

[9] W. Hu *et al.*, "Open Graph Benchmark: Datasets for Machine Learning on Graphs," 2020.

[10] D. Zheng, C. Ma, M. Wang, J. Zhou, and G. Karypis, "DistDGL: Distributed Graph Neural Network Training for Billion-Scale Graphs," 2020.

[11] W. Huang, T. Zhang, Y. Rong, and J. Huang, "Adaptive sampling towards fast graph representation learning," *arXiv preprint arXiv:1809.05343,* 2018.

[12] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International Conference on Machine Learning*, 2020: PMLR, pp. 1725-1735.

[13] F. Wu, T. Zhang, A. H. D. Souza, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying Graph Convolutional Networks," 2019.

[14] J. Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," *arXiv preprint arXiv:1810.05997,* 2018.

[15] J. Du, S. Zhang, G. Wu, J. M. Moura, and S. Kar, "Topology adaptive graph convolutional networks," *arXiv preprint arXiv:1710.10370,* 2017.

[16] H. Dai, Z. Kozareva, B. Dai, A. Smola, and L. Song, "Learning steady-states of iterative algorithms over graphs," in *International conference on machine learning*, 2018: PMLR, pp. 1106-1114.

[17] H. Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1416-1424.

[18] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," *arXiv preprint arXiv:1903.02428,* 2019.

[19] D. Zhang *et al.*, "AGL: a Scalable System for Industrial-purpose Graph Machine Learning," 2020.

[20] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," 2017.

[21] L. Ma *et al.*, "Neugraph: parallel deep neural network computation on large graphs," in *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, 2019, pp. 443-458.

[22] C. Tian, L. Ma, Z. Yang, and Y. Dai, "PCGCN: Partition-Centric Processing for Accelerating Graph Convolutional Network," in *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2020.

[23] W. Wei, Y. Wang, P. Gao, S. Sun, and D. Yu, "A Distributed Multi-GPU System for Large-Scale Node Embedding at Tencent," *arXiv preprint arXiv:2005.13789,* 2020.

[24] J. Li, J. Zhu, and B. Zhang, "Discriminative Deep Random Walk for Network Classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[25] P. Goyal *et al.*, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677,* 2017.

[26] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems,* 2019.

[27] P. Velikovi, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," 2017.

[28] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 793-803.

[29] M. Li *et al.*, "Scaling distributed machine learning with the parameter server," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 583-598.

[30] M. Li, D. G. Andersen, A. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," *Advances in neural information processing systems,* vol. 1, pp. 19-27, 2014.

[31] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315-323.

[32] D. Li, Z. Lai, K. Ge, Y. Zhang, and H. Wang, "HPDL: Towards a General Framework for High-performance Distributed Deep Learning," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019.

[33] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal Distributed Online Prediction Using Mini-Batches," *Journal of Machine Learning Research,* vol. 13, no. 1, 2012.

[34] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics,* pp. 400-407, 1951.

[35] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489,* 2017.

[36] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal on scientific Computing,* vol. 20, no. 1, pp. 359-392, 1998.

[37] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997,* 2014.

[38] T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes," *arXiv preprint arXiv:1711.04325,* 2017.

[39] M. Wang *et al.*, "Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks," 2019.

AUTHORS

Lizhi Zhang, born in 1996, MS candidate, his research interests include distributed machine learning and graph neural network.

E-mail: zhanglizhi15@nudt.edu.cn

Zhiquan Lai, born in 1986, assistant professor, his research interests include distributed machine learning and highly performance system software.

E-mail: laizhiquan@nudt.edu.cn

Zhejiang Ran, born in 1997, MS candidate, his research interests include distributed machine learning and graph neural network.

E-mail: ranzhejiang@163.com

Feng Liu, born in 1977, associate Professor, his research interest includes big data and distributed computing.

E-mail: liufeng@nudt.edu.cn

AUTHOR INDEX