AN INNOVATIVE RESEARCH FRAMEWORK ON INTELLIGENT TEXT DATA CLASSIFICATION SYSTEM USING GENETIC ALGORITHM

Dr. V V R Maheswara Rao[1], N Silpa[2] and Dr.Gadiraju Mahesh[3]

[1,2]Department of CSE, Shri Vishnu Engineering College for Women, Andhra Pradesh, India
[3]Department of CSE, S.R.K.R. Engineering College, Andhra Pradesh, India

ABSTRACT

Recent years have witnessed an astronomical growth in the amount of textual information available both on the web and institutional wise document repositories. As a result, text mining has become extremely prevalent and processing of textual information from such repositories got the focus of the current age researchers. Indeed, in the researcher front of text analysis, there are numerous cutting edge applications are available for text mining. More specifically, the classification oriented text mining has been gaining more attention as it concentrates measures like coverage and accuracy. Along with the huge volume of data, the aspirations of the user are growing far higher than the human capacity, thus, an automated and competitive intelligent systems are essential for reliable text analysis.

Towards this, the authors in the present paper propose an Intelligent Text Data Classification System (ITDCS) which is designed in the light of biological nature of genetic approach and able to acquire computational intelligence accurately. Initially, ITDCS focusses on preparing structured data from the huge volume of unstructured data with its procedural steps and filter methods. Subsequently, it emphasises on classifying the text data into labelled classes using KNN classification based on the selection of best features derived by genetic algorithm. In this process, it specially concentrates on adding the power of intelligence to the classifier using together with the biological parts namely, encoding strategy, fitness function and operators of genetic algorithm. The integration of all biological components of genetic algorithm in ITDCS significantly improves the accuracy and reduces the misclassification rate in classifying the text data.

KEYWORDS

Text Data, Classification, Genetic Approach, Learning Algorithm, Text Mining, KNN Classification.

1. INTRODUCTION

In the light of rapid advances in data storage, the presence of structured, semi-structured, unstructured, non-scalable and complex nature of data, evolved as a great text data repository and in-turn has motivated the advances in text mining along with soft computing paradigm, and this endorsed as a potential research area for the present researchers. Nowadays,text mining has attracted more attention from academia & industry and a great amount of progresses have been achieved in many applications also. Although much work has been done in text mining and a great amount of achievement has been made so far. However, still remain many open research problems to be solved in this area due to the fact of the complexity of text data, the diversity of various applications and the increased demands of the users in terms of efficiency and effectiveness.

DOI: 10.5121/ijaia.2016.7605

Text mining is also called as Knowledge Discovery from Text Data (KDTD)[27], is a process of extracting potential, previously hidden knowledge from large amount of text documents. Text Pre-processing, Text Transformation, Attribute Selection, Text Mining, Pattern Evolution and knowledge representation are the key phases of text mining as depicted in figure 1.

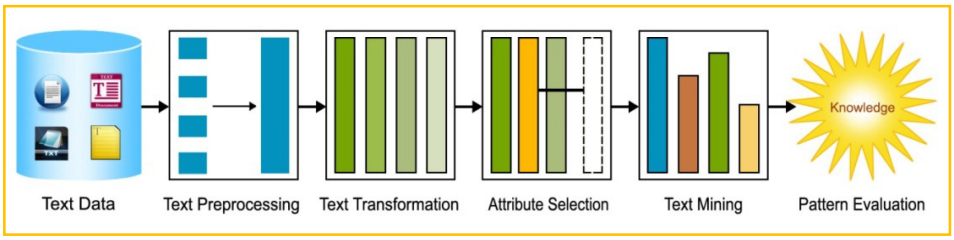

Figure 1.Phases of Text Mining

Text mining refers gaining knowledge from virtually infinite volumes of textual data, available in digital format such as transaction data in e-commerce applications or genetic expressions in bioinformatics research domains. At high level, the text mining techniques broadly categorised into statistical methods and linguistic methods. Statistical methods generally build on statistical or probabilistic framework and rely on mathematical representation of text, so called bag of words matrix. Linguistic methods, basically, developed on meaning semantics and natural language processing techniques. There are several methods used to structure the text documents. Principally, these methods are classified into supervised methods which perform the text mining task by assigning a given keyword to the documents and unsupervised methods which automatically group the similar text documents.

Text classification is a supervised technique which is used to build the classifier to classify the new text documents. In this approach, pre-defined class labels are assigned to the text documents. The aim is to train the classifier on the basis of known text documents and then new text documents are classified automatically. However, the classification is often used in the fields of information retrieval and information extraction. In overall process of classification, the performance and training of any text classification technique is highly depends on the selection of best features from possible set of features. The simple usage of feature selection process in conventional classification techniques proved to be inefficientas the dimensionality of the original feature set is high, drawn from the data preparation stage. Thus, it creates a need to deployintelligent learning feature selection techniques in order to enhance the performance of classification in the era of text data mining.

The computational intelligence models are relying on heuristic algorithms such as evolutionary computation, fuzzy systems and artificial neural networks. Basically these models acquire the intelligence by applying the combination of learning, adaption and evaluation techniques. The intelligent models enhance the performance of existing conventional techniques. In addition to that, these models are implemented by closely considering the biological inspiration of the nature. In a nutshell, neural networks, fuzzy systems, evolutionary computation, swarm intelligence, probabilistic reasoning, multi-agent systems and combination of these techniques created the era of soft computing paradigm. Among all, to attain global optimization of learning algorithms, Genetic algorithms, a biologically imitating technology is more suitable for feature selection problem in the classification of proposed work.

The remaining paper is organized as follows: In section 2,a brief related work is given. Then, in section 3 the proposed Intelligent Text Data Classification System is presented. Subsequently, the experimental analysis is showcased in section 4. Finally the conclusions are made.

2. RELATED WORK

Text mining or knowledge discovery from text (KDTD) is first time mentioned by Feldman et al. [27], which deals with the machine supported analysis of text data. Text mining is an interdisciplinary field, the confluence of a set of disciplines, including information extraction, information retrieval, data mining, soft computing and statistics. Over the last decade, the research contributions of many authors [21] are evident that the current text mining research mainly tackles the problems of information retrieval, text document categorization, and information extraction. Among all, text document categorization is a current and promising research area [19] of text mining and its future paths have triggered the present research in the field of text document categorization. Various authors [7,9,12] have found that the document categorization can be done in two ways: either try to assign keywords to documents based on a given keyword set (classification) or automatically structure document collections to find groups of similar documents (clustering). The proposed work has got the impression towards the classification and present brief related work from 2010 to the current year.

In 2010,the authors [21] are emphasized that text mining got a high commercial potential value as the most information is stored as text in real world applications. They mainly concentrated on presenting various issues such as information extraction, information retrieval, document clustering and classification with respect to text mining techniques. Among all issues, the authors [20] taken up the problem of document clustering based on semantic contents of document. They mostly pay their effort on preparation of feature selection by considering the meaning behind the text, then, they adapt the hierarchical agglomerative clustering algorithm to perform process of clustering. Other set of authors [19] shown the interest in other issue, namely, text document classification.They provide a review and comparative study on approaches and document representation techniques in document classification. In conclusion, they endorsed that the KNN algorithm is most appropriate for hybrid approaches for automatic classification.

After that in the year 2011, the authors [17] made a report on promising research results of data and text mining methods for real world deception detection. Their discussion revealed the combination of text and data mining techniques can be successfully applied to real world data to produce better results. Meanwhile, they felt that due to increasing demand of text mining, the extra effort is needed to achieve high accuracy in document segmentation process. With that objective, the authors [16, 18] analysed the problem of classification and clustering within the framework of genetic algorithm respectively. Empirical testing has been performed and the results are justified the need and relevance of genetic algorithm for both classification and clustering processes.

All the range in 2012 and 2013, the authors [10, 13, 14,15] discussed the significance of applying pre-processing techniques and presented the sequential stages of pre-processing in text data categorization. From the literature, they recognized that the KNN is one of the most accepted classification technique to segment the data efficiently.However, they noticed the traditional KNN has some limitations like high calculation complexity, dependence on the training set and no weight difference between samples. To overcome such limitations, the authors proposed numerous approaches with the combination of clustering techniques or by implementing weighting strategy with traditional KNN. In addition to that, they expressed that the combination of KNN with machine learning techniques definitely yield the optimal solutions.

In between 2014 and 2015, some of the authors[4,5,8] make use the text mining process on various applications such as identification of plagiaristic behaviour in online assignments as well as in research papers, and finding software vulnerabilities via bugs. Their results indicated that, text mining analysis require too much processing power and time and thus, it is a promoted as

great challenge in the large scale applications. They prominently conclude that further studies are required to improve the performance of text mining algorithm with soft computing paradigm for detecting text similarities to get faster results. In this path, the authors [6] use the genetic algorithm for efficient text clustering and text classification process. Their research is proved that the incorporation of genetic approach in text mining process can definitely yields the best results and it becomes one of the promising area of research.

The recent research review in 2016, carried out by many authors [1,2,3] evident that text classification is a fully potential area of research in text mining. In addition to that, the alone usage of KNN could not produce satisfactory results. To improve the performance of KNN in terms of efficiency and accuracy, it is necessary to employ machine learning techniques and then KNN is appropriate to deal with large amount of data. with this motivation, the authors in the present paper take-up the issue of text data classification with the combination of well-known KNN and genetic algorithm.

3. PROPOSED INTELLIGENT TEXT DATA CLASSIFICATION SYSTEM-ITDCS

In order to overcome the challenges involved in the earlier classifiers and to improve the accuracy and coverage of text data classifier the authors in the present paper propose an Intelligent Text Data Classification System (ITDCS) as shown in figure 2.

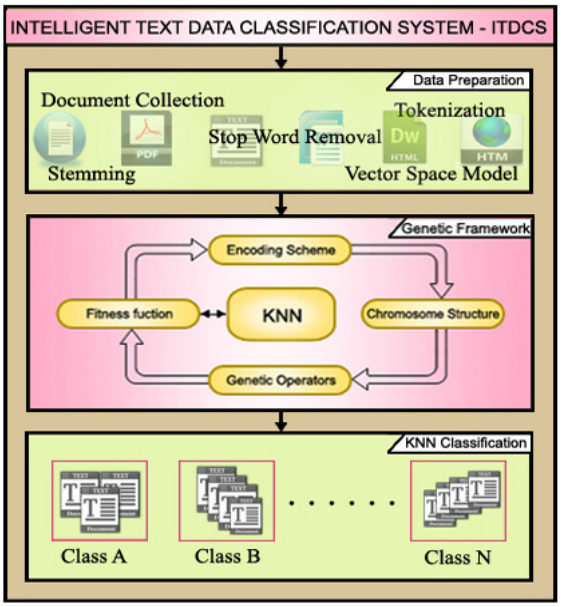

Figure 2. Architecture of ITDCS

The proposed ITDCS is designed and developed in the research framework of genetic approach with a focus to improve overall classification accuracy. At first, the ITDCS focuses on data preparation that includes document collection, tokenization, stop word removal, stemming and attribute selection. Later, it emphasises on classifying the text documents using the standard KNN classification algorithm. In this process of classification the authors integrated Genetic algorithm for feature selection so as to enhance the accuracy of the classifier and improve the coverage of all unclassified documents with its biological components. As a result, finally the classifier ITDCS comprehensively able to classify the text documents accurately into optimal labelled classes.

3.1. ITDCS - DATA PREPARATION

As the classification algorithms unable to understand the documents directly one has to represent the raw documents into a suitable structure. In addition, the success and accuracy of any text mining classifier is highly and directly depend on the effective preparation of input raw data. In the text mining it is important to identify the significant keywords that carry the meaning contribute towards the quality of the further stages of the classifier.The data preparation stage converts the raw original textual documents in a classifier ready structure. With this aim, the ITDCS initially concentrates on the data preparation with all possible procedural methods. The ITDCS data preparation covers all the procedural steps including text document collection, tokenization, stop word removal, stemming and attribute selection.

3.1.1 DOCUMENT COLLECTION

The ITDCS process starts with a collection of text data which include a set of documents with many extensions like .pdf, .txt or even flat file extension. They are normally collected from different sources like web, individual data stores, real world data, natural language text documents, online chat, emails, message boards, news-groups, blogs and web pages etc. Also, text dataset is created by processing spontaneous speech, printed text and handwritten text that contain processing noise. However, the data collection is purely dependent on the application domain of the text mining.

3.1.2 TOKENIZATION

In order to get all words in a given text, tokenization is required that splits the sentence into a stream of words by removing all punctuation marks. The present system ITDCS discovers the words as tokens by splitting a text into pieces and then remove all punctuation marks using it sparser. These token representations are separated by replacing a single white space between the words as word boundaries and then make use for further execution. The ITDCS create a dictionary with a set of tokens obtained by combining all text documents. The ITDCS parser finds significantly meaningful keywords and transforms the abbreviations and acronyms into a standard form. In addition, the tokenization helps in maintaining consistency in all the collected documents. The ITDCS successfully perform the phase of a tokenization and it allow for an accurate classification of text documents.

3.1.3 Stop Word Removal

In the next step, the data preparation of ITDCS pays an attention to remove the stop words because these words may misguide the entire text mining process. Stop words are words that are language specific words with very little or no semantic value as they are used to provide structure in the language rather than the content. Specifically, in English language,the stop word list may consist of pronouns, prepositions, and conjunctions etc. that are frequent words that carry no meaningful information. Additionally, the plain stop words such as names of day, names of month, one or two character terms, non-alphabetical characters are sensibly discarded during this step. In addition, the terms printed on each page of a document are treated as the template words, they also deleted as they negatively impact the text mining results. Specially, the advantage of elimination of these unwanted words reduces the dimensionality of input data of the classification. A sample list of English stop words is shown in table 1.

Table 1 : List of Stop Words in English

a	before	down	himself	her	other	than	to	where's
about	being	during	his	here	ought	that	too	which
above	below	each	how	here's	our	that's	under	while
after	between	few	how's	hers	ours	the	until	who
again	both	for	i	herself	ourselves	their	up	who's
against	but	from	i'd	him	out	theirs	very	whom
all	by	further	i'll	more	over	them	was	why
am	can't	had	i'm	most	own	themse	wasn't	why's
an	cannot	hadn't	i've	mustn't	same	lves	we	with
and	could	has	if	my	shan't	then	we'd	won't
any	couldn't	hasn't	in	myself	she	there	we'll	would
are	did	have	into	no	she'd	there's	we're	wouldn't
aren't	didn't	haven't	is	nor	she'll	these	we've	you
as	do	having	isn't	not	she's	they	were	you'd
at	does	he	it	of	should	they'd	weren't	you'll
be	doesn't	he'd	it's	off	shouldn't	they'll	what	you're
because	doing	he'll	its	on	so	they're	what's	you've
	don't	he's	itself	once	some	they've	when	your

3.1.4 STEMMING

Here the data preparation of proposed ITDCS concentrates further to reduce the size of the input data and helps in enhancing the intelligence of the proposed system. Stemming is an important process as it emphasises in eliminating the words having grammatically same meaning which are connected to the same root word. This process works with an objective to remove the words which are derived different forms such as noun, adjective, verb, adverb, plurals etc. from the root word. This procedure incorporates a great deal of language dependent linguistic knowledge but not on the domain. In general, the stemming process majorly considers two important points: one is the morphological form of a word that has the same base meaning should be mapped to the same stem. The other point is the words that do not have the same meaning treat as separate.

Towards this, the ITDCS data preparation stage rightly uses a standard statistical n-gram stemmer. It follows string similarity approach to covert word inflation to its stem and language independent. The main idea behind this approach is that, similar words have a high proportion of n-grams in common. In this method, an n-gram is a string of n, generally adjacent, characters extracted from a section of sentence. For n equals to 2 or 3, the words extracted are called di-grams or tri-grams, respectively.

For example, the word ‘INTRODUCTIONS’ results in the generation:

Di-grams:

I, IN, NT, TR, RO, OD, DU, UC, CT, TI, IO, ON, NS, S

Tri-grams:

I, *IN, INT, NTR, TRO, ROD, ODU, DUC, UCT, CTI, TIO, ION, ONS, NS*, S

Where '*' denotes a padding space. There are n+1 such di-grams and n+2 such tri-grams in a word containing n characters. Generally a value of 4 or 5 is selected for n. A well-known statistical analysis based on the Inverse Document Frequency (IDF) is used to identify them.

3.1.5 ATTRIBUTE SELECTION

The attribute selection is the key factor in building any classifier in the process of text mining. Though the previous stages of data preparation of ITDCS reduces the size of the input data by eliminating unwanted words, it is necessary to identify the attributes which are contributing more to the problem domain. With this aim, the authors in the present paper use the standard following measures to identify the required attributes and improve the overall performance of the proposed system.

Term Contribution: It is calculated on the basis of how much a specific term contributes to the similarity among all collected text data documents.

$$TC(t_k) = \sum_{i,j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j) \qquad (1)$$

Where $f(t_k, D_i)$ is the TF-IDF of the k^{th} term of the i^{th} document.

Term Variance: It calculates the variance of all terms among all the text data documents. Later, assign the maximum scores to terms that have more document frequency and a common distribution value.

$$v(t_i) = \sum_{j=1}^{N} \left[f_{ij} - \overline{f_i} \right]^2 \qquad (2)$$

Where f_{ij} is the frequency of the i^{th} term of the j^{th} in document and $\overline{f_i}$ is the mean frequency of the terms in the document collection.

Term Variance Quality: This measure calculates the quality of a term by using the total variance.

$$q(t_i) = \sum_{j=1}^{n} {f_{ij}}^2 - \frac{1}{n} \left[\sum_{j=1}^{n} f_{ij} \right]^2 \qquad (3)$$

Where f_{ij} is the frequency of the $i^{th\ term}$ of the j^{th} document.

As a final point, the filters assign ranks to all the attributes, so the top-ranked attributes are treated as features and give as an input for the next stages of ITDCS.

3.1.6 TEXT DATA ENCODING AND STRUCTURED REPRESENTATION USING VECTOR SPACE MODEL

In order to classify large number of text documents it is equally important to encode the words into numerical values using any standard approach in the preparation of text data. With this objective, the proposed work employs a proven vector space model which enables the classifier with efficient classification. This representation model usually improves the performance of the

classifier as it is designed based on term weighting scheme. Larger weights are assigned to the words that are frequently identified in the text documents, thus, it able to encode the input data appropriately.

The document weight w(d, t) for a term 't' in a document 'd' is computed by the product of Term Frequency TF(d, t) with Inverse Document Frequency IDF(t). Based on document weight, the ITDCS prepares vector space model that represents document as vectors in m-dimensional space. Here each document d is described by a numerical feature vector $W_d = \{x(d, t_1), x(d,t_2), \ldots\ldots x(d, t_m)\}$. Now, the documents are compared by performing simple vector operations for text mining process.

A sample dimensional vector space model is as shown in table 2, here, each term is a component of the vector and the value of each component is the number of times the corresponding term occurs in the document. The documents are the rows of this matrix, while the features are the columns.

Table 2 : Sample Dimensional Vector Space Model

	Team	Play	Coach	Ball	Game	Score	Win	Lost	Timeout	season
Doc 1	3	0	5	0	2	6	0	2	0	2
Doc 2	0	7	0	2	1	0	0	3	0	0
Doc 3	0	1	0	0	1	2	2	0	3	0
Doc 4	5	5	4	9	16	3	0	8	7	12
Doc 5	10	1	13	12	3	7	3	5	9	5
Doc 6	6	4	6	13	15	6	8	8	5	8
Doc 7	39	10	23	5	21	3	6	3	8	3
Doc 8	3	9	9	16	20	9	2	5	12	6
Doc 9	0	8	19	3	4	2	1	9	7	12

The prepared structured text data generated by vector space model is given as the basic input for the intelligent text data classification system.

3.2. ITDCS - CLASSIFICATION

The accuracy of building any classifier in the era of text mining refers the ability of predicting the known and unknown class labels of training data. The standard KNN classifier is unable to reach the expected level of accuracy in predicting the class labels. The uneven class distribution of conventional KNN technique is posed many challenges. In addition, the conventional classification techniques computationally expensive and leads to static solutions. To attain global optimum solution, one has to use an intelligent classification technique that can predict the class labels accurately and reaches the global solution. The accuracy and coverage are two measures for building the intelligent classifier. Moreover, relevance analysis of attribute selection is also key parameter in classification. The authors in the literature proposed many intelligent techniques, Genetic approach is rightly suitable for developing accurate supervised classifier. Towards this, the authors in the present paper propose an intelligent classification algorithm which is designed on the biological inspiration of Genetic approach.

The proposed intelligent text data classification system concentrates on modelling the classifier from the training data in the light of genetic algorithm with its all appropriate biological

components. The encoding strategy of Genetic approach model the global solution to the present text data classification problem. The continuous evaluation offitness function of Genetic algorithm builds the ITDCS classifier to overcome the challenges of "susceptible to overfitting" and "misclassification". The biologically inspired operators of genetic algorithm create an added intelligence to the ITDCS classifier and thus the boundaries of the classes are more flexible. As a whole, the comprehensive approach of all the components of Genetic makes the classifier to produce right prediction on finding the class labels accurately within time. Then, this classifier applies on the new testing data, fed by the data preparation stage of vector space model to predict the right class labels.

3.2. 1ITDCS - BASIC PRINCIPLES

Genetic algorithm is an efficient, robust, an adaptive evaluation and self-learning process which is generally applied on high volume, complex and multi-dimensional data. This is designed on the principles of natural Genetic systems, each individual text data document is encoded as chromosomes. This learning algorithm uses application domain dependent knowledge to compute the fitness function to create more promising solutions. According to the theory of evaluation, each initial individual chromosome is associated with fitness value to acquire the expected accuracy. Various biologically inspired genetic operators like selection, crossover and mutation are applied on these chromosomes to get potentially better global solution.

Genetic Algorithms (GA)are different from most of the normal learning algorithms in the following ways:

- GAs work with the coding of the parameter set, not with the parameters themselves
- GAs work simultaneously with multiple points, not with a single point.
- GAs optimize via sampling using only the payoff information
- GAs optimize using stochastic operators, not deterministic rules

3.3.2 ITDCS - ENCODING STRATEGY

The encoding strategy is a process to represent the data fed by the data preparation stage of ITDCS system into a right form of genetic algorithm. It is an important process in the genetic approach as it plays a key role to yield a best global performance of learning algorithm. Many encoding techniques like tree encoding, permutation encoding and binary encoding are used by the genetic algorithm.

The authors in the present paper for ITDCS, the Binary encoding technique is adapted to create initial population. The binary encoding encodes the chromosome with two bits as 0 and 1. Consider following example of documentfeatures {F1, F4, F5, F8} is encoded as a binary chromosome of length 10 and is represented in figure 3. The presence of a feature in a document is encoded as 1, otherwise encoded as 0.

Figure 3. Example of Binary Encoding Chromosome

3.3.3 ITDCS – FITNESS FUNCTION

The authors in the present paper employ KNN classifier to classify the text documents into predicted class labels which is generally works based on selected best features identified by genetic process. In this process the Genetic approach is intrinsically designed to train the ITDCS classifier to attain enough intelligence for accurate classification. Generally, to attain the accuracy in classification, the selection of features is playing an important role. Specifically, impact of accuracy of the classification, not only depends on the number-of-neighbours, but also dependent on weightage-of-individual neighbour.

Towards this, the fitness function of proposed ITDCS is designed with an intension to compute weightage-of-individual neighbour by its set of optimal parameter weights in the continuous evaluation process. The set of indicated parameters overcome the problem of susceptible to overfitting and reduce the rate of misclassification. In addition, it takes the responsibility of achieving the level of expected accuracy in predicting the exact class label of a text document. The ITDCS uses a robust fitness function which is formulated based on total documents to be classified, correctly classified documents and number of nearest neighbours.

$$Fitness\ Function\ (FFC) = \alpha \frac{(TotDocs - CCDocs)}{TotDocs} + \beta \left(\frac{NNN/_K}{TotDocs} \right)$$

Where,

FFC : Fitness Function of the Classifier
TotDocs : Total number of text Documents
CCDocs : Correctly Classified Documents
NNN : Number of Nearest Neighbours having less impact on classification
K : Number of Nearest Neighbours having high impact on classification
α, β are the constants to tune the learning algorithm

The goal of genetic algorithm is typically expressed by its fitness function that evaluates rate of accuracy in the labelled classes. The performance of ITDCS fitness function is highly proportional to the choosing number of times of individual as initial population of text document features. Therefore, the bestfit individual features have able to generate better population in the next generation while low fitness individuals likely to disappear.

3.3.4 ITDCS – OPERATORS

In addition to the choosing best individual population of features by the ITDCS fitness function, genetically inspired operators selection, crossover and mutation create the best fit solution. These operators are applied on the selected initial population to generate possible better new population. The selection, crossover and mutation operators designed intune of ITDCS that comprehensively transforms individual features of text document stochastically. Each feature chromosome has an associated fitness value that contributes in the generation of potential new population using operators. At each transformation, the ITDCS utilizes the fitness function value to evaluate the survival capacity of newly generated chromosome feature. As a whole, the operators of ITDCS create a new set of population chromosomes to improve the fitness function value.

SELECTION:

The mating pool is reproduced from choosing a particular individual chromosome feature by the selection operator with mimic nature of selection procedure. The number of iterations of choosing

the individual chromosome is also decided by the selection operation. Only the selected chromosomes in generating mating pool are able to take part in the subsequent genetic operations. Among the several available selection methods, ITDCS deploy the roulette wheel parent selection technique. The roulette wheel has as many slots as the population size, where the area of the slot is proportional to the relative fitness of corresponding feature chromosome in the population. An individual feature chromosome of the text document is selected by spinning the roulette wheel and record the position when the wheel stops. Therefore, the number of times a feature chromosome selected is proportional to its fitness value, in the given population. The procedure is illustrated as shown in figure 4.

Figure 4. Example of Roulette Wheel Parent Selection

The selection operator of ITDCSallots a probability of selection $P_j$to each individual feature j based on its fitness function value.

A series of N random numbers is generated and compared against the cumulative probability $C_i = \sum_{j=1}^{i} P_i$ of the feature population.

The suitable individual i is chosenfor the new population if $C_i < U(0,1) < C_i$.

The probability P_i for each individual chromosome is defined by:

$$P[Individualiischosen] = \frac{F_i}{\sum_{j=1}^{Popsize} F_j}$$

Where, F_i is the fitness of individual i.

CROSSOVER:

The genetic algorithm uses the crossover operation that mates two randomly selected parent chromosomes to generate two new child chromosomes. The objective of the crossover operator is producing new chromosomes that are potentially offer best optimal features than the parents by acquiring the best characteristics from their parents. Single point, two point and uniform crossover are the well-known approaches to evaluate the process according to defined crossover probability factor by the user. In this paper, to produce optimal feature chromosomes, the ITDCS elect the single point crossover function and define the crossover probability as Cp. The ITDCS crossover operator identify a single crossover point and then exchange portions of the two parent chromosomes lying to the right of the crossover point to produce two child chromosomes. For

chromosomes of length L, a random integer, called the crossover point, is generated in the range [1, L-1]. An example is as shown in figure 5 and 6 before and after crossover respectively.

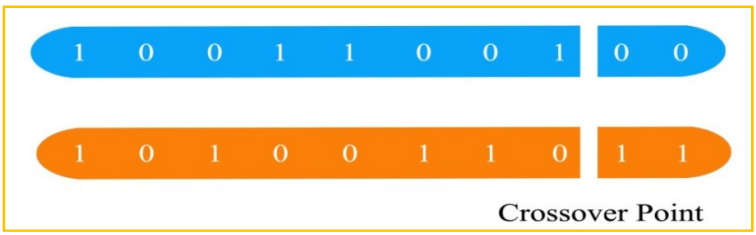

Figure.5 Single Point Crossover Operation Before Crossover

Figure.6 Single Point Crossover Operations After Crossover

MUTATION:

The mutation works with an aim to introduce genetic diversity into the newly produced population. In this process, a random alteration is takes place in the genetic structure of a chromosome. With these altered gene values the genetic algorithm able to arrive at a best fit solution than the previously possible solution. This mutation helps the genetic algorithm to prevent the population from stagnating at any local solution. Here, ITDCS uses bit flip mutation where each bit of the chromosome is given to mutation with a mutation probabilityMp. For example, for binary representation of chromosomes, a bit position is mutated by simply flipping its value. In the figure 7, a random position is selected in the chromosome and replace by negation of bit.

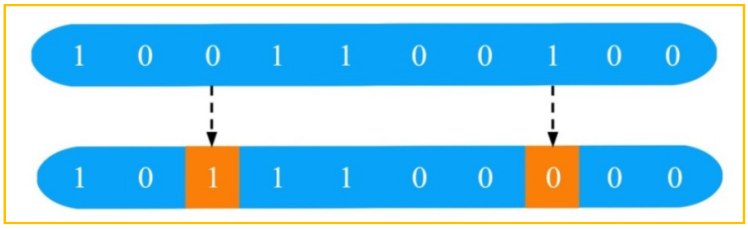

Figure 7. Process of bit-by-bit Mutation

3.3.5 ITDCS – ALGORITHM

The ITDCS is designed with the theories and techniques of genetic approach for identifying optimal features to classify text documents effectively. In this process, encoding technique along with fitness function tuned the length of the feature chromosome and the size of population. The biologically evaluation operators calculate and then adjust the probabilities of selection ,

crossover and mutation so that ITDCS acquires self-learning capability. The ITDCS algorithm is perfectly designed and implemented in the framework of genetic approach as follows,

ALGORITHM :: ITDCS

Step 01. Start
Step 02. Apply pre-processing techniques on the target text data documents
Step 03. Load the pre-processed documents in the database D.
Step 04. Build the classifier by applying KNN Classification algorithm based on identified features of the training data set TD and S is a set of identified features which are distributed in total documents.
Step 05. Set Q = φ where Q is the output set, which contains all labelled classes with optimal classification accuracy.
Step 06. Set the input termination condition of genetic algorithm.
Step 07. Represent each document of a labelled class of 'S' as binary encoding.
Step 08. Select two members from the labelled class.
Step 09. Repeatedly apply GA operators, crossover, and mutation functions on the selected members to generate optimal labelled classes.
Step 010. Find the fitness function value.
Step 011. If fitness function value meet the target classification accuracy then
Step 012. Set $Q = Q \cup Q^1$ where Q^1 is the set of accuracy labelled classes.
Step 013. If the desired number of generations is not met then go to Step 4.
Step 014. Stop

4. EXPERIMENTAL ANALYSIS

The authors in the present paper implemented the ITDCS on large number of text data documents under standard execution environment. The integrated methodology of KNN with genetic algorithm takes the input from pre-processed structured representation of vector space model after performing data preparation stages. In order to identify the performance of proposed ITDCS the authors conducted many experiments using standard datasets available on World Wide Web. A dataset consists of 52 categories and a total of 6532 documents for training and testing the proposed work. Among all 52 categories, the authors have chosen 36 categories which are distributed in total number of documents.

A series of experiments are conducted and results are showcased with respect to accuracy rate, crossover probabilities and execution time. The results are as follows,

a) Various crossover probabilities (Cp=0, 0.25, 0.5, 0.75, 1) are defined while iterating number of times to find the improvements in accuracy levels of proposed genetic based classifier. It is clearly evident that the ITDCS is performing well at the average mean Cp= 0.5 as shown in figure 8.

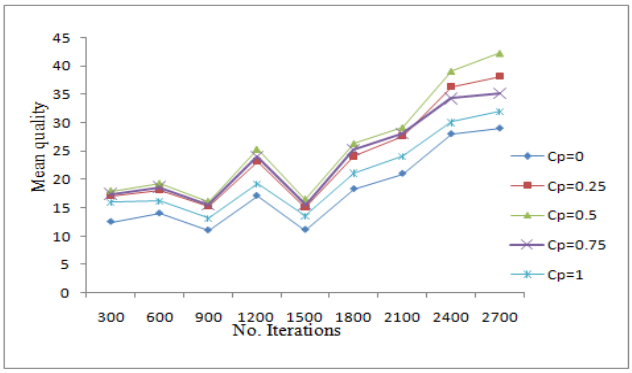

Figure 8. Mean quality of ITDCS

b) The experiment results noticed that intelligent text data classification system reduces the human intervention to classify the text data documents. The error rate between the testing data and training data is almost minimized in ITDCS and is found to be 0.2 on an average. The nature of relationship between testing and training text data is studied and both are proven as continuous and linear as shown in figure9.

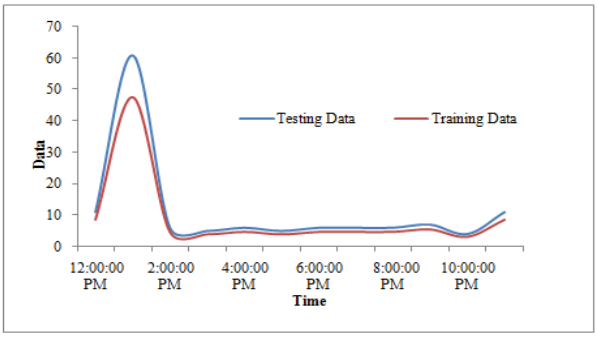

Figure 9. Error Rate between Training and Testing Data

c) From the results shown in figure 10, the performance of standard KNN and proposed ITDCS is not significantly varying for small values of K. However, as K value increases, the genetic approach of ITDCS noticeably improves its performance over standard KNN classification technique.

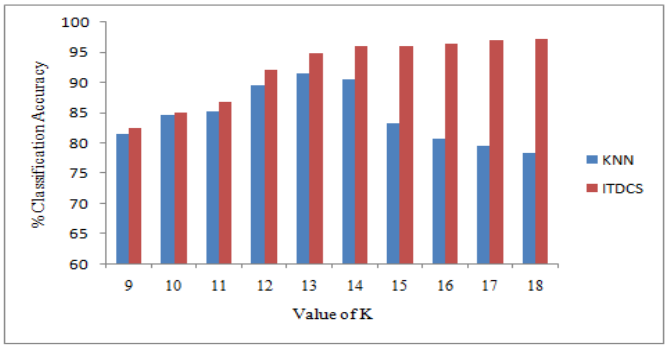

Figure 10.Classification Accuracy with Respect to K Value

d) The standard KNN classification technique, the intelligent classifier of ITDCS is tested on the same standard environment for a common test data. The test results are compared, the intelligent classifier of ITDCS after training takes less execution time when compared with conventional KNN classifier as presented in figure 11.

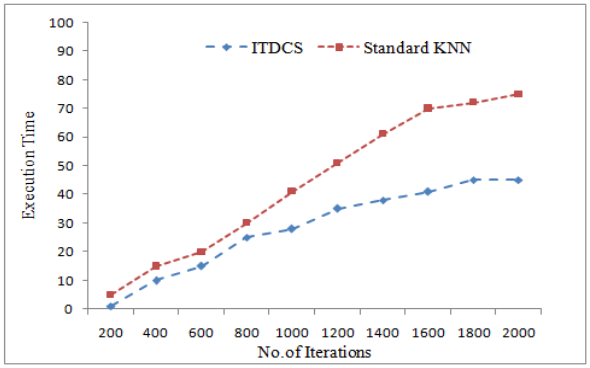

Figure 11. Execution Performance of ITDCS

e) The proposed ITDCS is also compared with the other existing classification techniques like J48, Naïve Bays (NB) and standard KNN on various data sets. The experimental results are evident that the proposed method performs better than other techniques, even though the dataset size is large.

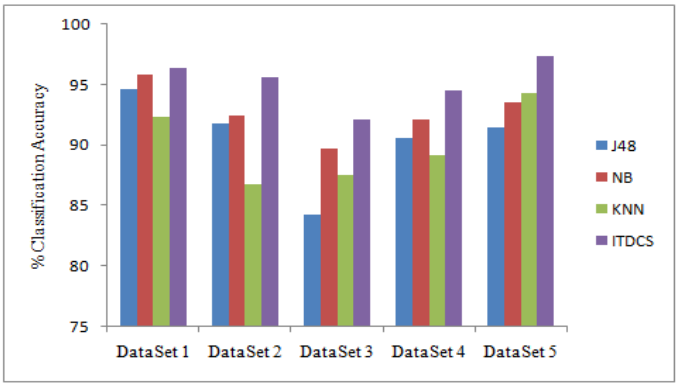

Figure 12.Comparison of Classification Accuracy

5. CONCLUSIONS

The authors in the present paper proposed an intelligent text data classification system which is designed with the integration of genetic based feature selection method to addresses the challenges of text data classification. The proposed system prepared the data using vector space representation for rightly suitable to KNN classification technique. The chosen encoding strategy of proposed ITDCS exactly represents each feature of the document as a chromosome and rightly populates the initial population. The measures taken in the fitness function optimally evaluate the survival fitness of the generated chromosomes. The biological imitation nature of ITDCS operators efficiently find the minimal feature subset and forwarded towards optimal solution. The integration of genetic algorithm select optimal feature set dynamically. The experimental results of ITDCS demonstrated the high learning performance of the classifier and the ability to converge

quickly. The proposed system has proven the need of genetic algorithm in the process of classification for selecting the optimal features. Moreover, the stochastic process of ITDCS significantly reduces the misclassification rate as well as improves the performance in terms of accuracy and efficiency in classifying text documents.

Applying the optimization techniques for text data classification and clustering is still in early stage as well as has open challenges. Thus, it creates a solid base for promising future research direction. To improve the performance genetic repetition, design genetic algorithm using parallel processing is also another future path.

REFERENCES

[1] RupaliP.Patil, R.P.Bhavsar, B.V.Pawar, "A Comparative Study of Text Classification Methods: An Experimental Approach", International Journal on Recent and Innovation Trends in Computing and Communication, Vol: 4, Iss: 3, pp:517-523, 2016.

[2] Shivani Sharma, Saurbh Kr. Srivastava, "Review on Text Mining Algorithms", International Journal of Computer Applications, Vol 134, No.8, pp: 39-43, 2016.

[3] ZhenyunDeng, XiaoshuZhu,et.al., "Efficient kNN classification algorithm For big data", Neurocomputing, Elsevier, vol. 195, pp: 143–148, 2016.

[4] GokhanAkcapınar, "How automated feedback through text mining changes plagiaristic behavior in online assignments", Computers & Education, Elsevier, Vol.87, pp: 123-130, 2015.

[5] Ranjeet Kumar, R.C.Tripathi, "Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset", Procedia Computer Science, ELSEVIER, Vol. 65, pp: 911 – 919, 2015.

[6] Sung-Sam Hong, Wanhee Lee, et.al, "The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification", Int. J. Advance Soft Compu. Appl, Vol. 7, No. 1, pp: 22-40, 2015.

[7] Charu C. Aggarwal, Yuchen Zhao, et.al., "On the Use of Side Information for Mining Text Data", Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, pp: 1415-1429, 2014.

[8] DumiduWijayasekara, Miles McQueen, "Vulnerability Identification and Classification Via Text Mining Bug Databases", IEEE, pp: 3612-3618, 2014.

[9] Deepankarbharadwaj, suneetshukal, "Text Mining Technique using Genetic Algorithm", International Conference on Advances in Computer Application, Proceedings Published in International Journal of Computer Applications, pp: 07-10, 2013.

[10] J. AlameluMangai, Satej Milind Wagle, and V. Santhosh Kumar, "A Novel Web Page Clasification Model using an Improved k Nearest Neighbor Algorithm"; rd International Conference on Intelligent Computational Systems pp: 49-53, 2013.

[11] M.Akhiljabbar, B.L DeekshatuluPriti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", Procedia Technology, ELSEVIER, Vol. 10, pp: 85–94, 2013.

[12] DivyaNasa, "Text Mining Techniques- A Survey", International Journal Of Advanced Research In Computer Science And Software Engineering, vol. 2, issue 4, pp: 50-54, 2012.

[13] PutuWiraBuana, SesaltinaJannet D.R.M, I KetutGede Drama Putra, "Combination of K-Nearest Neighbor And K-Means Based on Term Re-Weighting for Classify Indonesian News", International Journal of Computer Applications, Volume 50, No.11, pp: 37-42, 2012.

[14]Tsung-Hsien Chiang, Hung-Yi Lo, Shou-De Lin, "A Ranking-based KNN Approach for Multi-Label Classi_cation", JMLR: Workshop and Conference Proceedings,Vol. 25pp: 81-96, 2012.

[15] VandanaKorde, C NamrataMahender, "Text Classification And Classifiers: A Survey", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2,pp: 85-99, 2012.

[16] Chien-Pang Lee, Wen-Shin Lin, Yuh-Min Chen, Bo-JeinKuo, "Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method", Expert Systems with Applications, Elsevier, Vol. 38, pp: 4661–4667, 2011.

[17] Christie M. Fuller, David P. Biros, DursunDelen, "An investigation of data and text mining methods for real world deception detection", Expert Systems with Applications, Elsevier, Vol. 38, pp:8392–8398, 2011.

[18] N. El-Bathy, C. Gloster, I. Kateeb1, G. Stein, "Intelligent Extended Clustering Genetic Algorithm for Information Retrival Using BPEL", American Journal of Intelligent Systems, Vol. 1(1), pp: 10-15, 2011.
[19] Aurangzeb Khan, BaharumBaharudin, Lam Hong Lee, Khairullah khan, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances In Information Technology, Vol. 1, No. 1, pp: 04-20, 2010.
[20] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq, "Document Clustering based on Topic Maps", International Journal of Computer Applications, Vol. 12– No.1, pp: 32-36, 2010.
[21] Vidhya. K. A, & G. Aghila, "Text Mining Process, Techniques and Tools : an Overview", International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2, pp: 613-622, 2010.
[22] A. A. Shumeyko,, "S. L. Sotnik, Using Genetic Algorithms for Texts Classification Problems", Anale. SeriaInformatica, Vol. 7, pp: 325-340, 2009.
[23] Adriana Pietramala, Veronica L. Policicchio, Pasquale Rullo, and Inderbir Sidhu, "A Genetic Algorithm for Text Classification Rule Induction", Springer-Verlag Berlin Heidelberg, pp: 188–203, 2008.
[24] Milos Radovanovic, Mirjana Ivanovic, "Text Mining: Approaches And Applications", Novi Sad J. Math, Vol. 38, No. 3, pp:227-234, 2008.
[25] R. Gil-Pita and X. Yao, "Using a Genetic Algorithm for Editing k-Nearest Neighbor Classifiers", Springer-Verlag Berlin Heidelberg, pp:1141–1150, 2007.
[26] SanghamitraBandyopadhyay, Sankar k. Pal, "Classification and Learning Using Genetic Algorithms", Application in Bioinformatics and Web Intelligence, Springer, 2007.
[27] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.

AUTHORS

Dr. V V R Maheswara Rao is working as Professor in the Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women (Autonomous), Bhimavaram. He received M.Tech. CSE from JNTU Kakinada and Ph.D. from Acharya Nagarjuna University, Andhra Pradesh, India. He has 6 years of IT industry experience and 12 years of Teaching and Resea rch experience. He has more than 25 papers to his credit in many international and national journals and conferences. His research interests are Data Mining, Web Mining, Text Mining, Artificial Intelligence, Machine Learning and Big Data Analytics. He has associated with two projects funded by DST.

N Silpa is working as Assistant Professor in the Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women (Autonomous), Bhimavaram. She received M.Tech. CSE from JNTU Kakinada, Andhra Pradesh, India. She has 7 years of Teaching and Rese arch experience. Her research interests are Data Mining, Text Mining, Machine Learning and Big Data Analytics. She has associated with one project funded by DST.

Dr.Gadiraju Mahesh has been working as an Associate Professor in Department of Computer science and Engineering, S.R.K.R. Engineering College, Bhimavaram, India. Worked in Col. D.S. Raju polytechnic for 17 years and has a total teaching Experience of 27 years. Done M. Tech. (C.S.E.) from J.N.T.U., Kakinada and completed Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University in 2016 under the guidance of Prof. V. ValliKumari, Andhra University. Guided many M.Tech. and M.C.A. students in their thesis work.