Incremental Learning: Areas and Methods – A Survey

Prachi Joshi[1] and Dr. Parag Kulkarni[2]

[1]Assistant Professor, MIT College of Engineering, Pune
prachimjoshi@gmail.com
[2]Adjunct Professor, College of Engineering, Pune
paragakulkarni@yahoo.com

ABSTRACT

While the areas of applications in data mining are growing substantially, it has become extremely necessary for incremental learning methods to move a step ahead. The tremendous growth of unlabeled data has made incremental learning take up a big leap. Starting from BI applications to image classifications, from analysis to predictions, every domain needs to learn and update. Incremental learning allows to explore new areas at the same time performs knowledge amassing. In this paper we discuss the areas and methods of incremental learning currently taking place and highlight its potentials in aspect of decision making. The paper essentially gives an overview of the current research that will provide a background for the students and research scholars about the topic.

KEYWORDS: *Incremental, learning, mining, supervised, unsupervised, decision-making*

1. INTRODUCTION

An important problem that is faced in data mining process is continuously evolving new data. It is essential that existing approaches of classification and clustering tackle this in such a way that the classifier is tuned-in to accommodate it. This is where we need incremental learning; a learning that occurs with the new data and is taken further in the process. With the methods of machine learning like k-means clustering which is considered to be one of the pivot block of machine learning has to undergo multiple sweeps prior to stabilization of clusters [1][2]. Other techniques like hierarchical ones do not consider the overall size of the cluster [3].The methods of supervised and semi-supervised learning allow us to learn and classify with the help of training data. In the design of efficient learning algorithms, condensed pre-processed data along with new evolving data creates certain issues. Handling the problem of learning the knowledge while maintaining the previous one, is the most important target for the incremental learning methods [4]. Another issue is how to handle the learning when the labeled data is quiet low and is difficult to obtain with user expertise. Identifying the areas of whether the learning is to be done online with the data stream being continuous also affects the learning process. Time and memory constraint also play a crucial role in the learning process, hence making it necessary for the incremental learning to be effective at the same time be accurate.

DOI : 10.5121/ijdkp.2012.2504

The current research in incremental learning is not just about incremental update of the data, but how to utilize the updated data as knowledge for further mining and forecasting. The field of research has come up with various methods for incremental learning. In this paper we will discuss how incremental learning is used with existing learning environments of supervised to unsupervised methods. Though with the current work, lot of mining and incremental learning activity is carried out with respect to specific domains and applications, there are a varied methods and techniques that prove to be more useful and beneficial with regard to the type of application. This paper discusses about the methods and their domains clarifying the concepts of incremental learning.

2. EXISTING LEARNING METHODS AND INCREMENTAL LEARNING

First and foremost we will discuss about the methods of unsupervised learning to supervised techniques, with the need for incremental learning. The paper proceeds in stages putting light on how incremental learning evolved with these learning methods. The focus of our paper is on the methods, the approaches and their novelty that are used for incremental learning referencing the type of application as well.

In the course of understanding what precisely an incremental approach is, it is necessary to understand that incremental learning can be in terms of the newly gained knowledge as well as evolving new class or a cluster. It can even merge or reform the classes. Precisely we can frame the learning to be one that is –

1. Capable to learn and update with every new data – labeled or unlabeled.

2. Will use and exploit the knowledge in further learning.

3. Will not rely on the previously learned knowledge.

4. Will generate a new class/cluster as required and take decisions to merge or divide them as well.

5. Will enable the classifier itself to evolve and be dynamic in nature with the changing environment. Figure 1 shows representative diagram of incremental learning.

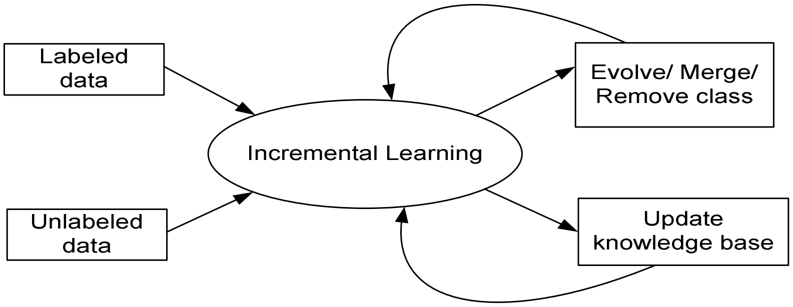

Figure 1: Working of an Incremental Learning

With these factors, it can be rightly said that the incremental learning forms a full package that is up all the time. Considering the traditional methods and where incremental learning stands with respect to each of them along with the applications, is discussed ahead in this section.

2.1 Mathematical Representation and algorithm

When we talk about incremental learning, it is all about the learning approach or the classifier who is capable to perform the activities with respect to the environment. Mathematically it is represented as-

Let $U = \{ud_1, ud_2, ud_3, ...ud_n\}$ be the new unlabeled data and $L = \{ld_i : C_j \mid i = 1 \; to \; \text{n , j = 1 to m}\}$.

Let *Ic* be the classifier that is used for incremental learning. So, we can have-

$$K = f(Ic(U_x), K_{prev}) \text{ where } K = \{C_x, KB\}$$

Here the value of C_x can be of existing class or new generated one. K governs the entire process. This is modeled and learned at every stage of new data availability. The learning process is summarized in the algorithm as follows:

```
1.      For every Dx| Dx Є U or L
2.      Do
                Use KBprev
                If (Dx Є U)
                Classify Dx, with $f$(IC)
                Generate K
                Update KBnew←K+KBprev
                Assign KBprev ←KBnew
```

3. CLUSTERING AND INCREMENTAL LEARNING

Typically, finding patterns that belong to same category or are probable to be in same group is the main task of clustering. In clustering, incremental learning finds a special identity. There is a need for incremental clustering where, new data is accommodated without re-clustering.

The clustering algorithms like k-means[2] are governed with the number of scans where as BIRCH is sensitive to data [5][6]. Other factors that influence the clusters are the selection of centroids and the shape of the cluster formed. The number of clusters to be formed is also a critical parameter that governs the learning. Fast and stable algorithms that are incremental in nature can overcome the difficulties faced by earlier clustering methods. The incremental clustering methods aim at restricting the re-clustering phase, accommodating the newly unlabeled data set, at the same time have the cluster feature updated efficiently and effectively [7].

The research on incremental clustering started with various factors along with various domains; one application domain is in document and image classification. For the same, [8][9] propose incremental clustering where dynamic clustering of points are considered. Comprising of merging and update stages to maintain the clusters, the approach is based on distance measure, where the diameter of the cluster is considered in the cut-off to take up decisions. Further, the clustering techniques tend to employ calculation of similarity measure between the clusters and the data sets, where the new samples are clustered incrementally. [10] propose incremental clustering approach on the same grounds, where threshold value plays an important role in deciding the groups. With the threshold value, an incremental divisive and agglomerative approach is proposed [11] that is used in relational data sets.

In some cases Bayesian methods combined with the similarity measure make the learning more effective [12]. [13] propose a GRIN algorithm in comparison to BIRCH (that supports incremental hierarchical clustering), where the approach is based on gravitational theory of physics that is capable to handle large databases.

Managing a warehouse with new data is a challenge that is handled by incremental learning methods. Existing approach of DBSCAN is enhanced further to be used in the changing environment incrementally, where part of the cluster affected is examined and the clusters are updated with the insertions and deletions considering their density [14].

Incremental methods for web pages categorization are also now making a place. Centroid based approach along with updating of documents and web pages re-assignment are proposed [15].

Adaptive Resonance Theory (ART) is again a popular concept that is been used with unsupervised neural network for incremental learn. Having different variants that are specific to the type of data, [16] propose modified ART, that handles mixed data attributes, that is again based on distance hierarchy. Further to improve the learning, [17] propose rough sets based approach. The emphasis here is on the clustering of interval data, where dissimilarity function between the representative points of the clusters is defined.
Taking it further, [18] propose incremental methods for trajectories. [19] propose methods defining minimum bounding boxes and rectangles to locate the path for clustering of mobile objects.

Methods of incremental clustering are also been used in pattern based reasoning, where new patterns are incrementally learnt with base of neural network [20] comprising of learning and reasoning phases to support the decisions. Further the technique is used in understanding of the document layout, where the papers belonging to journals of Elsevier, Machine Learning are categorized based on first order logic [21].

It is necessary to make a note that the data sets differ in memory requirements depending on the type of data. The approach selected should be such that the required outcomes or the categorization should be achieved at a faster pace.

From the related study it is worth to mention that most of the incremental clustering for pattern discovery rely on similarity measure between the data points, where as some are managed by threshold. Though we can always get and come up with other combination techniques so as to improve and manage the data in a better way, the outcome should not affect the existing knowledge.

4. Supervised and Semi-Supervised Incremental Learning

In case of incremental learning in supervised or semi-supervised environment, what needs to be looked at is that the training data can arise at later stages. Rather that restricting the environment to the specific data, incremental learning unfolds the learning.

With pattern based techniques, [22] propose learning of new chunk of patterns keeping intact the previous ones on the basis of neural network, where the same can be applied in text domain [23].

In order to avoid the training phase, and update with each new training data that is evolved over the time, ensemble base methods are used [4]. [24] suggest a Learn++, an approach inspired by Adaboost, working on neural network based ensemble classifiers working on digital optics database. Further [25] propose ADAIN, an adaptive framework which focus on use of nonlinear regressive models, but comparatively faster than Learn ++. [26] adopt the Gaussian mixture model and Resource allocating NN for the learning with its application in a dormitory to study the habits of the students.

Performance driven data selection model is another approach [27] where selective incremental learning occurs for the unlabeled data, taking decision to learn from specific data sets that are classified to learn further.

A wide application to have a robot to learn manipulative tasks by use of Markov methods is [28] proposed, where initial stage teaching occurs and later the robot learns.

Extending the approach in medical image segmentation, [29] propose Ripple down rules for knowledge acquisition, where as Bayesian approach is proposed [30] to detect emergency in health surveillance.

Alike [31] propose its use in sports video view classification explicitly in baseball presenting a new distance measure along with a threshold criteria building positive and negative model pools. The discovery of interesting patterns further has given rise to [32] learning in detection of objects occurring in the images in hierarchy. Incremental detection and classifying the new images with existing objects is applied here. In case of face recognition too, incremental learning has taken a step ahead, [33] propose method for adaptive learning of new features and classifiers. Here the feature space is tuned with use of neural networks where Resource Allocating Network and Long Term Memory model is used.

SVM are found to be effective in large number of classification and regression problem. [34] discuss of their use in the optical character recognition has found a new area, with ensembles of SVMs in operation. Ensemble based methods are further used with dynamic weighting scheme that is built for extra training models over the earlier ones giving incremental approach for new training data sets when available in batches as suggested by [35]. [36] focus on ensemble methods to learn concept drift; that are characterized by non-stationary environment applying the approach on weather prediction system where the Learn++ approach is extended.

5. CURRENT SCENARIO

With the related work that is been taking place, the objective of this paper is to make the researcher familiar with the concepts. Currently work that is been carried out focuses on query formulation and data distributions.

To focus a few, [37] present a survey on the different techniques of incremental learning of HMM parameters. The work reviews batch learning techniques for estimation of HMM parameters, trying to remove the impact of use of HMM as priori methods with limited data.

Further, incremental and reinforcement learning and whole system learning could be the next big thing. [38] propose use of incremental reinforcement learning designed for operation in multi-agent scenario. The work is based on modified version of Q-Learning, where the agent is faced with number of tasks that
it learns.

Incremental learning in query formulation is explored by [39], where it assists the user to form query, from arbitrary to structured one trying to bridge the gap between the queries formed to the effectiveness that it can be retrieved.

Very recently incremental approach for managing the concept drift of data distributions [40] is proposed. The domain identified was remote sensing images for land cover classifications. With the kernel function in operation and on the basis of Markov chain, the learning process is active, making the addition and deletion of the training vectors.

6. DISCUSSION AND CONCLUSION

The paper tries to highlight the areas and methodologies that presented incremental approach, thought still lot more are to be explored. Considering the various areas and the work that is taking place, it is for the researcher to have a broad view and work on domain where the process of incremental learning will help in making strong decisions. Incremental learning that is selective in terms of the datasets used at the same time adaptive and dynamic with ability to take decision accurately is what currently looked at. Accuracy that considers impact of the decisions taken should equally be taken into consideration for the same. The following table tries to summarize the domains and the methods that are in use. (Since it is impossible to cover each and every aspect of the methods and domain, there could be more applications and domain that have been remained unexplored in this paper.)

Methods/ Algorithms	Application Domain
Bayesian, GRIN, BIRCH, DBSCAN	Relational databases/ Warehouse
Neural Network, Centroid based methods	Web pages / Document layout
ART, NN, Rough sets	Document clustering
Minimum bounding boxes	Mobiles/ trajectories
Pattern matching – Neural Network	Text classification
Ensemble based methods, Learn++, ADAIN	Digital optics
Gaussian distributions, Neural Networks	Students behavior pattern
Markov chaining	Robots
Bayesian learning, Resource Allocation Network	Medical Image segmentation/ Sports video
SVM, Ensemble methods, Dynamic weighing	Optical character/ Text document
Concept drift, Ensemble methods	Weather Prediction/ Data streams

Table 1: Incremental methods and domains: A summary of survey

The point is not where one intends to work that is in supervised or unsupervised environment but to identify the domain where the learning will be effective at the same time can it be used further for forecasting. When we refer to forecast, it can be with respect to the weather, the sales of an industry, the attrition rate and so on. At this point the methods and approaches proposed aim at giving accuracy at the same time better decisions. The point of discussion here is not just the application that you want, but the reason you want to have incremental learning.

With the current work that has been done in various fields, one more factor that has to be discussed is to evolve incremental learning with a feedback mechanism. The impact of the decisions and learning from the impacts and incorporating these decisions for further learning is required. The main aspect that can be part of further research is this decision support mechanism, that itself will evolve with every new decision taken, irrespective of the application it is been employed at.

Finally to have this, it equally essential to identify what statistical methods would be used and for what purposes. Each and every existing method has its own flavor and pre-requisites and that is what should be exploited further to come up with novel methods that can be used for incremental learning.

REFERENCES

[1] Y. Lui, J. Cai, J. Yin, A. Fu, Clustering text data streams, Journal of Computer Science and Technology, 2008, pp 112-128.

[2] A. Fahim, G. Saake, A. Salem, F. Torky, M. Ramadan, K-means for spherical clusters with large variance in sizes, Journal of World Academy of Science, Engineering and Technology, 2008.

[3] F. Camastra, A. Verri, A novel kernel method for clustering, IEEE Transactions on Pattern Analysis and Machince Intelligence, Vol. 27, no.5, 2005, pp 801-805.

[4] F. Shen, H. Yu, Y. Kamiya, O. Hasegawa, An Online Incremental Semi-Supervised Learning Method, Journal of advanced Computational Intelligence and Intelligent Informatics, Vol. 14, No.6, 2010.

[5] T. Zhang, R. Ramakrishnan, M. Livny, Birch: An efficient data clustering method for very large databases, Proc. ACM SIGMOD Intl.Conference on Management of Data, 1996, pp.103-114.

[6] S. Deelers, S. Auwantanamongkol, Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with highest variance, International Journal of Electrical and Computer Science, 2007, pp 247-252.

[7] S. Young, A. Arel, T. Karnowski, D. Rose, A Fast and Stable Incremental Clustering Algorithm, Proc. of International Conference on Information Technology New Generations, 2010, pp 204-209.

[8] M. Charikar, C. Chekuri, T. Feder, R. Motwani, Incremental clustering and dynamic information retrival, Proc. of ACM symposium on Theory of Computeion, 1997, pp 626- 635.

[9] K. Hammouda, Incremental document clustering using Cluster similarity histograms, Proc. of IEEE International Conference on Web Intelligence, 2003, pp 597- 601.

[10] X. Su, Y. Lan,R. Wan, Y. Qin, A fast incremental clustering algorithm, Proc. of International Symposium on Information Processing, 2009, pp 175-178.

[11] T. Li, HIREL: An incremental clustering for relational data sets, Proc. of IEEE International Conference on Data Mining, 2008, pp 887 – 892.

[12] P. Lin, Z. Lin, B. Kuang, P. Huang, A Short Chinese Text Incremental Clustering Algorithm Based on Weighted Semantics and Naive Bayes, Journal of Computational Information Systems, 2012, pp 4257- 4268.

[13] C. Chen, S. Hwang, Y. Oyang, An Incremental hierarchical data clustering method based on gravity theory, Proc. of PAKDD, 2002, pp 237-250.

[14] M. Ester, H. Kriegel, J. Sander, M. Wimmer, X. Xu, Incremental Clustering for Mining in a Data Warehousing Environment, Proc. of Intl. Conference on very large data bases, 1998, pp 323-333.

[15] G. Shaw, Y. Xu,Enhancing an incremental clustering algorithm for web page collections, Proc. of IEEE/ACM/WIC Joint Conference on Web Intelligence and and Intelligent Agent Technology, 2009.

[16] C. Hsu, Y. Huang, Incremental clustering of mixed data based on distance hierarchy, Journal of Expert systems and Applications, 35, 2008, pp 1177 – 1185.

[17] S. Asharaf, M. Murty, S. Shevade, Rough set based incremental clustering of interval data, Pattern Recognition Letters, Vol.27 (9), 2006, pp 515-519.

[18] Z. Li, Incremental Clustering of trajectories, Computer and Information Science, Springer 2010, pp 32-46.

[19] S. Elnekava, M. Last, O. Maimon, Incremental clustering of mobile objects, Proc. of IEEE International Conference on Data Engineering, 2007, pp 585-592.

[20] S. Furao, A. Sudo, O. Hasegawa, An online incremental learning pattern -based reasoning system, Journal of Neural Networks, Elsevier, Vol. 23,(1), 2010.pp 135-143.

[21] S. Ferilli, M. Biba, T.Basile, F. Esposito, Incremental Machine learning techniques for document layout understanding, Proc. of IEEE Conference on Pattern Recognition, 2008, pp 1-4.

[22] S. Ozawa, S. Pang, N. Kasabov, Incremental Learning of chunk data for online pattern classification systems, IEEE Transactions on Neural Networks, Vo. 19 (6), 2008, pp 1061-1074.

[23] Z. Chen, L. Huang, Y. Murphey, Incremental learning for text document classification, Proc. of IEEE Conference on Neural Networks, 2007, pp 2592-2597.

[24] R. Polikar, L. Upda, S. Upda, V. Honavar, Learn ++: An incremental learning algorithm for supervised neural networks, IEEE Transactions on Systems, Man and Cybernatics, Vol.31 (4), 2001, pp 497-508.

[25] H. He, S. Chen, K. Li, X. Xu, Incremental learning from stream data, IEEE Transactions on Neural Networks, Vol.22(12), 2011, pp 1901-1914.

[26] A. Bouchachia, M. Prosseger, H. Duman, Semi supervised incremental learning, Proc. of IEEE International Conference on Fuzzy Systems, 2010 pp 1-7.

[27] R. Zhang, A. Rudnicky, A new data section principle for semi-supervised incremental learning, Computer Science department, paper 1374, 2006, http://repository.cmu.edu/compsci/1373.

[28] Z. Li, S. Watchsmuch, J. Fritsch, G. Sagerer, Semi-supervised incremental learning of manipulative tasks, Proc. of International Conference on Machine Vision Applications, 2007, pp 73-77.

[29] A. Misra, A. Sowmya, P. Compton, Incremental learning for segmentation in medical images, Proc. of IEEE Conference on Biomedical Imaging, 2006.

[30] P. Kranen, E. Muller, I. Assent, R. Krieder, T. Seidl, Incremental Learning of Medical Data for Multi-Step Patient Health Classification, Database technology for life sciences and medicine, 2010.

[31] J. Wu, B. Zhang, X. Hua, J, Zhang, A semi-supervised incremental learning framework for sports video view classification, Proc. of IEEE Conference on Multi-Media Modelling, 2006.

[32] S. Wenzel, W. Forstner, Semi supervised incremental learning of hierarchical appearance models, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol.37,2008.

[33] S. Ozawa, S. Toh, S. Abe, S. Pang, N. Kasabov, Incremental Learning for online face recognition, Proc. of IEEE Conference on Neural Networks, Vol. 5, 2005 pp 3174-3179.

[34] Z. Erdem, R. Polikar, F. Gurgen, N. Yumusak, Ensemble of SVMs for Incremental Learning, Multiple Classifier Systems, Springer Verlang,, 2005, pp 246-256.

[35] X. Yang, B. Yuan, W. Liu, Dynamic Weighting ensembles for incremental learning, Proc. of IEEE conference in pattern recognition. 2009, pp 1-5.

[36] R. Elwell, R. Polikar, Incremental Learning of Concept drift in nonstationary environments, IEEE Transactions on Neural Networks, Vol.22 (10), 2011 pp 1517- 1531.

[37] W. Khreich, E. Granger, A. Miri, R. Sabourin, A survey of techniques for incremental learning of HMM parameters, Journal of Information Science, Elsevier, 2012.

[38] O. Buffet, A. Duetch, F. Charpillet, Incremental Reinforcement Learning for designing multi-agent systems, Proc. of ACM International Conference on Autonomous Agents, 2001.

[39] E. Demidova, X. Zhou, W. Nejdl, A probabilistic scheme for keyword-based incremental query construction, IEEE Transactions on Knowledge and Data Engineering, 2012, pp 426-439.

[40] R. Roscher, W. Forestner, B. Waske, I^2VM: Incremental import vector machines, Journal of Image and Vision Computing, Elsevier, 2012.