International Journal of Data Mining Techniques and Applications
Volume: 02 Issue: 02 December 2013 Page No.71-75
ISSN: 2278-2419

Certain Investigation on Dynamic Clustering in Dynamic Datamining

S.Angel Latha Mary[1] ,Dr.K.R.Shankar Kumar[2]
[1]Associate Professor, CSE Department, Karpagam college of Engineering,Coimbatore, India
[2]Professor, ECE Department, Ranganathan Engineering College
Email : xavierangellatha@gmail.com, shanwire@gmail.com

Abstract - Clustering is the process of grouping a set of objects into classes of similar objects. Dynamic clustering comes in a new research area that is concerned about dataset with dynamic aspects. It requires updates of the clusters whenever new data records are added to the dataset and may result in a change of clustering over time. When there is a continuous update and huge amount of dynamic data, rescan the database is not possible in static data mining. But this is possible in Dynamic data mining process. This dynamic data mining occurs when the derived information is present for the purpose of analysis and the environment is dynamic, i.e. many updates occur. Since this has now been established by most researchers and they will move into solving some of the problems and the research is to concentrate on solving the problem of using data mining dynamic databases. This paper gives some investigation of existing work done in some papers related with dynamic clustering and incremental data clustering.

Keywords- Clustering, incremental data clustering dynamic clustering, dynamic data mining, Cluster evaluation, Cluster validity Index.

I. INTRODUCTION

Clustering is the process of grouping a set of objects into classes of similar objects. This data objects are similar to one another within the same cluster and dissimilar to the objects in the other clusters [1]. Clustering is a challenging problem in scalability, ability to deal with different types of attributes, discovery of cluster with arbitrary shape, domain knowledge to determine input parameters, updating new dataset and visualizing high-dimensional sparse data simultaneously [2]. Dynamic clustering comes in a new research area that is concerned about dataset with dynamic aspects. It requires updates of the clusters whenever new data records are added to the dataset and may result in a change of clustering over time. For example, the bank customer is interested in obtaining his current account status. An economic analyst can receive a lot of new articles every day and he would like to update the relevant associations based on all current articles. Recent developments of clustering systems uses dynamic data which are concerned about the clustering process in dynamically [3][4] .

II. DYNAMIC DATA MINING (DDM)

Data mining, the process of knowledge discovery in databases (KDD), is concerned with finding patterns in the raw data and finds useful information or to predict trends. Recently, the data are growing with unpredictable rate. Discovering knowledge in these data is a very expensive operation [5]. Running data mining algorithms each time when there is a change in data is a challenging problem. Therefore updating knowledge dynamically will solve these problems. Dynamic data mining is a shift from static analysis to dynamic analysis which discoverers and updates knowledge along with new updated data [3]. Dynamic data mining is very useful to obtain high quality results in the field of time series analysis, telecommunications, mobile networking, nanotechnology, physics, chemistry, biology, health care, sociology and economics [6]. When there is a continuous update and huge amount of dynamic data, rescan the database is not possible in static data mining. But this is possible in Dynamic data mining process [7].

Dynamic data mining applies data mining algorithms on dynamic database. It updates existing set of clusters dynamically. A data warehouse is not updated immediately when insertions and deletions takes place in the databases. These updates are applied to the data warehouse periodically, e.g. each night. This dynamic data mining occurs when the derived information is present for the purpose of analysis and the environment is dynamic, i.e. many updates occur [8][9]. The data mining tasks clustering, classification and summarization use this dynamic data mining. Some examples are some set of customers (clustering sales transactions), Symptoms of distinguishing disease A from disease B (classification in a medical data warehouse), Description of the typical WWW access patterns (summarization in the data warehouse of an internet provider).

Some factors for dynamic data mining.

- Whenever a database may have frequent updates.
- Database modified for every insertion or deletion.
- For every modification (insertion and deletion) the database requires rescanning again. So the time complexity is high.
- Hence the incremental clustering algorithm includes the logic for insertion and deletion of the databases as separate dynamic operation.
- After insertions and deletions to the database, the existing clusters have to be updated.
- When the data is inserted or deleted the rescanning of the whole database increases the time complexity. The system with dynamic concept will not rescan the database, but it updates the new data existing data. So it is more efficient and suitable to use in a large multidimensional dynamic database.

III. DYNAMIC CLUSTERING

International Journal of Data Mining Techniques and Applications
Volume: 02 Issue: 02 December 2013 Page No.71-75
ISSN: 2278-2419

Clustering and visualizing high dimensional dynamic data is a challenging problem in the data mining. Most of the existing clustering algorithms are based on the static statistical relationship among data. In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [2]. The databases will dynamically change due to frequent insertions and deletions which changes clustering structure over time. Completely reapplying the clustering algorithm to detect the changes in the clustering structure and update the uncovered data patterns following such deletions and insertions is very expensive for large high dimensional fast changing dataset. Dynamic clustering is a mechanism to adopt and discover clusters where a data set is updated periodically through insertions and deletions. There are many applications such as incremental data mining in data warehousing applications, sensor network which relies on dynamic data clustering algorithms.

IV. EVALUATING THE QUALITY OF THE RESULT CLUSTERS

The choice of number of the clusters is an important sub problem of clustering. Since it needs a priori knowledge in general and the vector dimensions are often higher than two, which do not have visually apparent clusters. The solution of this problem directly affects the quality of the result. If the number of clusters is too small, different objects in data will not be separated. Moreover, if this estimated number is too large, relatively regions may be separated into a number of smaller regions [10]. Both of these situations are to be avoided. This problem is known as the cluster validation problem. The aim is to estimate the number of clusters during the clustering process. The basic idea is the evaluation of a clustering structure by generating several clustering for various number of clusters and compare them against with some evaluation criteria. The procedure of evaluating the results of a clustering algorithm is known as c*luster validity.* In general terms there are three approaches to investigate cluster validity. The first is based on *external criteria.* This implies that the results of a clustering algorithm is evaluated based on a pre-specified structure, which is imposed on a data set and reflects user intuition about the clustering structure of the data set [11]. The indices *R*and Statistic, Jaccard Coefficient, Fowlkes–Mallows index, *Hubert's* statistic are used to measure the degree of similarity based on external criteria.

The second approach is based on *internal criteria.* In this case the clustering results are evaluated in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix). The following methods can be used to assess the quality clustering algorithms based on internal criterion: Davies–Bouldin index, Dunn index and F-measure. The third approach of clustering validity is based on *relative criteria.* Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes resulting by the same algorithm but with different input parameter values. The two first approaches are based on statistical tests and their major drawback is their high computational cost. The third approach aims at finding the best clustering scheme that a clustering algorithm can define under certain assumptions and parameters [11].

V. WORKS RELATED WITH DYNAMIC CLUSTERING

The most perspective direction is based on the attempts to model the work of human brain, which is highly complex, nonlinear and parallel information processing system. Complex cortex structure is modeled and formed by artificial neuron lattices, which are joined by great amount of interlinks. This global link of simple neurons provides their collective behavior. Each neuron carries out the role of a processor. That's why neuron network structure is the most appropriate base for parallel computing. There is no need to prepare data (in neural network input data is already parallelized). For parallel computing to work correctly software should be able to partition its work and data it operates on over hundreds of processors. High speed and with the same time high quality solution of most various complicated problems can be received by means of microsystem's collective behavior property. The main idea of self organization is in distributed character of data processing, when one element dynamics means nothing, but at the same time group dynamics define macroscopic unique state of the whole system, that allows this system to reveal capabilities for adaptation, learning, data mining and as one of the results high computation effectiveness [5].Advances in experimental brain science give evidence to the hypothesis that cognition, memory, attention processes are the results of cooperative chaotic dynamics of brain cortex elements (neurons). Thus the design of artificial dynamic neural networks on the base of neurobiological prototype seems to be the right direction of the search for innovative clustering techniques. Computer science development predetermined promising possibilities of computer modeling. It became possible to study complex nonlinear systems. Dynamics exponential unpredictability of chaotic systems, their extreme instability generates variety of system's possible states that can help us to describe all the multiformity of our planet.

It is assumed to be very advantageous to obtain clustering problem solution using effects produced by chaotic systems interaction. This research tries to make next step in the development of universal clustering technique. Dynamic data mining combines modern data mining techniques with modern time series analysis techniques.Saka and Nasraoui [2] designing a simultaneous clustering and visualization algorithm and the Dynamic FClust Algorithm which is based on flocks of agents as a biological metaphor. This algorithm falls within the swarm based clustering family, which is unique compared to other approaches, because its model is an ongoing swarm of agents that socially interact with each other and is therefore inherently dynamic.Yucheng Kao and Szu-Yuan Lee [12] presented a new dynamic data clustering algorithm based on K-Means and combinatorial particle swarm optimization, called KCPSO. Unlike the traditional K-Means method, KCPSO does not need a specific number of clusters given before performing the clustering process and is able to find the optimal number of clusters during the clustering process. In each iteration of KCPSO, a discrete PSO (Particle Swarm Optimization) is used to optimize the number of clusters with which the K-Means is used to find the best clustering result. KCPSO has been developed into a software system and evaluated by testing some datasets. Encouraging results show

that KCPSO is an effective algorithm for solving dynamic clustering problems.

Elghazel Haytham et al [13] presented a dynamic version for the b-coloring based clustering approach which relies only on dissimilarity matrix and cluster dominating vertices in order to cluster new data as they are added to the data collection or to rearrange a partition when an existing data is removed. A real advantage of this method is that it performs a dynamic classification that correctly satisfies the b-coloring properties and the clustering performances in terms of quality and runtime, when the number of clusters is not pre-defined and without any exception on the type of data. The results obtained over three UCI data sets have illustrated the efficiency of the algorithm to generate good results than Single-Pass and k-NN (k-nearest neighbor) algorithms. There are many interesting issues to pursue: (1) leading additional experiments on a larger medical data set where a patient stay typology is required and an inlet patient stay is regular and has to be incorporate to the typology. (2) Extending the incremental concept to add or remove simultaneously sets of instances and (3) to define some operators which permit to combine easily different clustering's constructing on different data.Ester et al [14] presented an incremental clustering algorithm based on the clustering algorithm DBSCAN for mining in a data warehousing environment which is applicable to any database containing data from a metric space, e.g., to a spatial database or to a WWW-log database. Due to the density based nature of DBSCAN, the insertion or deletion of an object affects the current clustering only in the neighborhood of that object. Thus efficient algorithms could be given for incremental insertions and deletions to an existing clustering. Based on the formal definition of clusters, this incremental algorithm yields the same result as DBSCAN. Incremental DBSCAN yields significant speed-up factors over DBSCAN even for large numbers of daily updates in a data warehouse. The authors were assumed that the parameter values Eps and MinPts of incremental DBSCAN did not change significantly when inserting and deleting objects.

Elena and Sofya [5] described about centuries humans admire animate nature and accessories applied by life creatures to fulfill various functions. At first it was just formal resemblance and mechanistic imitation, then along with sciences maturity the focus shifted on inner construction of living systems. However due to the complexity of a living system it is reproduced partly. Separate subsystems embody limited set of functions and principals. Just independently showed up artificial neural networks (attempts to mimic neural system), genetic algorithms (data transfer by means of inheritance), artificial immune systems (partial reproduction of immune system), evolutionary modeling (imitation of evolution development principals). The idea of natural self-organization within individuals is the basis for swarm and ant colony technologies. It is important to note that nearly all mentioned technologies deal with distributed parallel data processing thanks to numerous simple processing units comprised into self-organized networks that adapt to ever-changing environment (input information).Of course there exist substantial peculiarities in the types of local cooperation and global behavior mechanisms predetermined by system's goal (as it is well-known systems demonstrate not only interconnectivity of elements but their ability to serve one purpose). Evolution of society, new computer technologies have in common idea of small worlds modeling. Communities of various natures (interests clubs, computer clusters, marketing networks, etc.) speak up for strong local linkage of units and weak connectivity outward nearest neighbors (nodes of the net).

Recent research on brain activities gives evidence for its cluster organization. So generally the small world models reflect both animate nature and abiocoen. Originally the notion bio-inspired comprised problem solving approaches borrowed from living systems but nowadays it is understood more widely. Results in physics in the field of chaos theory and nonlinear dynamics contribute greatly to bio-inspired methodology as soon as nonlinear chaotic models find their application in data mining - first and fore most bio-inspired scientific area. It proposes to classify bio inspired methods on different issues:

a. Structure and connection: neural networks (macro level) and artificial immune systems (micro level);
b. Collective behavior: ant-based networks, swarm methods, multi agent systems, small- world networks;
c. Evolution and selection: genetic algorithm, evolutionary programming and evolutionary modeling, evolutionary computations;
d. Linguistics: fuzzy logic.

To step forward with generalization one can note that nearly all mentioned methods realize collective data processing through adaptation to external environment. Exception is fuzzy logic more relative to classical mathematics (interval logic reflects the diversity of natural language descriptions) Though bio inspired methods are applied to solve a wide set of problems it focus on clustering problem as the most complex and resource consuming. The division of input set of objects into subsets (mainly non-overlapping) in most cases is interpreted as optimization task with goal function determined by inter and inner cluster distances. This approach obliges the user to give them a priori information about priorities: what is of most importance - compactness of clusters and their diversity in feature space or inner cluster density and small number of clusters. The formalization process of clustering problems in terms of optimization procedures is one of the edge one in data mining.

Recent modifications of bio inspired methods are developed as heuristics. It desires to enlarge the abilities of intellectual systems a separate knowledge domain within artificial intelligence field revealed.Soft computing (SC) considers various combinations of bio-inspired methods. As a result there appeared such hybrid methods like: neural fuzzy methods, genetic algorithms with elements of fuzzy logic (FL), hybrid comprised by genetic algorithms (GA) and neural networks (NN); fuzzy logic with genetic algorithm constituent, fuzzy systems with neural network constituent, etc. One of the main ideas of such combinations is to obtain flexible tool that allow to solve complex problems and to compensate drawbacks of one approach by means of cooperation with another.For example, FL and NN combination provides learning abilities and at the same time formalized knowledge can be represented due to fuzzy logic element. Fuzzy logic is applied as soon as to

add some flexibility to a data mining technique. One of the main drawbacks of all fuzzy systems are absence of learning capabilities, absence of parallel distributing processing and what is more critical the rely on expert's opinions when membership functions are tuned. In advance to input parameters sensitivity almost all methods suffer from dimension curse and remain to be resource consuming. The efficiency of these methods depends greatly on the parallel processing hardware that simulate processing units: neurons of neural networks, lymphocyte in artificial immune systems, ants and swarms, agents in multi-agent systems, nodes in small-world networks, chromosomes in genetic algorithms, genetic programming and genetic modeling.

The benefit from synergetic effects considers not only collective dynamics but also physical and chemical nature of construction elements - nonlinear oscillators with chaotic dynamics. As it is shown in numerous investigations on nonlinear dynamics: the more is the problem complexity. The more complex should be the system dynamics. All over the world investigations on molecular level take place to get new materials, to find new medicine, to solve pattern recognition problem etc. Most of them consume knowledge from adjacent disciplines: biology, chemistry, math, informatics, nonlinear dynamics and synergetics.

VI. RELATED WORKS

Due to the continuous, unbounded, and high speed characteristics of dynamic data, there is a huge amount of data and there is not enough time to rescan the whole database or perform a rescan as in traditional data mining algorithms whenever an update occurs [4]. Ganti et al [15] examine mining of data streams. A block evolution model is introduced where a data set is updated periodically through insertions and deletions. In this model the data set consists of conceptually infinite sequence of data blocks D1, D2, ... that arrive at times 1, 2, ... where each block has a set of records. The authors highlight two challenges in mining evolving blocks of data: change detection and data mining model maintenance. In change detection, the differences between two data blocks are determined. Next, a data mining model should be maintained under the insertions and deletions of blocks of the data according to a specified data span and block selection sequence.Crespoa and Weberb [3] presented a methodology for dynamic data mining using fuzzy clustering that assigns static objects to dynamic classes. Changes that they have studied are movement, creation and elimination of classes and any of their combination. Once a data mining system is installed and is being used in daily operations, the user has to be concerned with the system's future performance because the extracted knowledge is based on past behavior of the analyzed objects If future performance is very similar to past performance (e.g. if company customers files do not change their files over time) using the initial data mining system could be justified. If, however, performance changes over time (e.g. if hospital patients do not change their files over time), the continued use of the early system could lead to an unsuitable results and (as an effect) to an unacceptable decisions based on these results. Here dynamic data mining could be extremely helpful in making the right decision in the right time and affects the efficiency of the decision.

There are three strategies if a user is to keep applying his/her data mining system in a changing environment.

- The user can neglect changes in the environment and keep on applying the initial system without any further updates. It has the advantage of being "computationally cheap" since no update to data mining system is performed. Also it does not require changes in subsequent processes. Its disadvantage is that current updates could not be detected.
- Every certain period of time, depending on the application, a new system is developed using all the available data. The advantage in this case is the user has always a system "up-to-date" due to the use of current data. Disadvantages of this strategy are the computational costs of creating a new system every time from scratch.
- Based on the initial system and "new data" an update of data is performed. This will be shown to be available method in this dissertation.

In the area of data mining various methods have been developed in order to find useful information patterns within data. Among the most important methods are association rules, clustering and decision trees methods. For each of the above data mining methods, updating has different aspects and some updating approaches have been proposed:Decision trees: Various techniques for incremental learning and tree restructuring.Neural networks: Updating is often used in the sense of re-learning or improving the net's performance by learning with new examples presented to the network.Clustering: Chung and Mcleod describes in more detailed approaches for dynamic data mining using clustering techniques.Association rules :Raghavan et al developed a system for dynamic data mining for association rules. Chung and Mcleod proposed mining framework that supports the identification of useful patterns based on incremental data clustering, they focused their attention on news stream mining, they presented a sophisticated incremental hierarchical document clustering algorithm using a neighbourhood search.
Reigrotzki et al (2001) presented the application of several process control-related methods applied in the context of monitoring and controlling data quality in financial databases. They showed that the quality control process can be considered as a classical control loop which can be measured via application of quality tests which exploit data redundancy defined by Meta information or extracted from data by statistical models. Appropriate processing and visualization of the tests results enable Human or automatic detection and diagnosis of data quality problems. Moreover, the model-based methods give an insight into business-related information contained in the data. The methods have been applied to the data quality monitoring of a real financial database at a customer site, delivering business benefits, such as improvements of the modeling quality, a reduction in the number of the modeling cycles, and better data understanding.
These benefits in turn lead to financial savings.In many situations, new information is more important than old information, such as in publication database, stock transactions, grocery markets, or web-log records. Consequently, a frequent itemset in the dynamic database is also important even if it is infrequent in the updated database.

Incremental clustering is the process of updating an existing set of clusters incrementally rather than mining them from the scratch on each database update. A brief overview of work done on incremental clustering algorithms is given next. COBWEB was proposed by Fisher [16]. It is an incremental clustering algorithm that builds taxonomy of clusters without having a pre-defined number of clusters. Gennary et al [17] proposed CLASSIT which associates normal distributions with cluster nodes. The main drawback of both COBWEB and CLASSIT is that they results in highly unbalanced trees.

Charikar et al [18] introduced new deterministic and randomized incremental clustering algorithms while trying to minimize the maximum diameters of the clusters. The diameter of a cluster is its maximum distance among its points and is used in the restructuring process of the clusters. When a new point arrives, it is either assigned to one of the current clusters or it initializes its own cluster while two existing clusters are combined into one. Ester et al [14] presented Incremental DBSCAN suitable for mining in a data warehousing environment. Incremental DBSCAN is based on the DBSCAN algorithm which is a density based clustering algorithm. It uses R* Tree as an index structure to perform region queries. Due to its density based qualities, in Incremental DBSCAN the effects of inserting and deleting objects are limited only to the neighborhood of these objects. Incremental DBSCAN requires only a distance function and is applicable to any data set from a metric space. However, the proposed method does not address the problem of changing point densities over time, which would require adapting the input parameters for Incremental DBSCAN over time. Another limitation of the algorithm is that it adds or deletes one data point at a time. An incremental clustering algorithm based on SWARM intelligence is given in Chen and Meng[19].

VII. CONCLUSION

The ability to solve complex clustering problems in terms of oscillations clustering language in future research can be extended by dynamic inputs or at the beginning of the road, as there are many aspects of data mining that have not been tested. Up to date most of the data mining projects have been dealing with verifying the actual data mining concepts. Since this has now been established most researchers will move into solving some of the problems and in this case, the research is to concentrate on solving the problem of using data mining dynamic databases. This chapter gives existing work done in some papers related with dynamic clustering and incremental data clustering.

REFERENCES

[1] Han, J. and Kamber, M. “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
[2] Saka, E. and Nasraoui, O. “On dynamic data clustering and visualization using swarm intelligence”, In Data Engineering Workshops (ICDEW), 26th International Conference on IEEE, pp. 337-340, 2010.
[3] Crespoa, F. and Weber, R. “A methodology for dynamic data mining based on fuzzy clustering”, Fuzzy Sets and Systems, Elsevier, Vol. 150, pp. 267-284, 2005.
[4] Hebah, H.O. Naseredd, “Stream Data Mining”, in International Journal of Web Applications, Vol. 1, No. 4, pp. 183-190, 2009.
[5] Elena N. Benderskaya and Sofya V. Zhukova, “Dynamic Data Mining: Synergy of Bio-Inspired Clustering Methods”, Knowledge-Oriented Applications in Data Mining.
[6] Xu, R., Wunsh, D. II, “Survey of clustering algorithms”, IEEE Trans. on Neural Networks, Vol. 15, pp. 645-678, 2005.
[7] Zhang, J., Tianrui Li, Da Ruan and Dun Liu, “Neighborhood Rough Sets for Dynamic Data Mining”, International Journal of Intelligent Systems, Wiley Periodicals, Inc., Vol. 27, pp. 317-342, 2012.
[8] Sanjay Chakraborty and Nagwani, N.K. “Analysis and Study of Incremental DBSCAN Clustering Algorithm”, International Journal of Enterprise Computing and Business Systems, Vol. 1, Issue 2, pp. 2230-8849, 2011.
[9] Sanjay Chakraborty, Nagwani, N.K. and Lopamudra Dey, “Performance Comparison of Incremental K-Means and Incremental DBSCAN Algorithms”, International Journal of Computer Applications (0975 – 8887) , Vol. 27, No. 11, 2011.
[10] Svetlana Cherednichenko, Outlier Detection in Clustering, 2005.
[11] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. “Cluster Validity Methods: part I”, In Proceedings of the ACM SIGMOD International Conference on Management of Data, Vol. 31, Issue 2, pp. 40-45, 2002.
[12] Yucheng Kao and Szu-Yuan Lee, “Combining K-Means and particle swarm optimization for dynamic data clustering problems”, This paper appears in: Intelligent Computing and Intelligent Systems, ICIS 2009, IEEE International Conference on, Vol. 1, pp. 757-761, 2009.
[13] Elghazel Haytham, Hamamache Kheddouci, Véronique Deslandres and Alain Dussauchoy, “A Partially Dynamic Clustering Algorithm for Data Insertion and Removal”, Discovery Science ,Lecture Notes in Computer Science, Vol. 4755, pp. 78-90, 2007.
[14] Ester, M. and Wittmann, R. “Incremental Generalization for Mining in a Data Warehousing Environment”, Proc. 6th Int. Conf. on Extending Database Technology, Valencia, Spain, 1998, in: Lecture Notes in Computer Science, Springer, Vol. 1377, pp. 135-152, 1998.
[15] Ganti, V., Gehrke, J., Ramakrishnan, R. and Loh, W. “Mining data streams under block evolution”, In ACM SIGKDD Explorations, Vol. 3(2), pp. 1-10, 2002.
[16] Fisher, “Knowledge acquisition via incremental conceptual clustering”, Machine Learning, Vol. 2, pp. 139-172, 1987.
[17] Gennary, J., Langley, P. and Fisher, D. “Model of Incremental Concept Formation”, Artificial Intelligence Journal, Vol. 40, pp. 11-61, 1989.
[18] Charikar, M., Chekuri, C., Feder, T. and Motwani, R. “Incremental clustering and dynamic information retrieval”, 29th Symposium on Theory of Computing, pp. 626-635, 1997.
[19] Chen, Z. and Meng, Q.C. “An incremental clustering algorithm based on SWARM intelligence theory”, Proc. of the 3rd Int. Conf. on Machine Learning and Cybernetics, Shanghai, 26-29 August, 2004