TRANSFORMING TECHNOLOGY
EXPLORING KNOWLEDGE

Volume 1 Issue 4
October 2013 - December 2013

ISSN: 2322-0821

INTERNATIONAL
RESEARCH JOURNAL
OF ENGINEERING
& APPLIED SCIENCES

www.irjeas.com

Dynamic Data Mining – A Survey to New Approach

Anurag Shrivastava[1]

[1]*Anurag Shrivastava, M. Tech Scholar, Dept. of Information Technology, KSOU, Mysore, India*

Abstract **- "Data mining refers to extracting or "mining" knowledge from large amounts of data." Data mining should have been more appropriately named knowledge mining from data, so it is also popularly referred to as knowledge discovery in databases (KDD). Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Its applications can use a variety of parameters to examine the data. As the databases tend to be large and dynamic, the classical and customary approach of generating a fixed set of data to analyze often misses the opportunity to handle uncertainty and to respond to the range of problem characteristics that can be captured through the use of simulation. A more responsive and effective approach involves conducting data mining for dynamic processes by analyzing the processes in action, as they are unfolding. This paper provides the insight into Dynamic Data Mining that uses simulation optimization process that provides the means to produce the classification. The simulation optimization process may be viewed as a set of rules, but the rules are more complex than customary rules, and they have an iterative character. Dynamic association rule mining approach which is applicable on dynamic data or data mining method with time series method is also introduced.**

Keywords **-** Data Mining, Dynamic Data Mining (DDM), Association rule mining

I. INTRODUCTION

Data mining is one of the rapidly growing fields in the computer industry. The data base system industry has an evolutionary path in the development of the functionalities like Data collection and database creation, data management and advanced data analysis which includes Data mining and data warehousing. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, etc. Data mining tasks have the many categories like: Class descriptions, Association rule mining, Classification, Clustering, Outlier detection and analysis, Evolution analysis.

The architecture of a typical data mining system may have the following major components [1]:

(a) Database, data warehouse, or other information repository.
(b) Their server which is responsible for fetching the relevant data based on the user's request for mining.
(c) Knowledge base which is used to guide the search, or evaluate the significance of resulting patterns.
(d) data mining engine which consists of a set of functional modules for tasks.
(e) pattern evaluation module which interacts with the data mining modules to search interesting patterns.
(f) graphical user interface which interfaces between users and the data mining system.

Dynamic Data Mining (DDM) combines modern data mining techniques with modern time series analysis techniques. Standard time series analysis deals with the type of time based sequences of data points and forecasting of future data points that can be used for forecasting data. Standard data mining, on the other hand, incorporates many methods to handle large numbers of input variables, but it is typically not suitable for temporal data. The Dynamic Data Mining technology combines the best of both worlds: it uses state of the art nonlinear time series analysis and prediction techniques with state of the art data mining techniques for handling and analyzing large amounts of input data [2].

The remaining of the paper is organized as follows. The Section II briefly reviews the related literature. In Section III the overview and need of Dynamic Data Mining is illustrated. The dynamic association rule mining approach which is applicable on dynamic data is introduced in Section IV. Section V, concludes the paper.

II. LITERATURE REVIEW

Association rule mining aims to explore large transaction databases for applying association rules. Classical Association Rule Mining model assumes that all items have the same significance without taking their weight into account. It also ignores the difference between the transactions and importance of each and every itemsets. But, the Weighted Association Rule Mining does not work on databases with only binary attributes. It makes use of the importance of each itemset and transaction.

WEIGHTED ASSOCIATION RULE MINING ON DYNAMIC CONTENT

WARM [3] proposes assignment of weight to each item to reflect their importance to the user. The weights may correspond to special promotions on some products, or the profitability of different items. The research work focused on a weight assignment based on a directed graph where nodes denote items and links represent association rules. This research then uses enhanced HITS algorithm whose generalized version is applied to the graph to rank the items, where all nodes and links are allowed to have weights. The method in [3] uses development of online eigenvector calculation method that can compute the results of mutual reinforcement voting in case of frequent updates. For Example in Share Market Shares price may go down or up. So we need to carefully observe the market and our association rule mining has to produce the items that have undergone frequent changes. These are done by estimating the upper bound of change and postponing of the updates whenever possible.

The results proved that enhanced algorithm is more efficient than the original HITS under the context of dynamic data as the construction of static graph is replaced by the construction of dynamic graph by finding out the Eigen gap. Also an online eigenvector calculation method estimating the upper bound is applied in case of frequent updates whenever possible.

DYNAMIC DATA MINING BASED ON FUZZY CLUSTERING

Dynamic data mining is increasingly attracting attention from the respective research community. In [4] authors present a methodology for dynamic data mining based on fuzzy clustering, which allows updates of the underlying classifier. Using the implementation of the proposed system its benefits are shown in two application areas: customer segmentation and transaction management. In it fuzzy c-means along with other fuzzy or possibility clustering technique is used. The proposed methodology applied for dynamic customer segmentation and dynamic task state identification using real-world data. The obtained results provide updated class structures and even more important potential insights a user gets by analyzing changes in his or her application domain.

At the same time author stated that more work is necessary in order to understand better the potentials

and limitations of dynamic data mining using clustering techniques and the parameters of methodology needs further investigation. An automatic adaptation of parameters in a given application domain by employing an appropriate learning technique would be an interesting additional feature. Applying the proposed methodology in order to solve different problems would also give further information regarding future research.

INTEGRATING DYNAMIC DATA MINING WITH SIMULATION OPTIMIZATION

Authors in [5] introduced a simulation optimization approach that is effective in guiding the search for optimal values of input parameters to a simulation model. The proposed approach, which includes enhanced data mining methodology and state-of-the-art optimization technology, is applicable to settings in which a large amount of data must be analyzed in order to discover relevant relationships. The suggested approach makes use of optimization technology for data mining as well as for optimizing the underlying simulation model itself. A market research application embodying agent based simulation is used to illustrate the approach in [5]. The relationship among the input variables and their effect on the performance of the system that is being modeled is studied. As an alternative to extensive trial and error for an optimal configuration of the system, a practical approach that includes a dynamic data mining module to identify the relevant inputs and discover the nature of their relationships to the performance of the system.

The dynamic data mining model makes use of information learned during the optimization process to distinguish good-quality solutions from bad, so that only promising solutions need to be evaluated during future iterations. Underlying and supporting this module is an optimization engine that makes use of state-of-the-art algorithms to aid in the data mining and to ultimately guide the search for optimal configurations for the simulation model. A wide range of applications can benefit from the proposed approach, includes business process management, portfolio management, project life-cycle management, health care, prevention and control of epidemics, bioterrorism detection and control, vaccination benefits assessment, clinical trial simulations, and numerous applications in the social sciences, physical sciences, and materials sciences. It is shown through a detailed market research example that the approach can be used to find the best scenario with respect to a user's desired performance measures. Simulation optimization is growing and increasingly rich field of application in data mining that is used to represent and respond to complex relationships in ways that cannot be achieved by alternative approaches.

DYNAMIC DATA MINING ON A SELF-ADAPTIVE MULTIPLE EXPRESSIONS

More recent advancements in collecting massive evolving data streams created a crucial need for dynamic data mining. In [6], genetic algorithm based on a new representation mechanism that allows several phenotypes to be simultaneously expressed to different degrees in the same chromosome is presented. The gradual multiple expression mechanism offers a simple model for a multiplied representation with self-adaptive dominance, co-dominance and incomplete dominance. Based on this model, a data mining approach that considers the data as a reflection of a dynamic environment, and investigate a new evolutionary approach based on continuously mining non-stationary data sources that do not fit in main memory is proposed.

Preliminary experiments are performed on real Web click stream data. For many data mining tasks, the subjective objective functions and/or dissimilarity measure may be non-differentiable. Evolutionary techniques can handle a vast array of subjective, even non-metric dissimilarities. In it a new framework that considers evolving data, such as in the context of mining stream data, as a reflection of a dynamic environment and therefore requires dynamic learning. This approach can be generalized to mining huge data sets that do not fit in main memory. Massive data sets can be mined in parts that can fit in the memory buffer, while the evolutionary search adapts to the changing trends automatically. The approach in

[6] is compared against other standard dynamic optimization strategies.

DYNAMIC DATA MINING: EXPLORING LARGE RULE SPACES

As pruning techniques either do not sufficiently reduce the space of rules, or they are overly restrictive. So, authors in [7] propose a new solution to this problem, called Dynamic Data Mining (DDM). DDM foregoes the completeness offered by traditional techniques based on downward-closed measures in favor of the ability to drill deep into the space of rules and provide the user with a better view of the structure present in a data set. Instead of using a downward-closed measure such as reducing rule space, DDM uses a user defined measure called weight, which is not restricted to be downward closed. The exploration is guided by a heuristic called the Heavy Edge Property. The key to sampling the space of itemsets is the Heavy Edge Property, which hypothesizes that interesting itemsets are likely to be near to other interesting itemsets.

The system incorporates user feedback by allowing weight to be redefined dynamically. Then the system is tested on a particularly difficult data set. Furthermore, dynamic data mining allows for quick user feedback. Instead of waiting for a long data mining job to run, the user can get results as they happen from the system and adjust the mining based on those results. It is mentioned that the dynamic data mining adds several significant enhancements to the process of data mining.

III. DYNAMIC DATA MINING

OVERVIEW

Classical data mining, assumes the existence of a collection of data in some repository, such as a computer database, whose elements are often referred to as points / vectors in some sort of space.
The data can be generated by historical records or by various types of deterministic or stochastic processes, which may in some cases be viewed as a dynamic source for the data. However, in classical data mining, the data elements once generated constitute a static collection that is analyzed by taking its elements as fixed. In other words, the data is not modified or updated, which is in contrast to the dynamic case, in which we may update the data by incorporating new features or factors as a result of information gained during the optimization process. Thus a more responsive and effective approach involves conducting data mining for dynamic processes by analyzing the processes in action, as they are unfolding. In case of Dynamic data mining a procedure is activated at certain intervals during the optimization process in order to make use of information obtained during that process, with the goal of speeding the search for optimal solutions. To accommodate data that changes over time, successive snapshots or samples are taken using updated forecasts or other information.

The gains of being able to handle data mining considerations in this dynamic fashion are substantial. For example, the classical and usual approach of generating a fixed set of data to analyze often misses the opportunity to handle vagueness and to respond to the range of problem characteristics that can be captured through the use of simulation. Moreover, the methods designed for analyzing fixed data sets have a significantly different characteristics than those designed for simulation optimization.

Dynamic data mining has several other fine properties. First, it can work on continuous data streams. This means that it is possible to mine data from live data sources or to mine data which one cannot afford to store on disk. For example, it is possible to mine the entire World Wide Web without storing a single web page. Another use of this property is mining where quick reaction to new relationships is important such as data feeds from stock and currency exchanges. In summary, dynamic data mining is a very effective tool for knowledge exploration, particularly when it is not feasible to completely explore the space of rules. It provides results quickly independent of the size of the data set and refines them as it progresses [7].

International Research Journal of Engineering & Applied Sciences
www.irjeas.com, ISSN: 2322-0821, Volume 1, Issue 4, Oct. 2013 – Dec. 2013, Page 43-49

NEED

As Databases tend to be large and dynamic thus their contents usually do change; new information might need to be inserted, current data might need to be updated and/or deleted. Due to the continuous, unrestrained, and high speed characteristics of dynamic data, there is a huge amount of data in both online and offline data streams. In traditional data mining algorithms, there is unavailability of approaches and time to rescan the whole database or perform a rescan as whenever an update occurs.

In view of importance and effects of dynamic data mining on decision making it is vital to put forward a solution to run data mining without the need to restart the whole process every time there are changes on the data being used. In other words the running process should focus solely on the amendments taking into consideration that the mining run is held constant. The problem with this, from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information.

Also the learning system has to be time sensitive as some data values vary over time and the discovery system is affected by the correctness of the data. So, in this regard many researchers and developers have proposed various process models to guide the user through a sequence of steps that will lead to good results. Some of these assumed that this is possible with Dynamic data mining process.

IV. ASSOCIATION RULE MINING

One of the most important approaches to detect relationships or associations between specific values of categorical variables in large data sets is mining association rules.

Association mining that discovers dependencies among values of an attribute. The problem of association mining, also referred to as the market basket problem, is formally introduced in Ref. [8] and can be stated as follows.

Let $I=\{i_1, i_2,. . ., i_m\}$ be a set of items and D be a set of transactions, where each transaction T (a data case) is a set of items so that $T \subseteq I$. An association rule is an implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ holds in the transaction set T with confidence c, if c% of transactions in T that support X also support Y. The rule has support s in T if s% of the transactions in T contains $X \cup Y$. Given a set of transactions D (the database), the problem of mining association rules is to discover all association rules that have support and confidence greater than the user-specified minimum support (called minsup) and minimum confidence (called minconf).

The key element that makes association-rule mining practical is MINSUP. It is used to prune the search space and to limit the number of rules generated.

DYNAMIC ASSOCIATION RULE MINING APPROACH

let the transaction set T be divided into n transaction subsets Ti 's, $1 \leq i \leq n$. S is a large itemset if $\sum_{i=1}^{n} F_i * \delta_i \geq 0$, where Fi is the number of transactions in Ti and δi = SUPPORTi(S) - MINSUP, $1 \leq i \leq n$. $-MINSUP \leq \delta i \leq 1-MINSUP$, $1 \leq i \leq n$.

For those cases where $\sum_{i=1}^{n} F_i * \delta_i < 0$, there are two options, either

- discard S as a large itemset (a small itemset with no history record maintained), or
- Keep it for future calculations (a small itemset with history record maintained). In this case, we are not going to report it as a large itemset, but its $\sum_{i=1}^{n} F_i * \delta_i$ formula will be maintained and checked through the future intervals.

For $\sum_{i=1}^{n} F_i * \delta_i < 0$, the two options described above could be combined into a single decision rule that says discard S if

$$\sum_{i=k}^{n} F_i * (MINSUP + \delta_i) \Big/ \sum_{i=k}^{n} F_i \geq MINSUP/\alpha$$

where $1 \leq \alpha < \infty$, and k≥1.

$\alpha = 1$ Discard S from the set of a large itemsets (it becomes a small itemset with no history record)

$\alpha \to \infty$ Keep it for future calculations (it becomes a small itemset with a history record)

The value of α determines how much history information would be carried. This history information along with the calculated values of locality can be used to

- Determine the significance or the importance of the generated emerged-large itemsets.
- Determine the significance or the importance of the generated declined-large itemsets.
- Generate large itemsets with less SUPPORT values without having to rerun the mining procedure again.

The choice of which value of α to choose is the essence of our approach. If the value of α is chosen to be near the value of 1, we will have less declined-large itemsets and more emerged-large itemsets, and those emerged-large itemsets are more to be occurred near the latest interval episodes. For those cases where the value of α is chosen to be far from the value of 1, we will have more declined-large itemsets and less emerged-large itemsets, and those emerged-large itemsets are more to be large itemsets in the apriori-like approach.

In this section, we introduce the notions of declined-large itemset, emerged-large itemset, and locality. Let S be a large itemset a emerged-large itemset in a transaction subset Tl , $l \geq 1$. S is called a declined-large itemset in transaction subset Tn , n > l, if

$$MINSUP > \sum_{i=1}^{m} F_i^* (MINSUP + \delta_j) \Big/ \sum_{i=k}^{m} F_j \geq MINSUP/\alpha$$

for all l < m ≤ n, where $1 \leq k \leq m$, and $1 \leq \alpha < \infty$, S is called a emerged-large itemset in transaction subset Tn , n > 1, if S was a small itemset in transaction subset Tn-1 and $F_n * \delta_n \geq 0$, or S was a declined-large itemset in transaction subset Tn-1, n > 1, and $\sum_{i=k}^{n} F_i * \delta_i \geq 0, k \geq 1$. For an itemset S and a transaction subset Tn, locality(S) is defined as the ratio of the total size of those transaction subsets where S is either a large itemset or a emerged large itemset to the total size of transaction subsets Ti , $1 \leq i \leq n$.

$$\sum_{\forall i \text{ s.t.S is a large or a emerged-large itemset}} F_i \Big/ \sum_{i=1}^{n} F_i$$

Clearly, the locality(S) =1 for all large itemsets S. The dynamic data mining approach generates three sets of itemsets, large itemsets, that satisfy the rule $\sum_{i=1}^{n} F_i * \delta_i \geq 0,$ where n is the number of intervals carried out by the dynamic data mining approach

- Declined-large itemsets, that was large at previous intervals and still maintaining the rule for some value α.

$$MINSUP > \sum_{i=1}^{m} F_i^* (MINSUP + \delta_j) / \sum_{i=k}^{m} F_j \geq MINSUP/\alpha$$

- Emerged-large itemsets, that was either small itemsets and at a transaction subset Tk they satisfied the rule $F_i * \delta_i \geq 0.$ and still satisfy the rule $\sum_{i=1}^{n} F_i * \delta_i \geq 0$.
- Or they were declined-large itemsets, and at a transaction subset Tm they satisfied the rule.

$\sum_{i=1}^{n} F_i * \delta_i \geq 0$ and still satisfy the rule $\sum_{i=1}^{n} F_i * \delta_i \geq 0$.

V. CONCLUSION

This paper briefly reviews the literature related to Dynamic Data Mining. We have also defined Dynamic Data Mining approach along with a significant potential usage. The DDM approach performs periodically the data mining process on data updates during a current occurrence and uses that knowledge captured in the previous occurrence to produce data mining rules. We have tried to put up that the Dynamic Data Mining approach is much efficient than classical and customary approach of generating a fixed set of data. The association rule mining and dynamic association rule mining approach is also discussed briefly. This work will significantly support a large domain of different data mining applications, such as, web site access analysis for improvements in e-commerce advertising, fraud detection, screening and investigation, retail site or product analysis, and customer segmentation etc.

VI. REFERENCES

[1]. Hebah H. O. Nasereddin," Stream Data Mining", Department of computer Information system Faculty of IT Amman Arab University for Graduate Studies Amman – Jordan, 18 August 2009.

[2] Joseph Kielman,"The real-time nature and value of homeland security information", CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management, 2006.

[3] Fernando Crespoa, Richard Weberb, "A methodology for dynamic data mining based on fuzzy clustering", Fuzzy Sets and Systems 150 (2005) 267–284, October 2002.

[4] Better," Advances in analytics: Integrating dynamic data mining with simulation optimization "IBM Journal of Research and Development, May 2007

[5] P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal, Lalit M. Patnaik," Dynamic Association Rule Mining using Genetic Algorithms", Intelligent Data Analysis, Volume 9, Pages439-453, Number 5/2005.

[6] George Rzevski, Peter Skobelev, Igor Minakov, Semen Volman," Dynamic Pattern Discovery using Multi-Agent Technology", Published in Proceedings of the 6th WSEAS International Conference on Telecommunications and Informatics, Dallas, USA, , pp 75-81, March, 2007.

[7] Weiwei He, Shuanghua Yang , Lili Yang, "Real-time data mining methodology and emergency knowledge discovery in wireless sensor networks", Department of Computer Science, Loughborough University, 2010.

[8] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, Proceedings of the 20th Very Large DataBases Conference (VLDB'94), Santiago de Chile, Chile, 1994, pp. 487– 499.