KNOWLEDGE DISCOVERY and SAMPLING TECHNIQUES with DATA MINING for IDENTIFYING TRENDS in DATA SETS

[#1]Prof. Punam V. Khandar, [*2]Prof. Sugandha V. Dani
[#1]*Dept. of M.C.A., Priyadarshini College of Engg., Nagpur, R.T.M.N.U., Maharashtra, INDIA.*
[*2]*Dept. of M.C.A., Priyadarshini College of Engg., Nagpur, R.T.M.N.U., Maharashtra, INDIA.*
[#1]punam.khandar@gmail.com [*2]sugandhadani@yahoo.com

***Abstract* - Complex search and data analysis can be automated using Knowledge Discovery. Extracting hidden knowledge in large volume of raw data can be accomplished by exploiting theoretical and practical issue in Data Mining techniques for Knowledge Discovery. Complex iterative and sequential steps formulate Knowledge Discovery in Databases.**

The performance of Data mining tools may be degraded due to poor training set and bulky practical data set. The Data Mining process can be optimized by focusing on the efficient method to estimate optimal training set. It is looked upon as a quite challenging problem.

In our work, we are aiming to focus on Knowledge Discovery techniques by improving approach with efficiently identifying the sampling dataset and data mining algorithm to mine the sampled data.

***Keywords*- Knowledge Discovery, Hidden Knowledge, Data Mining, Training Set, Sampling Dataset.**

I. INTRODUCTION

Over the past years, data is being collected and accumulated at a tremendous volume by various organizations. It is the need of time to have new computational techniques and tools for searching and fetching useful knowledge from the dramatically growing pile of data. This has created a challenge for the knowledge workers and hence consequently data mining has become a research area to find hidden, valid and actionable information. Data mining is part of Knowledge Discovery in Databases (KDD) for extracting trends and patterns from data [1]. Knowledge discovery in databases (KDD) area is consistently striving for such techniques and tools.

It is found that scalability of the data mining algorithms with the size of source data set is one of the greatest limitations. With the rise in size of the data set, the computation time of the algorithm rises exponentially. Hence accurate models often require the large data sets that help algorithms to discover complex structure and make accurate parameter estimates.

In this paper we view the knowledge discovery process as a set of various activities for making sense of data. The fundamental of this paper is to handle an important data mining problem to determine a reasonable upper bound of the data set size needed for building sufficiently accurate model.

II. A LOOK AT KNOWLEDGE DISCOVERY IN DATABASES

Knowledge Discovery in Databases is the process of identifying unknown, valid, actionable and ultimately understandable patterns in data. Pattern in the given data is an expression in interpreting the data or a model applicable to the subset in given data. Extracting a pattern designates fitting a model to data, finding structure from data, or in general any high-level description of a set of data. KDD Process performs selection, preprocessing, sub-sampling, and transformations on provided space elements to evaluate the products of data mining to identify the subset of patterns – nothing but "knowledge". Specifically, KDD is comprised of data preparation, search for patterns, knowledge evaluation, and refinement of the result found.[2] The discovered patterns should be unknown previously so that new trends in data are found. It must be valid too on new data with some degree of certainty. We also want patterns to be actionable so that it can be utilized further to take decisions. Finally, the patterns should be understandable, if not immediately then after some post-processing. The overall KDD process includes the evaluation and possible interpretation of the "mined" patterns to determine which patterns may be considered as "new knowledge".

The measures should be defined to evaluate the patterns. The measures of certainty/prediction accuracy and utility/cost saved due to better predictions or speed-up in response time must be applied over the result. An important feature of target data is usually

taken as an overall measure of pattern, validity, usefulness and simplicity.

Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, and anomalies, and significance structures, from large amount of data stored in databases, data warehouses, or other information repositories [3]. It is a step in the KDD process consisting of applying data analysis and discovery algorithms under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. The data mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The patterns available in given space are often infinite and actual computational constraints place certain limits on the search by a data mining algorithm.

In general, a knowledge discovery process consists of an iterative sequence, as follows:

A. Goal Identification:

Developing an understanding of the application domain

B. Data Integration:

Various heterogeneous data sources may be integrated into one.

C. Data Cleaning and Preprocessing:

Handling noisy, erroneous, missing, or irrelevant data.

D. Data Transformation:

Data are transformed into forms appropriate for mining by performing aggregation operations.

E. Data Mining:

Intelligent mining methods are applied for searching patterns of interest.

F. *Interpreting Mined Patterns*:

Identify the truly interesting patterns representing knowledge using some visualization aid.

G. Utilize Discovered Knowledge:

Documenting gained knowledge or applying into another system for further action. Compare for conflicts with previously believed knowledge.

III. THE PROBLEM DEFINED

Practical data sets are bulky enough that many of the algorithms involved in data mining may not be scaleable with it. Volume of the data set causes increased computation time. Accurate models need the use of large data sets that enables algorithms to discover complex patterns. Hence, the challenging task is to determine a reasonable upper bound of the data set size needed for building fairly accurate model. The data mining process can be optimized by estimating the optimal training set which is one of the most promising directions to resolve the problem with large data sets.

Association rule mining is a popular focal point of research in knowledge discovery research area. But the major technical problem in association rule mining is frequent item set identification. An item set is frequent if and only if it is contained by β% of the transactions. Mining association rules suffers from a number of different shortcomings including the lack of user exploration and control, and the lack of focus.

IV. THE PROPOSED WAY TO SOLUTION

In the proposed approach, the objective is to partition the problem space. The problem is divided into two phases: sample and mining. The first phase focuses on determining the appropriate dataset needed to be mined, whereas the second phase insists on effective mining algorithm selection for above captured dataset. It ultimately results in optimal mining outcome with minimal cost.

The central objective for the first step above is to choose a sample from target database which can represent the major characteristics of the entire database. While the key focus of the second step is selecting a mining method(s) that supports constraint based human centered exploratory mining of associations that enables the user to clearly specify what associations are to be mined and generate an informative set of rules with a high efficiency [4].

Sampling is the process of selecting representative which indicates the whole data set by examining a part. Sampling is needed in order to make abstraction of complex problem as well as it is used to acquire a sub set that is inferring a larger data set. It is widely accepted that a fairly modest-sized sample can sufficiently characterize a much larger population.

The virtue of the sample for the entire database is determined by two characteristics: the size and the quality of the sample. The sample size should not be too small as not to actually represent the entire data set or too large to be overloading the data mining algorithms. Also the quality sample for one problem may not be a quality sample for another problem due to the different problem definitions as per the requirements. The highest quality sample would preserve the distributions of individual variables and the relationships among variables i.e. unbiased.

There are several reasons why a sample is preferred to a complete collection [5]:

A. Helpful to work around constraints

B. Greater economy

- Sampling can reduce I/O costs.
- Data cleansing can be very time-consuming. The total cost of cleansing a sample will be much less than that of the whole database.
- Shorter time-lag: smaller number of observations [6].

C. Generalization Samples

Representative of the entire database with little (if any) information loss.

D. Greater scope

Variety of information by virtue of its flexibility and adaptability.

V. FRAMEWORK SUGGESTED

In our work we aim in building an efficient and flexible knowledge discovery system for discovering interesting associations from databases [4].

As we have stated above our approach include two phases: sample and mine. The following subsections explain them in details.

A. Sampling Phase

1) Introduction: Researchers have been proposing sampling algorithms according to their needs. Traditionally, Zaki et al. [7] state that simple random sampling can reduce the I/O cost and computation time for association rule mining. Instead of a static sample, john and Langley [8] use a dynamic sample, where the size is selected by how much the sample represents the data, based on the application. More recently, [9] Wang proposed sampling algorithm and applied it to the selection of a concise training set for multimedia classification. In [10] Akcan proposed two algorithms in his research, named Biased-L2 and DRS, to find a sample S which optimizes the root mean square (RMS) error of the frequency vector of items over the sample. Also, in [11] CHUANG proposed Feature-Preserved Sampling technique for Streaming Data. Now, "Can one sampling technique be valid for all situations?" Unfortunately, the answer is negative. Consequently, we should have a way to permit using more that one sampling technique in our framework [12]. The question is then who, and how the appropriate sampling technique will be selected. To solve this problem we have identified a set of selection criteria for each sampling technique. And the framework should also permit the representation for these criteria to allow a flexible selection. The following sub sections describes some used sampling techniques together with there selection criteria.

2) Sampling Techniques: Simple Random Sampling: Each data record has the same chance of being included in the sample.

- *First N Sampling*: The first *n* records are included in the sample.
- *Weighted Sampling*: In which the inclusion probabilities for each element of the population is not uniform, each element in the population has a different probability of being selected in the sample according to a defined criteria [13].
- *Stratified Sampling* : In stratified sampling, one or more categorical variables are specified from the input data table to form strata (or subsets) of the total population by dividing the area up into a number of strata such that within each of the strata the values of the variable of interest are expected to be relatively similar.
- *Proportional Sampling*: In probability sampling [14], every observation in the population from which the sample is drawn has a known probability of being selected into the sample.
- *Cluster Sampling*: This method builds the sample from different clusters [16]. Each cluster consists of records that are similar

in some way. Clustered samples are generated by first sampling cluster unit and then sampling several elements within the cluster unit.

- *Multi-Stage Sampling*: Multistage sampling is sampling where the elements are chosen in more than one stage [17,18]. Initially large areas selected then progressively smaller areas within larger area are sampled, this process may continue till a sample of sufficiently small ultimate area units (UAUs) is obtained [19].

3) Criteria for Selecting the Appropriate Technique: This section identifies the criteria for selecting the appropriate sampling technique; this criterion must be able to appropriately discriminate between different techniques to be applied on the data for providing the best sample which can be applied on the mining algorithm and propose the optimal mining result. The following subsections present the main four categories of the features.

- *Task Domain Features:-*

It defines the application domain of the task including the type of processes which will be applied to the input in order to improve the result. It also includes the related features which clears the degree of complexity of the processes.

- *Process Type*: This feature defines the type of the process is the task about; it may be one of the mining tasks like clustering, association, and prediction.
- *Domain Knowledge Generality*: This feature describes the degree that the domain may be described in terms of general domain knowledge. The domain may have generalized knowledge, or specialized knowledge.

- *Task Purpose Features*:-

Task purpose specifies the goal of the task with respect to the role of input and output and describes the applied method in terms of the relation between the input and the output.

- *Data size*
- *User directed format*
- *Relevancy degree to the domain*
- *Completeness*
- *Data correctness*
- *Noise*

- *Task Environment Features:-*

The task environment is the organization in which the systems have to operate and it restricts the set of task behaviors that are acceptable. The following are the task environment features.

- *User interaction*
- *Costs*
- *Task Grounding Features*

In general, grounding concerns the relation between the actual system which the task is about and the model of the system, when identifying grounding relations. The initial focus is on the way of interacting with the task reality. The following is the grounding features related to our work.

- *Complexity of the computation*
- *Task Type*
- *Processing speed*
- *Ease in implementation*
- *Usability*
- *Ability to resample*

B. Mining Phase:

The second phase concerns the applied algorithm to mine the data, in this phase we aim in integrating more than one method of finding the frequent item sets according to the needs of the user. This can be accomplished by applying an integrated two mining algorithms for using different constraints on the selected sample. This step is applied by integrating two algorithms, CAP[20] and FIC[21]. CAP algorithm is used for mining association rules with the concept of 'constraint association queries' which is formulated with the only part of the database which is interesting to the user. FIC algorithm is used to mine association rules by pushing support constraints in a better performance than CAP, but with more limits in the results. So integrating both algorithms will yield to a better performance with maintaining the structure of the CAP algorithm to obtain optimal results in finding rules.

VI. CONCLUSION

The objective of this research is to build an efficient knowledge discovery system. Our approach concerned improving the utility of the data mining process.

To achieve this goal, we have proposed a framework that divides the discovery task into two main tasks: 'sampling' which considers the dataset to be mined and 'mining' which considers the applied algorithm to mine the data and integrating more than one method to accomplish the mining task efficiently. More than one sampling techniques have been proposed in the work and suggested a set of criteria for selecting the suitable sampling technique to be applied on the database.

For more preciseness, future work can focus on extending the criteria for selecting the appropriate sampling technique and thus estimating the dataset for mining leads to raise the effectiveness and efficiency of the mining task. Thus the result is user focused.

REFERENCES

[1] N. Megiddo and R. Srikant. "Discovering Predictive Association Rules". Proc. of the 4th Int'l Conference on Knowledge Discovery in Databases and Data Mining, New York, Auguest 1998.

[2] Some Trends in Knowledge Discovery and Data Mining M.Lobur, Yu. Stekh, A.Kernytskyy, Faisal M.E. Sardieh

[3] Sam Y. Sung, Zhao Li, Chew L. Tan, Peter A. Ng, "Forecasting association rules using Existing Data sets", IEEE Transactions on knowledge and data engineering, vol 15, No 6, November 2003

[4] Sampling Technique Selection Framework for Knowledge Discovery Hesham Ahmed Hassan #1, Amira Mohamed Idrees *2 # Faculty of Computers and Information, Cairo University 1 hesham@claes.sci.eg* Central Lab. for Agriculture Expert Systems 2 amira@claes.sci.eg.

[5] George H. John, Pat Langley, "Static Versus Dynamic Sampling for Data Mining", in the proceeding of the second international conference on knowledge discovery in databases and data mining" AAAI/MIT Press, 1996

[6] Pedro Domingos, Geoff Hulten , "Mining high-speed data streams", Conference on Knowledge Discovery in Data Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages: 71 – 80, ACM Press New York, NY, USA, 2000

[7] M. J. Zaki, S. Parthasarathy, W. Lin and M. Ogihara. "Evaluation of sampling for data mining of association rules." Technical Report 617, University of Rochester, Rochester, NY, 1996.

[8] G.H. John and P. Langley. "Static versus dynamic sampling for data mining." Proc. 2nd ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD), pp. 367370, 1996.

[9] Wang, S., Dash, M., Chia, L.-T., and Xu, M. 2007. "Efficient sampling of training set in large and noisy multimedia data" ACM Trans. Multimedia Comput. Commun. Appl. 3, 3, Article 14 (Aug. 2007).

[10] H¨useyin Akcan Alex Astashyn Herv´e Br¨onnimann Leonid Bukhman "Deterministic Sampling beyond EASE: Reducing Multi-Dimensional Data", KDD'05.

[11] Kun-ta Chuang, Hung-leng Chen, and Ming-syan Chen, "Feature-Preserved Sampling over Streaming Data", ACM Transactions on Knowledge Discovery from Data, Vol. 2, No. 4, Article 15, Publication date: January 2009.

[12] Srinivasan Parthasarathy: Efficient Progressive Sampling for Association Rules. ICDM 2002: 354-361

[13] Saar-Tsechansky, M. and F. Provost. "Active Sampling for Class Probability Estimation and Ranking." Machine Learning 54:2 2004, 153-178

[14] H. Leung, T. Y. Chen, " a new prespective of the proportional sampling strategy", the computer journal, Vol 42, No 8, 1999

[15] Tobias Scheffer, Stefan Wrobel, "Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling" Journal of Machine Learning Research 3 (2002) 833-862.

[16] Palmer, C.R., Faloutsos, C.,, "Density Biased Sampling: An Improved Method for Data Mining and Clustering" SIGMOD International Conference on Management of Data (SIGMOD 2000), Dallas, TX, May 1419, 2000

[17] Ashwin Srinvasan, "A study of two sampling methods for analyzing large datasets with ILP", pp. 95-123 Journal Special Issues on ILP, published in 1999 KLuwer Academic Publishers, Boston

[18] Bin Chen, Peter Haas and Peter Scheuermann, "A New Two-Phase Sampling Based Algorithm for Discovering

[19] Ke Wang, Yu He, Jiawei Han, "Mining Frequent Itemsets Using Support Constraints", Proceeding of the 26th VLDB conference, Cairo, Egypt, 2000

[20] Pei, J., Han, J., 2000. "Can we push more constraints into frequent pattern mining?", In Proceedings of ACM International Conference on Knowledge Discovery and Data Mining, 350–354.

[21] Pei, J., Han, J., Lakshmanan, L.V.S., 2001. "Mining frequent itemsets with convertible constraints." In Proceedings of IEEE International Conference on Data Engineering, pp. 433–442.