International Journal of Electrical, Electronics and Computer Engineering **3**(2): 149-153(2014)

ISSN No. (Online): 2277-2626

A Review of Data Reduction/ Extraction in Data mining from the Large set of Database

Jaya Shrivastava* and Prof Neelesh Shrivastava*

**Department of Computer Sciences,*
Vindhya Institute of Technology & Science Satna, (MP), INDIA

(Corresponding author: Jaya Shrivastava)
(Received 05 October, 2014 Accepted 18 November, 2014)

ABSTRACT: Data reduction or extraction from large set of the database is the key challenge nowadays which is gathered from the temporal or spatial domain source such as remote sensing, medical and satellite data. Data mining is the widely used technology to uncover the hidden information from the huge datasets of database. Various approaches such as clustering and association rule are used to extract the information. In this paper review of various approaches proposed by different researchers and also explained the data reduction techniques to extract the essential information with its advantages and disadvantages.

Keywords: Association rule, Clustering, Data reduction, Mining, Spatial or Temporal dataset.

I. INTRODUCTION

Evolution in current technologies for data input through such medium as data from satellite, remote sensing, medical data and organizational data can be stored easily and ease to access with less cost. But the problem is with large-scale mass data and traditional data analysis work with only some surface treatment but cannot get the intrinsic relationship among the data and the principal information from fall into the "data rich, knowledge poor" problem [1]. To get away this problem, people immediately need a species can astutely and robotically transform the data into useful information and acquaintance of techniques and tools, which are on the strong force the vital needs of data analysis tools compose data mining (Data Mining) technology emerged [2]. Data mining in recent years with the database and artificial intelligence developed a new technology that the big amount of raw data to discover the hidden, useful information and knowledge to help policy makers to find the potential between the data Associated factors found to be ignored. While data mining methods are faster when used on smaller data sets, the demand for accurate models often requires the use of large data sets that allow algorithms to discover complex structure and make accurate parameter estimates. Therefore, one of the most important data mining problems is to determine a reasonable upper bound of the data set size needed for building sufficiently accurate model.

Oates and Jensen [3] found that increasing the amount of data used to build a model often results in a linear increase in model size, even when additional complexity causes no significant increase in model accuracy. Despite the promise of the better parameter estimation, models built with large amounts of data are often needlessly complex and cumbersome. Data reduction can also be extremely helpful for data mining from very large distributed databases.

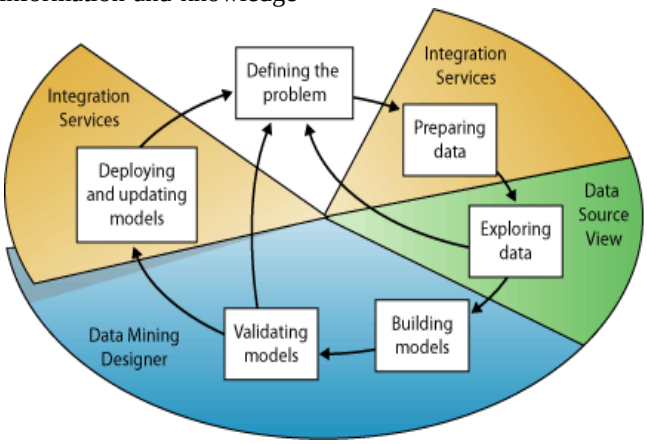

Fig. 1. Data mining process for task oriented.

In the contemporary data mining community, the majority of the work for learning in a distributed environment considers only two possibilities: moving all data into a centralized location for further processing, or leaving all data in place and producing local predictive models, which are later moved and combined via one of the standard machine learning methods [4]. With the emergence of new high cost networks and huge amounts of collected data, the former approach may be too expensive, while the latter too inaccurate. Therefore, reducing the size of databases by several orders of magnitude and without loss of extractable information could speed up the data transfer for a more efficient and a more accurate centralized learning. In this paper we present the literature of the approaches proposed/implemented by different researchers together with its advantages and disadvantages.

The rest section of the paper is arranged in such manner: Section second Data reduction techniques in data mining. Section third presents the related work done in the extraction of data and last section discusses about the whole paper.

II. DATA REDUCTION TECHNIQUES IN DATA MINING

Data mining is a large number of incomplete, noisy, fuzzy, random the practical application of the data found in hidden, regularity, people not known in advance, but is potentially useful and ultimately understandable information and knowledge of non-trivial process [5]. "Known in advance" means the information is pre- unanticipated, or novelty. The information unearthed more surprising, the more likely value. "potential the usefulness of the "means of knowledge found in actual effect of future, that information or knowledge of the business discussed or research to be effective, there are practical and achievable [6]. The definition of data mining is closely related to another commonly used term knowledge discovery [7]. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. In data mining there are various techniques has been developed to reduce/extract the data from the large set of database such as clustering, association, classification, genetic algorithm, and neural network etc. in which some of them are describe below:

A. Clustering

Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments [8].

B. Classification

Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified. Classification [8] routines in data mining also use a variety of algorithms - and the particular algorithm used can affect the way records are classified. A common approach for classifiers is to use decision trees to partition and segment records. New records can be classified by traversing the tree from the root through branches and nodes, to a leaf representing a class. The path a record takes through a decision tree can then be represented as a rule

C. Association Rule

The task of mining association rules over market basket data is considered a core knowledge discovery activity. Association rule mining provides a useful mechanism for discovering correlations among items belonging to customer transactions in a market basket database. Essentially, association mining is about discovering a set of rules that is shared among a large percentage of the data. Association rules mining tend to produce a large number of rules. The goal is to find the rules that are useful to users. There are two ways of measuring usefulness, being objectively and subjectively. Objective measures involve statistical analysis of the data, such as support and confidence.

Support. The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have as greater than a user-specified support is said to have minimum support.

Confidence. The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

Almost all association algorithms are objective and use some form of statistical analysis to determine the usefulness of a rule. Thus the set of all transactions used in the analysis must be sufficiently large in order for association rules to be concluded from it. Therefore, for the rest of this article, the term large will be used to describe a set of data with enough transactions to obtain association rules.

There are a few commonly used terms that must be defined:

(i) **Item-set:** An items-et is a set of items. A k-item-set is an item-set that contains k number of items.

(ii) **Frequent item-set:** This is an item-set that has minimum support.

(iii) **Candidate set:** This is the name given to a set of item-sets that require testing to see if they fit a certain requirement [9] and [10].

D. Genetic Algorithm

The genetic algorithms (GA) efforts with a population of the potential solutions [11]. In computing terms, genetic algorithms map strings of numbers to each potential solution. Every solution becomes an individual in the population, and every string becomes a representation of an individual [11]. Nearby ought to be a mode to derive each individual from its string representation. The genetic algorithm then again manipulates the most promising strings in its search for an enhanced solution. This algorithm uses the following cycle.

(i) Creation of a population of strings.
(ii) Evaluation of each string.
(iii) Selection of the best strings.
(iv) Genetic manipulation to create a new population of strings.

Objective Function. As we knew that the genetic algorithm is good at exploring or extracting for undetermined solutions, which is unusual to see that genetic algorithm is used to mine association rules. The genetic algorithm (GA) is used to mine association rule, along with the all measurements, one measurement is accuracy or confidence factor. In this another three measures of the rules is explained such as comprehensibility [12], interestingness [13] and completeness, in addition to predictive accuracy. By using these measures, some formerly unknown, simply understandable and compressed rules can be generated. It is very difficult to quantify understandability or comprehensibility. A vigilant study of an association rule will deduce that if the number of conditions involved in the antecedent part is fewer hence the rule is more comprehensible. For replicating such behavior, an expression was derived as comp = N – (Number of circumstances in the antecedent part). These expressions fit for the classification rule generation where the number of attributes in the consequent part is always one. We require an expression where the number of attributes involved in both the parts of the rule has some effect. The following expression can be used to quantify the comprehensibility of an association rule,

$$\text{Comprehensibility} = \log\left(1 + \frac{|C|}{(|D| - |A|)}\right) * \frac{1}{|A|}$$

Here, $|A|$ and $|C|$ are the attributes implicated in the consequent part and the antecedent part, respectively and |D| is the total number of records in the database.

This is most essential that whatever rule will be selected for valuable one this rule should signify all useful attributes or components. So for which we have to select compact association rule with all useful characteristics. Therefore we have to find out the frequent itemsets with maximum length. An antecedent part and consequent part for an association rule must cover all useful features as well as the two parts should be frequent. The following expression can be used to quantify the completeness of an association rule:

Completeness = (log (1 + |C| + |A|) / |D|)* Supp (A) * Supp(C)

Here, |C| and |A| are the number of attributes involved in the consequent part and the antecedent part, respectively and |D| is the total number of records in the database. Supp (A) and Supp(C) are the occurrences of Antecedent part and consequent part respectively.

The most important is that finding interestingness of the data set is to be separated based on each attribute nearby in the consequent part.

Since a number of attributes can shows in the consequent part and they are not pre-defined, such method may not be viable for association rule mining. The subsequent expression can be used to describe as interestingness of an association rule;

$$\text{Interestingness} = \text{Supp}(A) * \left[\left(1 - \text{Supp}(C) \Big/ 1 - \text{Supp}(AUC)\right)\right] * \left[\text{Supp}(AUC) \Big/ \text{Supp}(A)\right] * \left[\text{Supp}(AUC) \Big/ \text{Supp}(C)\right]$$

$$\text{Supp}(A) * \left[\left(1 - \text{Supp}(C) \Big/ 1 - \text{Supp}(AUC)\right)\right]$$

The expressions consist of two parts. The first one is that and then compares the probability that A appears without C if they were dependent with the actual frequency of the appearance of A. The second part measures the difference of A and C appearing mutually in the data set and what would be expected if A and C where statistically dependent.

E. Artificial Neural Network

The Artificial Neural Network (ANN) is a very commonly used technique to solve data mining problems. Neural Network is a set of processing units which are assembled in a tightly interconnected network, based on some features of the biological neural network. As biological neural network or human brain learns by its surrounding, ANN learns by its past experience. The structure of neural network provides an opportunity to the user to implement parallel concept at each layer level. A very significant characteristic of ANN is that they are fault tolerant in nature. ANNs are very good in situations where information is uncertain and noisy. ANN are an information processing methodology that differs drastically from conventional methodologies in that it employ training by examples to solve problem rather than a fixed algorithm [14,15]. Training of a NN can be categorized in to two methods: Unsupervised training and Supervised training. Supervised networks that are require the actual desired output for each input and require a teacher for training, where as unsupervised networks does not require the desired output for each input and does not require a teacher for training.

Learning process of neural network is an iterative learning process in which every data cases are presented to the network one by one, and the weights adjustment is done for all the input values coming to network [16].

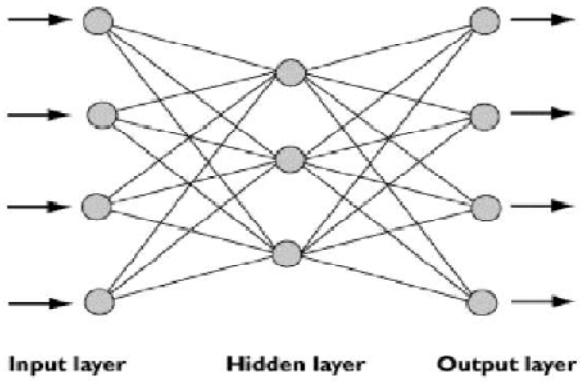

Fig. 2. Artificial Neural Network.

When all the cases are presented, the process starts again from its beginning. In the learning phase, weight adjustment is done to make the network learn so that it may able to predict the correct class label of input samples whose class label is unknown. Once the structure of network is ready for a specific problem, than the network is ready to be trained.

Initial weights are chosen randomly to start the training process. Then the learning or training, begins. The most popular and commonly used neural network training algorithm is back-propagation algorithm. Although many types of neural networks are available for classification purposes [17].

III. LITERATURE SURVEY

Compieta *et al.* [18] proposed a strategy that is to be incorporated in a system of exploratory spatio-temporal data mining, to improve its performance on very large spatiotemporal datasets. This system provides a data mining engine that can integrate different data mining algorithms and two complementary 3D visualisation tools. Within this system, there is a 2-layer architecture; a mining layer that provides newly developed techniques to efficiently support the data mining process, address the spatial and temporal dimensions of the dataset, and a visualisation layer to visualise and interpret results. Wilson *et al.* [19] proposed a k-nearest neighbour approach for the reduction of data.

In general the k-nearest neighbour rule, in particular. The effectiveness of the reduced set is measured in terms of the classification accuracy. These methods attempt to derive a minimal consistent set, i.e., a minimal set which correctly classifies all the original samples. The very first development of this kind is the condensed nearest neighbour rule (CNN) [20]. Other algorithms in this category including the popular IB3, IB4 [21].

Akhil Kumar *et al.* [22] presented techniques that can be employed effectively for exact and approximate reduction in a database system. These techniques can be implemented efficiently in a database system using SQL (structured query language) commands. We tested their performance on a real data set and validated them. The results showed that the classification performance actually improved with a reduced set of attributes as compared to the case when all the attributes were present. They also discussed how our techniques differ from statistical methods and other data reduction methods such as rough sets. Lowe *et al.* [23**]** presented a Variable Similarity Metric (VSM) learning system that produces a confidence level of its classifications. In order to reduce storage and remove noisy instances, an instance t is removed if all k of its neighbours are of the same class, even if they are of a different class than t (in which case t is likely to be noisy). This removes noisy instances as well as internal instances, while retaining border instances.

The instance is only removed, however, if its neighbours are at least 60% sure of their classification. The VSM system typically uses a fairly large k (e.g., k D 10), and the reduction in storage is thus quite conservative, but it can provide an increase in generalization accuracy. Also, the VSM system used distance-weighted voting, which makes a larger value of k more appropriate.

IV. CONCLUSION

From the large set of database it is essential to retrieve or extract the needful data. By using the data mining techniques it not only reduces the size of data storage but also enhance the accessing capacity of the data. In this paper we present the literature study of different approaches and techniques with its disadvantages and disadvantages.

REFERENCE

[1]. Esehrieh S. Jingwei Ke, Hall L. O. etc. Fast accurate fuzzy clustering through data reduction *IEEE transactions on fuzzy systems 2003* ’**11**(2):262-270.

[2]. Zahid N ’Abouelala O ’Limouri M ’etc. Fuzzy clustering based on K-nearest-neighbours rule [J].Fuzzy Sets and Systems ’2001 ’120(2)

[3]. Oates, T., Jansen, D.: Large Datasets Lead to Overly Complex Models: An Explanation and a Solution, *Proc. Fourth International Conference On Knowledge Discovery and Data Mining*, (1998), 294-298.

[4]. Grossman R, Turinsky A. A Framework for Finding Distributed Data Mining Strategies That Are Intermediate Between Centralized Strategies and In-Place Strategies, KDD Workshop on Distributed Data Mining, (2000)

[5]. Jung-Hua Wang ’Jen-Da Rau ’Wen-Jeng Liu. Two-stage clustering via neural networks [J]. IEEE ransactions on Neural Networks ’2003 ’**14**(3):606-615.

[6]. Veenman C.J. ’Reinders M.J.T. ’Backer E. “A maximum variance cluster algorithm” ’[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence* ’2002 ’24(9):1273-1280.

[7]. Cattell Raymond B. The Three Basic Factor Analytic Research Designs-Their Interrelations and Derivatives [J]. Psychological Bulletin ’1952 ’49:499-520

[8]. Ed Colet “Clustering and Classification: Data Mining Approaches” http://www.virtualgold.com

[9]. Rupali Haldulakar and Prof. Jitendra Agrawal “Optimization of Association Rule Mining through Genetic Algorithm”, *International Journal on Computer Science and Engineering (IJCSE)*, Vol. **3** No. 3 Mar 2011, pp. 1252-1259.

[10]. Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba, “Extraction of Interesting Association Rules Using Genetic Algorithms”, Advances in Systems Modelling and ICT Applications, pp. 101-110.

[11]. Anshuman Singh Sadh, Nitin Shukla: “Association Rules Optimization: A Survey”, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Vol. **3** Number-1 Issue-9 March-2013.

[12]. Sufal Das, Bhabesh Nath, “Dimensionality Reduction using Association Rule Mining”, IEEE Region 10 Colloquium and Third International Conference on Industrial and Information Systems (ICIIS 2008) December 8-10, 2008, IIT Kharagpur, India

[13]. Hsu, W., B. Liu and S. Chen, “Ggeneral impressions to analyze discovered classificationrules”,. *Proc. Of 3rd Intl. Conf. On Knowledge Discovery & Data Mining (KDD-97),* pp: 31–36.AAAI Press.(1997)

[14]. Anil Jain k., Jianchang Mao and K.M. Mohiuddi, “Artificial Neural Networks: A Tutorial”, *IEEE Computers*, (1996) pp.31-44.

[15]. George Cybenk, “Neural Networks in Computational Science and Engineering”, *IEEE Computational Science and Engineering*, (1996), pp.36-42.

[16]. Rojas, “Neural Networks: a systematic introduction”, *Springer-Verlag* (1996).

[17]. R. P. Lippmann, “Pattern classification using neural networks,” *IEEE Commun. Mag*, (1989), pp. 47–64.

[18]. Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., and Kechadi, T., Exploratory Spatio-Temporal Data Mining and Visualization. *Journal of Visual Languages and Computing*, **18**, 3, pp.255-279, June, 2007.

[19]. Wilson, D.R. and Martinez, T.R., Reduction Techniques for Instance-based Learning Algorithm. *Machine Learning*, **33**, 3, 257–286, 2000.

[20]. Angiulli, F. “Fast Condensed Nearest Neighbor Rule” *Proceedings of the 22nd International Conference on Machine Learning (Bonn, Germany, August 07 - 11*, 2005), vol. **119**. ACM, New York, NY, pp.25-32. 2005.

[21].Aha, D. W., Tolerating Noisy, Irrelevant and Novel Attributes in Instance-based Learning Algorithms. International Journal of Man-Machine Studies, 36-2, pp.267-287, February, 1992.

[22]. Akhil Kumar “New Techniques for Data Reduction in a Database System for Knowledge Discovery Applications”, *Journal of Intelligent Information Systems* January 1998, Volume **10**, Issue 1, pp 31-48

[23]. Lowe, D. G. “Similarity Metric Learning for a Variable-Kernel Classifier” Neural Computation, **7**(1), 72–85 (1995).