International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 6, June 2016)

Predictive Analytics of Cluster Using Associative Techniques Tool

Jyoti[1], Savita Bisnoi[2]
[1]M. Tech Scholar, [2]Associate Professor, CSE Dept, RIEM, Rohtak, MDU University Rohtak, India

Abstract: **- Mining the data means fetching out a piece of data from a huge data block. The basic work in the data mining can be categorized in two subsequent ways. One is called classification and the other is called clustering. Although both refers to some kind of same region but still there are differences in both the terms. The goal of the thesis is to experimentally exploring clustering and classification technique for predictive analysis. Apriori Algorithms is implemented using C#. Apriori partitioning and sampling algorithms have been implemented and their performance is evaluated extensively. We are also analysis diabetic data on WEKA tool using clustering and classification technique.**

Keyword: **Data Mining, Clustering, Classification, Apriori Algorithm, WEKA tool**

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing [1], model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

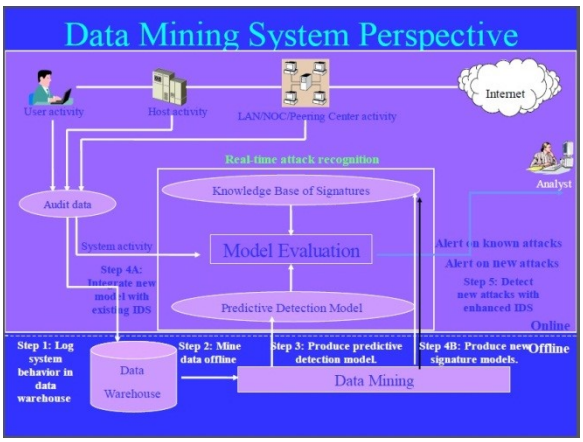

Figure 1: Basic of Data Mining

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons [2]. Often the more general terms "(large scale) data analysis", or "analytics" or when referring to actual methods, artificial intelligence and machine learning are more appropriate.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 6, June 2016)

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, and data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

II. Applying Data Mining To CRM

The Two Crows data mining process model described below is similar to other process models such as the CRISP-DM model, differing mostly in the emphasis it places on the different steps. The initial models you build may provide insights that lead you to create new variables [3]. The basic steps of data mining for effective CRM are:

1. Define business problem
2. Build marketing database
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model and results

One key issue you must deal with in applying a model to new data is the transformations you used in building the model. Thus if the input data (whether from a transaction or a database) contains age, income and gender fields, but the model requires the age-to-income ratio and gender has been changed into two binary variables, you must transform your input data accordingly.

III. Association Rule In Data Mining

Association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements [4]. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

IV. Classification In Data Mining

This technique of data mining is done on data base sets with a record of information.

- *Step1:* with the help of training set of data producing association rule set.
- *Step2:* eliminate all those rules that may cause over fitting.
- *Step3:* finally we predict the data and check for accuracy and this is said to be the classification phase.

The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time [5]. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. See "Testing a Classification Model".

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 5, Issue 6, June 2016)

V. CLUSTERING IN DATA MINING

Cluster analysis is a traditional statistical method that was also initially used for simple data mining. This method is used for classifying a collection of things into segments whose members have similar characteristics. Unlike Classification, Clustering belongs to the unsupervised learning techniques, which entail the modeling of data with the involvement of inputs without having a predefined output [6]. All data inputs are treated similarly in order to obtain information for the determination of groups or associations. Therefore, in clustering the characteristics according to which the objects are categorized into segments or else classes are initially unknown. The clusters are determined when relevant clustering algorithms are applied on the data set under investigation and the similarities in the characteristics of the objects are identified. Some popular algorithms for cluster analysis are: [7, 8] K-means which is a quite fast algorithm applicable in large and wide datasets that requires predetermination of the number of clusters by the user; Two Step which processes the records in two sets and can automatically determine clusters; Kohonen network/Self Organizing Map (SOM), which is a unique neural network architecture that produces a two-dimensional map of the clusters, and is slower than Two Step and K-means. Clustering techniques are often used for customer segmentation. This technique can be applied to huge datasets. There are already software tools in the market able to perform several clustering algorithms.

VI. APRIORI ALGORITHM

Apriori is a seminal algorithm for finding frequent item-sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item-sets, "if an item-set is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or item-set are sorted in lexicographic order.

Let the set of frequent item-sets of size k be Fk and their candidates be Ck. Apriori first scans the database and searches for frequent item-sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent item-sets.

1. Generate Ck+1, candidates of frequent item-sets of size k +1, from the frequent item-sets of size k.
2. Scan the database and calculate the support of each candidate of frequent item-sets.
3. Add those item-sets that satisfy the minimum support requirement to Fk+1.

VII. WEKA

Weka is a program that evaluates the data analysis to develop models to support on decision making on business warehouse management. There are many data mining techniques for model developing. Among the most popular ones are Classification, Clustering and Association Rule Discovery which are applied in model developing (Weka Machine Learning Project, 2010; Wass, 2007). For the models supporting decision making in business warehouse management in this essay, Classification is used in analyzing nominal data and prediction analysis for numeric data.

Processes in model developing with classification technique are applying training data in developing mode land testing the model with the evaluation data. Then the model is used in real practice by applying the unseen data that there is still no answer class with this model, developed by Weka program.

VIII. CONCLUSION

Successful completion of the clustering for the given files Apriori algorithms is providing results. Along with this we have the completion of clustering for the algorithm of modified Apriori clustering techniques. Comparison is performed by these algorithms on database. Only some samples of data are taken and classify according their properties. In future the proposed algorithm can be applied on any type files. The variety and quantity of data is constant in this work so in future we can vary these issues also.

REFERENCES

[1] W. Pieczynski, Chaînes de Markov Triplet, Triplet Markov Chains, Comptes Rendus de l' Académie des Sciences – Mathématique, Série I, Vol. 335, No. 3, pp. 275-278, 2002.

[2] W. Pieczynski, Multisensor triplet Markov chains and theory of evidence, International Journal of Approximate Reasoning, Vol. 45, No. 1, pp. 1-16, 2007.

[3] Bertram, Raymond, Alexander Pollatsek, and Jukka Hyönä. "Morphological Parsing and the Use of Segmentation Cues in Reading Finnish Compounds." Journal of Memory and Language 51.3 (2004): 325-345. 27 Apr. 2014.

[4] Nisbet, Robert A. (2006); Data Mining Tools: Which One is Best for CRM? Part 1, Information Management Special Reports, January 2006.

[5] Azevedo, A. and Santos, M. F. KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182–185

[6] Goebel, Michael; Gruenwald, Le (1999); A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, Vol. 1, Issue 1, pp. 20–33.

[7] Kobielus, James; The Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010, Forrester Research, 1 July 2008.

[8] Biotech Business Week Editors (June 30, 2008); Biomedicine; HIPAA Privacy Rule Impedes Biomedical Research, Biotech Business Week, retrieved 17 November 2009 from LexisNexis Academic.

[9] Potu Narayana , "Association Rule Mining Based On Apriori Algorithm", International Journal For Development Of Computer Science & Technology ISSN-2320-7884 (Online)Volume-1, Issue-Iii (April-May 2013).