International Journal of Advanced Science and Research
ISSN: 2455-4227
Impact Factor: RJIF 5.12
www.allsciencejournal.com
Volume 3; Issue 1; January 2018; Page No. 20-23

Data Mining of the large dataset for classification based on rule and tree based classifiers: A Review

Ankur Gupta
Assistant Professor, Department of Computer Science, RSD College, Ferozepur City, Punjab, India

Abstract
Data mining is defined as the procedure of extracting information from huge sets of data. Data mining is mining knowledge from data. Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms. In the process of data mining various types of classifiers have been used for decision evaluation process. In this paper various approaches have been discussed that can be used for classification of different datasets. On the basis of rules, and trees various classifiers have been reviewed and there process of classification of data has been discussed in this paper.

Keywords: data mining, decision table, decision tree, SVM, naïve byes

1. Introduction

1.1 Data Mining

It is the process of fetching hidden knowledge from a wide store of raw data. The knowledge must be new, and one must be able to use it. Data mining has been defined as "It is the science of fetching important information from wide databases". It is one of the tasks in the process of knowledge discovery from the database. Data Mining is used to discover knowledge out of data and present the data in an easy and understood able form. It is a process to examine large amounts of data routinely collected. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. Two goals of data mining are prediction and description. Prediction tells us about the unknown value of future variables.

1.2 Architecture for data mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data.

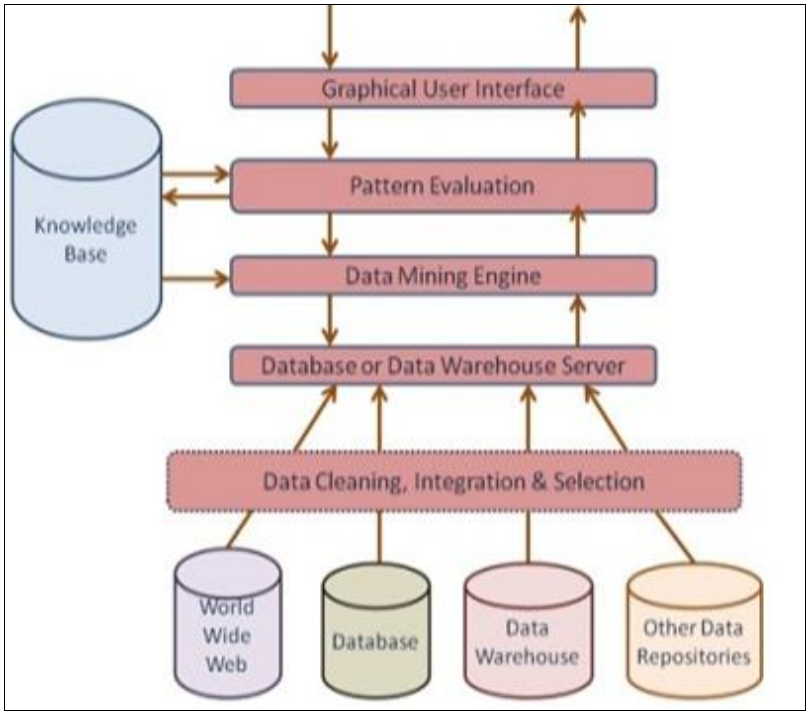

Fig 1: Integrated Data Mining Architecture

Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data ware house can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business - summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

1.3 Key phases in data mining process

1.3.1 Information association

The one of the most familiar and straightforward feature of this system is that here we made association between two or more items or often of the same type to formulate particular pattern. Like it is very well known etiological association between smoking and lung cancer. We have to collect data concerned with smoking habit details including numbers of smoke per day, duration of smoking, type of smoking either bidis, cigarettes, specific brands, lifestyle and age of patient Etc.

1.3.2 Information classification

This is the second phase in this we can classify the collected information according to our objectives like etiological factors, investigation purpose, drug treatment plans and results. For example the etiological information collected from lung cancer patients can be classified on the basis of duration of smoking habit, type of exposure, number of exposure, age of patient etc.

1.3.3 Pattern Sequencing

This is the next step in module preparation. The pattern sequencing can be prepared with the help of readymade software packages available in market.

1.3.4 Preparation of decision tree

This is final step of prediction system.

1.3.5 Implementation

This is directly concerned with last step. You may have option either long term or short term data processing. Each data mining system has their different objectives. Data mining process are broadly formulated either as supervised run supervised learning. Supervised learning is that type of learning in which a training set is used to learn model parameters but in Unsupervised learning no training set is used. These are broadly dived either classification or prediction based pattern. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

1.4 Data Mining Techniques

Data mining technique is linked with data processing, identifying patterns and trends in information. Or we can say that data mining simply means collection and processing data in systemic manner by using computer based programs and subsequent formation of disease prediction or patient management system aid. With the invention of information technology, now these days it is even more prevalent. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages. This technique is routinely use in large number of industries like engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining.

2. Review of Literature

Thuraisingham, B *et al.* [1] "Data Mining for Malicious Code Detection and Security Applications "In this paper author want to say that data mining is the process of posing queries and fetching patterns from large quantities of data using pattern matching or some other reasoning techniques. Data mining has many applications in security including for national security as well as for cyber security. Threats include in national security attacking buildings, destroying critical infrastructures such as power grids and telecommunication systems. Datamining techniques are being investigated to find out who the suspicious people are and who is capable of carrying out terrorist activities. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. Datamining is also being applied to provide solutions such as intrusion detection and auditing.

Thuraisingham, B *et al.* [2] "Data mining for security applications" Author want to proposed that the presentation will provide an overview of datamining and security threats and then discuss the applications of data mining for cyber security and national security including in intrusion detection and biometrics. Privacy considerations including a discussion of privacy preserving data mining will also be given.

Asghar, S *et al.* [3] "Automated Data Mining Techniques: A Critical Literature Review" In this paper author want to proposed that data mining has emerged as one of the major research domain in the recent decades in order to extract implicit and useful knowledge. This knowledge can be

comprehended by humans easily. This knowledge extraction was computed and evaluated manually using statistical techniques. Subsequently, semi-automated data mining techniques emerged because of the advancement in the technology. Such advancement was also in the form of storage which increases the demands of analysis. In such case, semi-automated techniques have become in efficient. So automated data mining techniques were introduced to synthesis knowledge efficiently. Consequently

RanaAlaa El-DeenAhmeda *et al*. [4] "Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining", Author proposed eleven data mining classification techniques that are comparatively tested to find the best classifier fit for consumer online shopping attitudes and behavior according to obtained dataset for big agency of online shopping. The results show that decision table classifier and filtered classifier give the highest accuracy and the lowest accuracy is achieved by classification via clustering and simple cart. Also, this paper provides a recommender system based on decision table classifier helping the customer to find the products he/she is searching for in some e-commerce websites. Recommender system learns from the information about customers and products and provides appropriate personalized recommendations to customers to find the desired products.

PareshTanna *et al*. [5] "A performance comparison between classification techniques with CRM application" Author stated knowledge exploration from the large set of data generated as a result of the various data processing activities due to data mining only. Frequent Pattern Mining is considered a very important undertaking in data mining. Apriori approach applied to generate frequent item set generally espouse candidate generation and pruning techniques for the satisfaction of the desired objective. This paper shows how the different approaches achieve the objective of frequent mining along with the complexities required to perform the job. This paper demonstrates the use of WEKA tool for association rule mining using Apriori algorithm.

Ila Padhi *et al*. [6] "Predicting Missing Items in Shopping Cart using Associative Classification Mining" Author presented a technique called the "Combo Matrix" whose principal diagonal elements represent the association among items and looking to the principal diagonal elements, the customer can select what else the other items can be purchased with the currently contents of the shopping cart and also reduce the rule mining cost. The association among items is shown through Graph. The frequent item sets are generated from the Combo Matrix. Then association rules are to be generated from the already generated frequent item sets. The association rules generated form the basis for prediction. The incoming item sets i.e. the contents of the shopping cart will be represented by set of unique indexed numbers and the association among items is generated through the Combo Matrix. Finally the predicted items are suggested to the Customer.

3. Approaches Used

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Decision tables are a precise yet compact way to model complex rule sets and their corresponding actions. Decision tables, like flowcharts and if-then-else and switch-case statements, associate conditions with actions to perform, but in many cases do so in a more elegant way.

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

4. Conclusion

Data mining is the field of data processing or data warehousing that has been used for extraction of valuable information from the raw data based on various set of rules. In the process of data mining clustering, classification and attribute selection has been done. Attribute selection is used for selection of best set of attributes that have minimum dependency on other attributes that are available in the dataset. In this paper various classification approaches have been discussed. Classification has been done for prediction of various data attributes so that best set of the rules can be extracted that can be used for extraction of best decision making process. On the basis of classification rule based classifiers, tree based classifiers and probability based classifiers have been reviewed. On the basis of these classifier one can say that rules based classifiers provide better efficiency for small scale datasets whereas tree based classifiers can be used for classifica5tion of the dataset that contain large instances.

5. References

1. Thuraisingham. Data Mining for Malicious Code Detection and Security Applications**,** 978-0-7695-4406-9, 4-5, IEEE, 2011.

2. Asghar S. Automated Data Mining Techniques: A Critical Literature Review 978-0-7695-3595-1, 75-79, IEEE, 2009.
3. Rana Alaa El-Deen Ahmeda. Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining, Fifth International Conference on Communication Systems and Network Technologies, 2015, 1344-1349.
4. Dalia Ahmed Refaat Mohamed. A performance comparison between classification techniques with CRM application, IEEE International Conference on AI Intelligent Systems, 2015, 112-119.
5. Hossin M. A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process, 2015, 1-6.
6. Nedaabdel Hamid. Emerging trends in associative classification data mining International journal of electronics and electrical engineering, 2015, 56-62.
7. ShreyBavisi Ȧ. A Comparative Study of Different Data Mining Algorithms, International Journal of Current Engineering and Technology, 2015, 3248-3252.
8. Kamal R. Adaptive Pointing Theory (APT) Artificial Neural Network, International Journal of Computer and Communication Engineering, 2014, 212-215.
9. Meenakshi. Survey on Classification Methods using WEKA, International Journal of Computer Applications, 2014, 16-19.
10. Mohammed Al-Maolegi. An Improved Apriori Algorithm For Association Rules, International Journal on Natural Language Computing (IJNLC), 2014, 21-29.
11. Paresh Tanna. Using Apriori with WEKA for Frequent Pattern Mining, International Journal of Engineering Trends and Technology (IJETT), 2014, 127-131.