P.DIVYA *et al*, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 10, October- 2017, pg. 71-82 **ISSN: 2321-8363**
Impact Factor: 4.123

A SURVEY OF COMPARISON STUDY OF CLASSIFICATION FOR HEMATOLOGICAL DATA

P.DIVYA, Dr. K.PALANIVEL
M.Phil. Scholar, Associate Professor
AVC College (Autonomous), Mayiladuthurai

ABSTRACT: ***Research focus on to analysis which calculation is most appropriate for user working on hematological information using Weka tool Environment for Knowledge Analysis. Data mining is defined as analyzing very large amount of data for getting some useful information. In Medical environment, the information which is retrieved from the dataset can be used for accuracy prediction.comparison of various Classification algorithms using Waikato Environment for Knowledge Analysis to analysis hematological data. This work can be able to compare the selective classification algorithm based on weka, thus investigate which algorithm provides efficient result.***

Keywords: classification, weka, Hematological data, Data mining, Knowledge analysis.

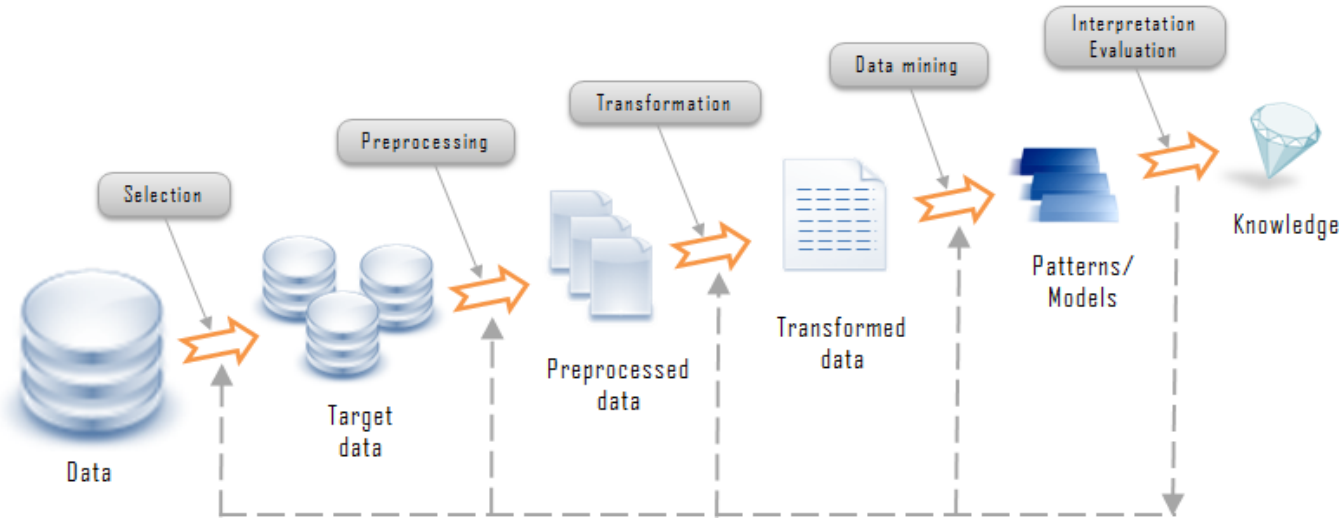
INTRODUCTION

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, valid patterns and relationships in huge data set. These tools are nothing but the machine learning methods, statistical models and mathematical algorithm. Data mining consists of more than collection and managing the data, it also includes analysis and prediction. Classification technique in data mining is capable of processing a wider variety of data than regression and is growing in popularity. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, potentially useful and previously unknown information from data in

databases. 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.Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining, *the extraction of hidden predictive information from large databases,* is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

Knowledge Discovery

Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data. The knowledge discovery process is described as follows:



Knowledge Discovery Process

Let's examine the knowledge discovery process in the diagram above in details:

- Data comes from variety of sources is integrated into a single data store called target data
- Data then is pre-processed and transformed into the standard format.
- The *data mining* algorithms process the data to the output in the form of patterns or rules.
- Then those patterns and rules are interpreted to new or useful knowledge or information.

The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.In some industries, such as retail, these numbers can be much larger. The

accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology.

Integrated Data Mining Architecture

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse 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. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

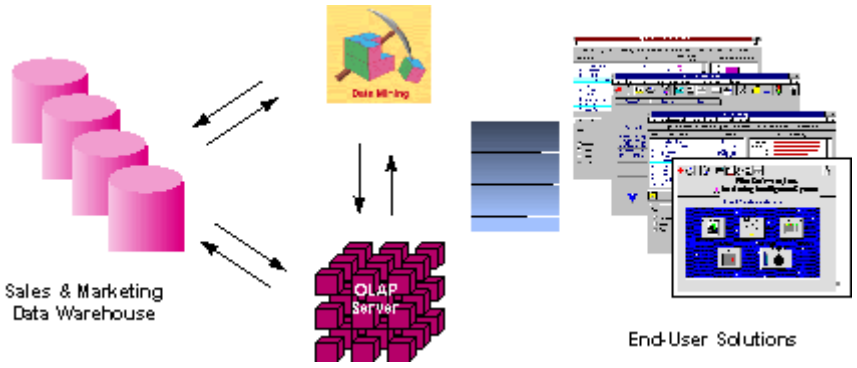


Figure 1 - Integrated Data Mining Architecture

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.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

Data Mining Algorithms

An algorithm in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the results of this analysis over many iterations to find the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. The mining model that an algorithm creates from your data can take various forms, including:

A set of clusters that describe how the cases in a dataset are related.

- A decision tree that predicts an outcome, and describes how different criteria affect that outcome.
- A mathematical model that forecasts sales.
- A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together.

The algorithms provided in SQL Server Data Mining are the most popular, well-researched methods of deriving patterns from data. To take one example, K-means clustering is one of the oldest clustering algorithms and is available widely in many different tools and with many different implementations and options. However, the particular implementation of K-means clustering used in SQL Server Data Mining was developed by Microsoft Research and then optimized for performance with Analysis Services. All of the Microsoft data mining algorithms can be extensively customized and are fully programmable, using the provided APIs.

Choosing an Algorithm by Type

SQL Server Data Mining includes the following algorithm types:

- **Classification algorithms** predict one or more discrete variables, based on the other attributes in the dataset.
- **Regression algorithms** predict one or more continuous numeric variables, such as profit or loss, based on other attributes in the dataset.
- **Segmentation algorithms** divide data into groups, or clusters, of items that have similar properties.
- **Association algorithms** find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- **Sequence analysis algorithms** summarize frequent sequences or episodes in data, such as a series of clicks in a web site, or a series of log events preceding machine maintenance.

Improved Apriori Algorithm

Uses Apriori, the classification algorithm of association rule, for data mining analysis of medical data. According to the characteristics of medical data, it improved the Apriori algorithm. Using the improved Apriori algorithm, it finds frequent item sets in a database of medical diagnosis, and generates strong association rules, in order to find out the useful association relationship or pattern between the large data item sets. By improving the Apriori algorithm of association rules, this study mined decision rules from breast cancer diagnosis database and found out the relationship between breast cancer and the factors such as age, position, etc.

Naïve Bayes Classifier Algorithm

The main objective of this research work is to find the best classification algorithm among Bayesian classifiers. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns which is novel and not known earlier. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text.Methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatics classification of documents.

SVM:

WEKA for Hematological Data Comment is a challenging and interesting task in medical research area. To find out which classification algorithms is better it is very difficult to compare different classification algorithms in different dataset. Our dissertation concerns with to make a mobile App, which is capable to Diagnose Hematological data comments.They divide this problem of Hematology Data comments into three phases: Data Collection, Classification algorithm, and developed App.

STUDY ON RELATED WORK

K. Tamizharasi,Dr. UmaRani,K.Rajasekaran,September 2014

The retail industry collects vast amounts of data on sales, customer buying history, goods, andservice with ease of use of modern computing technology. This thesis elaborates the use of data mining technique to help

retailers to identify customer profile for a retail store and behaviors, improve better customer satisfaction and retention. The aim is to judge the accuracy of different data mining algorithms on various datasets. The performance analysis depends on many factors encompassing test mode, different nature of data sets and size of data set. They analyzed important characteristics of the applications when executed in well-known toolWEKA. The work described in this thesiscomparatively evaluates the performance of algorithms on three testmodes that is hold out method, percentage split, and full training.Our current work is focusing on evaluating the applications on different data sets to allow the retailersto increase customer understanding and make knowledge- driven decisions in order to provide personalized andefficient customer service.

Vaithiyanathan,K. Rajeswari, KapilTajane, and Rahul Pitale, May 2013

In this thesis different classification techniques of Data Mining are compared using diverse datasets from University of California, Irvine(UCI). Accuracy and time required for execution by each technique is observed. The Data Mining refers to extracting or mining knowledge from huge volume of data. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. This work has been carried out to make a performance evaluation of J48, MultilayerPerceptron, Naïve BayesUpdatable, and BayesNet classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The thesis sets out to make comparative evaluation of classifiers J48, Multilayer Perceptron, Naïve BayesUpdatable, and BayesNet in the context of Labour, Soyabean and Weather datasets. The experiments are carried out using weka 3.6 of Waikato University.

Ms.Vijayarani, Ms M. Muthulakshmi,August 2013

Data mining is the non-trivial extraction of implicit, earlier unknown and potentially useful information aboutdata. There are several data mining techniques have been developed and used in data mining thesis which includesclassification, clustering, association rules, prediction, and sequential patterns. Data mining applications are used in variousareas such as sales, marketing, banking, finance, health care, insurance and medicine. There are various research domains indata mining namely web mining, text mining, image mining, sequence mining, privacy preserving data mining, etc. Textmining is a technique which extracts information from both structured and unstructured data and also finding patternswhich is novel and not known earlier. It is also known as knowledge discovery from text (KDT), deals with the machinesupported analysis of text. Text mining is used in various areas such as information retrieval, document similarity, and naturallanguage processing and so on. Searching for similar documents is an important problem in text mining. Thesis demonstrate that the efficiency of j48 and Naive bayes is good.

Nikhil N. Salvithal, Dr. R. B. Kulkarni,October 2013

Data mining is a technique that uses different types of algorithms to find relationships and trends in large datasets topromote decision support. The data sizes accumulated from various fields are exponentially increasing; Generally arff datasets have 2 types of attributes nominal & numeric. There is need to find suitable classifiers for datasetswith different type of class (either nominal or numeric), so they focused on evaluating performance of different classifiersin WEKA on datasets with numeric & nominal class attribute. During the evaluation, the input datasets and the numberof classifier used are varied to

measure the performance of Data Mining algorithm. Datasets are varied with mainly typeof class attribute either nominal or numeric. They present the results for performance of different classifiers based oncharacteristics such as accuracy, time taken to build model identify their characteristics in acclaimed Data Mining tool-WEKA.

CONCLUSION

In this paper, we have presented a survey of Comparison for classification Techniques using WEKA Tool. A lot of research has been done in this field. Still the work is going on to improve the accuracy of Hemotology and medical field. However the different methods of classifications techniques are very effective and useful for new researchers.

References:

[1]. K. Tamizharasi, Dr. UmaRani, K.Rajasekaran,"Performance analysis of Data Mining algorithms in Weka." IOSR Journal of Computer Engineering (IOSRJCE) ISSN (2014): 2278-0661, Vol.6, Iss.3.

[2]. Zhang, Wenjing, Donglai Ma, and Wei Yao. "Medical Diagnosis Data Mining Based on Improved Apriori Algorithm." Journal of Networks 9, no. 5 (2014): 1339-1345.

[3]. Vaithiyanathan, V., K. Rajeswari, KapilTajane, and Rahul Pitale. "Comparison of Different Classification Techniques Using Different Datasets."Vol.6, no. 2 (2013).

[4]. Ms S. Vijayarani ,Ms M. Muthulakshmi, International Journal of Advanced Research in Computer and Communication Engineering:

"Comparative Analysis of Bayes and Lazy Classification Algorithms". Vol.2, Issue. 8, (2013).

[5]. Nookala, Gopala Krishna Murthy, Bharath Kumar Pottumuthu, NagarajuOrsu,"Performance analysis and evaluation of different data mining algorithms used for cancer classification." International Journal of Advanced Research in Artificial Intelligence (IJARAI) 2, no. 5 (2013).

[6]. Pankajsaxena&sushmalehri, International Journal of Computer & Communication Technology ISSN (PRINT): "Analysis of various clustering algorithms of data mining on health informatics"Vol. 4, Issue. 2. (2013).

[7]. Salvithal, Nikhil N., and R. B. Kulkarni. "Evaluating Performance of Data Mining Classification Algorithm in Weka." Vol2., no. 10 (2013).

[8]. Kaushik H. Raviya, BirenGajjar "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA"Vol. 2, Issue. 1. (2013).

[9]. Sharma, Narendra, AmanBajpai, and MrRatneshLitoriya. "Comparison the various clustering algorithms of weka tools."Volume2, no.5 (2012).

[10]. Rajesh, K., and V. Sangeetha. "Application of data mining methods and techniques for diabetes diagnosis." International Journal of Engineering and Innovative Technology (IJEIT) Volume. 2, Issue. 3 (2012).

