INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

Class based associative classification algorithm for the identification of the Chest Pain using Chest pain dataset

Mr.K.Selvaraj[1], V.Anbarasi[2]

[1]Associate Professor & Head,Department of Computer Science
Arignar Anna Govt.Arts College, Attur,
selvaraj_kumaravel@yahoo.com
[2]Research Scholar, Department of Computer Science
Arignar Anna Govt.Arts College, Attur,
anbarasi.ramya@gmail.com

Abstract: *Data Mining is one of the most motivating area of research that is become increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn helpful for all the parties associated with this field. The mining of the input data is the most important task in all type of the database management. Associative rule mining is commonly used for mining the input data. The relations between the attributes were estimated and based on that the input data can be represented in binary format.The rule based mining can be most useful in medical research inorder to efficiently analyze the symptoms and treatments for the diseases.In the proposed approach a classifier based approach is employed for the classification of the input data based on the Ordered Rule(OR) tree.For generation of rules associative rule mining is employed. The process is done in medical dataset. The Chest pain dataset is selected for this process.The chest pain is normally identified based on the attributes like age, sex, serum cholesterol level, Blood pressure.The main objective of the process is to classify the data based on OR tree.To generate rules for the classification problem based on associate rule mining.To prune the rules based on each attributes of the input data.To check a redundant rule for input database and identify the class for the data inorder diagnose chest pain.To employ Ordered Rule based approach for the identification of the arrangement of the attributes.*

Keywords: *Association, Classification, Clustering and Forecasting.*

1. INTRODUCTION

Data Mining is to store and manage the data in a multidimensional database system by using application software analyze the data, provide data access to business analysts and information technology professionals, present the data in a useful format, like a graph or table[1]. Data Mining is a form of knowledge discovery essential for solving problems in a specific domain. Individual data sets may be gathered and studied collectively for purposes other than those for which they were originally created [2]. Data Mining is a technique which is used to identify relationships between various large amounts of data in many areas include scientific research, business planning traffic analyze, clinical trial data mining, mathematics, cybernetics, genetics and marketing [3]. Data Mining also involves the retrieval and analysis of data that is stored in a Data ware house. Some of the major techniques of Data Mining are Association, Sequence or path analysis, Classification, Clustering and Forecasting etc. Data mining parameters include:

- Association – Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets.
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model.
- Clustering - clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).
- Forecasting - discovering patterns in data that can lead to reasonable predictions about the future.

The main focus of this paper is the classification of different types of datasets that can be performed to determine if a person is diabetic. A diagnosis is a continuous process in which a doctor gathers information from a patient and other sources, like family and friends, and from physical datasets of the patient. The process of making a diagnosis begins with the identification of the patient’s symptoms. The symptoms will be the basis of the hypothesis from which the doctor will start analyzing the patient. This is our main concern, to optimize the

task of correctly selecting the set of medical tests that a patient must perform to have the best, the less expensive and time consuming diagnosis possible. In this paper our main goal is to show the cluster analysis on the diabetes dataset.
The implementation is doing in WEKA tool.

2. LITERATURE SURVEY

Fahab Shahbaz et.al [4] proposed the use of decision trees to extract the clinical reasoning in the form of medical expert's actions that is inherent in large number of EMRs (Electronic Medical records). The extracted data can be used to teach students of oral medicine a number of orderly processes for dealing with patients who represent with different problems within the practice context over time. Shusaka Tsumoto[5] examined the new approach to extract plausible rules which consists of three procedures. They are the characterization of decision attributes and classes. Finally the two kinds of sub-rules, characterization rules and discrimination rules for each class in the group are induced. The two parts are integrated into one rule for each decision attribute. The proposed method correctly represents experts' decision processes. Bing Liu and Wynne Hsu[6] proposed a fuzzy matching technique for rule comparison in the context of classification rules. It allows the user to compare the generated rules in order to find out the correct knowledge and to state that what changes is done during the last learning. This proposed technique plays a major rule in data mining for solving more interest problem. Krishnapriya et al[7] developed a project called classification algorithm which is used to classify the Pima Indian diabetes dataset. Results have been obtained using Android Application.

3. ARCHITECTURE FOR DATA MINING

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses…etc. The knowledge gains a lot of benefits to business strategies, scientific, medical research, governments and individual. Business data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports that help business users to make decisions. Data is now stored in databases and/or data warehouse systems. 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 architecture for advanced analysis in a large data warehouse.

4. APPLICATIONS OF DATA MINING

A wide range of companies have deployed successful applications of data mining [8-10]. The critical factors for success with data mining includes a large well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied.

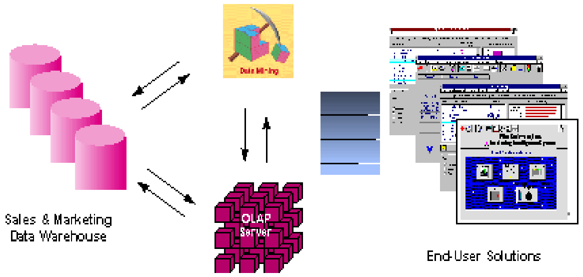

Figure 1: Integrated Data Mining Architecture

Some successful application areas include:

- A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.
- A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.
- A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.
- A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

5. EXISTING SYSTEM

Discover all rules first and then allow the user to query and retrieve those he/she is interested in. The representative approach is that of templates. This approach lets the user to specify what rules he/she is interested as templates. The system then uses the templates to retrieve the rules that match the templates from the set of discovered rules. Use constraints to constrain the mining process to generate only relevant rules. Proposes an algorithm that can take item constraints specified by the user in the association rule mining processor that only those rules that satisfy the user specified item constraints are generated.

6. PROPOSED SYSTEM

In our proposed work, the data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 1865 instances with different attributes. The objective of this study is to evaluate and investigate the classification algorithms based on WEKA. The best algorithm in WEKA is J48 classifier. we had been use the data mining classifiers to generate decision tree format. We identify the diabetic patient's behavior using the classification algorithms of data mining. The analysis had been carried out using a standard blood group data set and using the J48 decision tree algorithm implemented in WEKA. The research work is used to classify the diabetic patient's based on the gender, age, height & weight, blood group, blood sugar(F), blood sugar(PP), urine sugar(F), urine sugar(PP). The J48 derived model along with the extended definition for identifying regular patients provided a good classification accuracy based model

7. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. Our project modules are given below:

7.1 DATA PREPROCESSING

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in medical database is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task.

7. 2 DATA MINING STAGES

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter interesting patterns representing knowledge were identified.

7.3 CLASSIFICATION

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".Popular classification techniques include decision trees and neural networks.

7.4 IDENTIFY THE CLUSTERING STRUCTURE

(**OPTICS**) is an algorithm for finding density-based clusters in spatial data. Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster. This is represented as a dendrogram

8. SOFTWARE REQUIREMENT

Weka is a collection of visualization tools and algorithms for data analysis and predictive modeling. It is used to designed graphical user interfaces for easy access to these functions. The main advantages of Weka include:

- Free availability.
- Portability.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks mainly data preprocessing, clustering, classification, regression, visualization, and feature selection. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

9. ADVANTAGES

Our research work based on the concept from Data Mining is the knowledge of finding out of data and producing it in a form that is easily understandable and comprehensible to humans in general. These further extended in this to make an easier use of the data's available with us in the field of Medicine. The main use of this technique is the have a robust working model of this technology. The process of designing a model helps to identify the different blood groups with available Hospital

Classification techniques for analysis of Blood group data sets. The ability to identify regular diabetic patients will enable to plan systematically for organizing in an effective manner. Development of data mining technologies to predict treatment errors in populations of patients represents a major advance in patient safety research.

10. CONCLUSION AND FUTURE ENHANCEMENT

With the Help of this WEKA tool effective and efficient execution of the Diabetes data set has been done and in future we can extend this work by using other techniques like classification, Association rules etc . not only for this dataset but to any other data sets also The objective of this study is to evaluate and investigate FIVE selected classification algorithms based on WEKA. The best algorithm in WEKA is J48 classifier with an accuracy of 70.59% that takes 0.29 seconds for training. They are used in various healthcare units all over the world. The research work is used to classify the diabetic patient's based on the gender, age, height & weight, blood group, blood sugar(F), blood sugar(PP), urine sugar(F), urine sugar(PP). The J48 derived model along with the extended definition for identifying regular patients provided a good classification accuracy based model. The distribution of blood groups in both positive and negative are shown in Table-1. Overall blood group A was the commonest (24.03 %), followed by B (18.77%), AB (19.11%), O (23.65) and A1B(17.14%). In the present blood group-A was the predominant (24.03%) while A1B was the least common (17.14%). Blood group "A" was the most predominant (24.03%) in both positive and negative subjects, followed by blood group A, B, O, A1B and AB.

11. FUTURE ENHANCEMENT

The future work will be focused on using the other classification algorithms of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning will be explored in future. The future work can be applied to blood groups to identify the relationship that exits between diabetic, diagnosing cancer patients based on blood cells or predicting the cancer types on the blood groups, blood pressure, personality traits and medical.

References

[1] Anand V, Saurkar, Vaibhav Bhujade, Priti Bhagat and Amit Kharplarde, "A Review paper on various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering", ISSN:2277 128X Vol.4,Issue 4,pp. 98-101,April 2014

[2] Sushmita Mitra, Sankark. Pal and Pabitra Mitra, "Data Mining in Soft Computing Framework: A Survey", IEEE Transactions on Neural Networks, Vol.13, No.1, pp.3-14, January 2002.

[3] Chandana Napagoda, "WebSite visit Forecasting Using Data Mining Techniques", International Journal of Scientific and Technology Research, ISSN: 2277-8616, pp.170-174,Vol.2, Issue.12, December 2012.

[4] Fahab Shabhaz Khan et.al, "Data Mining in Oral Medicine Using Decision Trees" ,World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol.2,No.1, pp.113-118, 2008.

[5] Shusaka Tsumoto, "Automated Discovery of Plausible Rules based on Rough Sets and Rough Inclusion", Third Pacific-Asia Conference, PAKDD'99,Beijing, China, April 26-28,1999, Proceedings.

[6] Bing Liu and Wynne Hsu, "Post-Analysis of Learned Rules", From AAA1-96 Proceedings 1996, pp. 828-834.

[7] V. Krishnapriya, Monika, P. Kavitha, "Android Application to Predict and Suggest Measures for Diabetes Using DM Techniques", International Journal of Computer Applications Technology and Research (IJCATR), ISSN:2319-8656, Vol.4, issues. 4, April 2015.

[8] Simmi Bagga & Dr. G. N. Singh, "Applications of Data Mining", International Journal for Science and Emerging Technologies with Latest Trends", Vol.1, Issue 1, pp.19-23, 2012.

[9] Jiban K Pal, "Usefulness and applications of data mining in extracting information from different perspectives", Annuals of Library and Information Studies, Vol. 58, pp.7-16, March 2011.

[10] S. D. Gheware, A. S. Kejkar & S. M. Tondare, "Data Mining: Task, Tools, Techniques and Applications", International Journal of Advanced Research in Computer and Communication Engineering, Vol.3, Issue. 10, October 2014.