Analysis of a Population of Diabetic Patients Databases in Weka Tool

P.Yasodha, M. Kannan

Abstract - Data mining is an important tool in many areas of research and industry. Companies and organizations are increasingly interested in applying data mining tools to increase the value added by their data collections systems. Nowhere is this potential more important than in the healthcare industry. As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed. Taking into account the prevalence of diabetes among men and women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using weka tool. In this paper the data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 249 instances with 7 different attributes. The instances in the Dataset are pertaining to the two categories of blood tests, urine tests. WEKA tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

Keywords- Data Mining, Diabetics data, Classification algorithm, Association algorithm Weka tool

1. Introduction

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. The solution for this problem will also include the cost of the different types of datasets. For this reason, the goal of this paper is classifier in order to correctly classify the datasets, so that a doctor can safely and cost-effectively select the best datasets for the diagnosis of the disease. The major motivation for this work is that diabetes affects a large number of the world population and it's a hard disease to diagnose. 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. A solution like this one, will not only assist doctors in making decisions, and make all this process more agile, it will also reduce health care costs and waiting times for the patients. This paper will focus on the analysis of data from a data set called Diabetes data set.

2. Related Work

The few medical data mining applications as compared to other domains. [4] reported their experience in trying to automatically acquire medical knowledge from clinical databases. They did some experiments on three medical databases and the rules induced are used to compare against a set of pre-defined clinical rules. Past research in dealing with this problem can be described with the following approaches: (a) 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 [3]. 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. (b) Use constraints to constrain the mining process to generate only relevant rules. [12] proposes an algorithm that can take item constraints specified by the user in the association rule mining process so that only those rules that satisfy the user specified item constraints are generated. This also does not work well because doctors often do not have any specific rules to mine. (c) Find unexpected rules. This approach first asks the user to specify his/her existing knowledge about the domain. The system finds those unexpected rules [5, 6, 13].

3. Data Mining In The Expert System

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different

dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. In our thesis, some data mining methods on social-demographic data of the users were applied. Correction and actuality of the data is very important for data mining for diabetes.

3.1 Data Mining by WEKA Engine

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is developed by the University of Waikato. In our paper we used the Weka as data mining engine, and made a bridge between the Diabetes Expert System interface and Weka. No user needs to install Weka in his workstation, but it do already enough to be installed on server machine.

3.1.1 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.

Bayes Network Classifier

Bayes Network learning uses various searching algorithms and quality measures. This is the base module for a Bayes Network Classifier, and also provides data structures (network structure, conditional probability distributions, etc.) and facilities common to Bayes Network learning algorithms like K2 and B

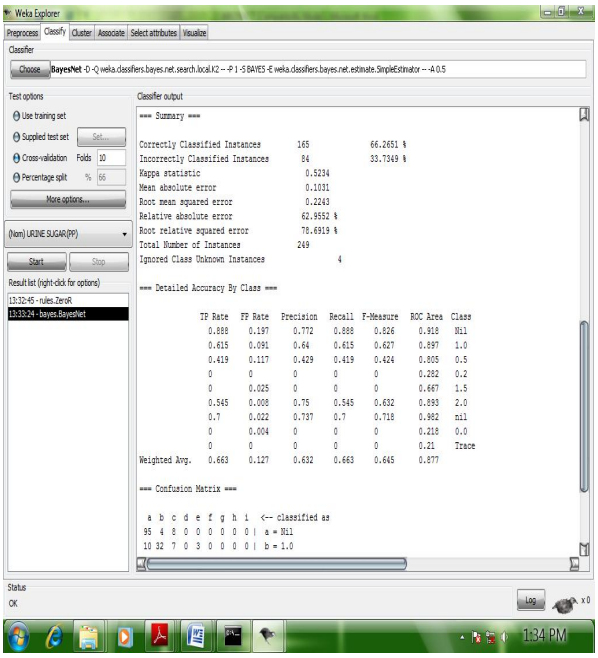

Figure1 BayesNet

J48 Pruned Tree

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When we applied J48 onto refreshed data, we got the results shown as below on Figure2 .

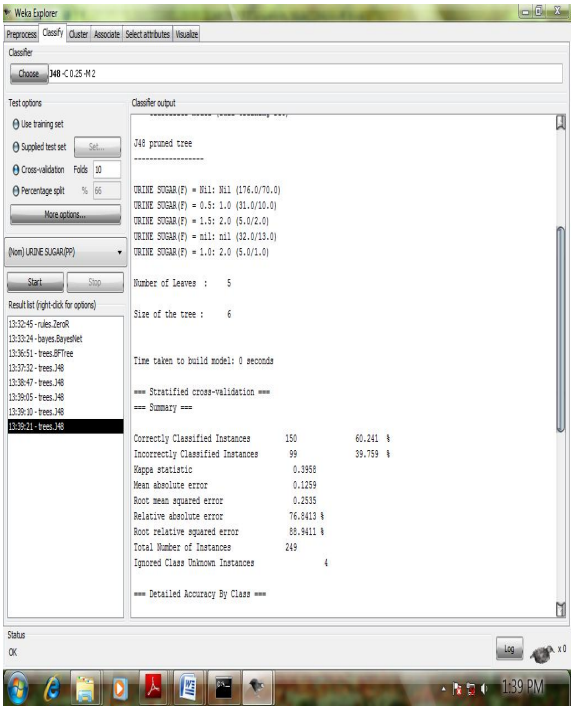

Figure2 J48 Tree

REP Tree

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

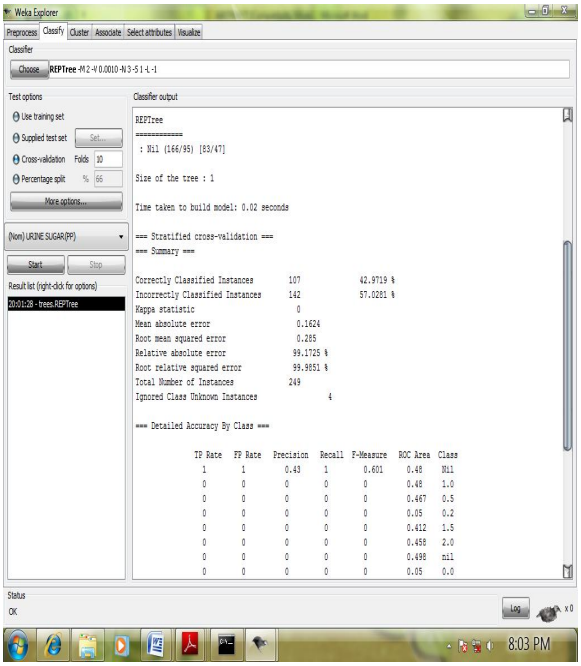

Figure3 REP tree

Random Tree

Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities based on a hold-out set (backfitting).

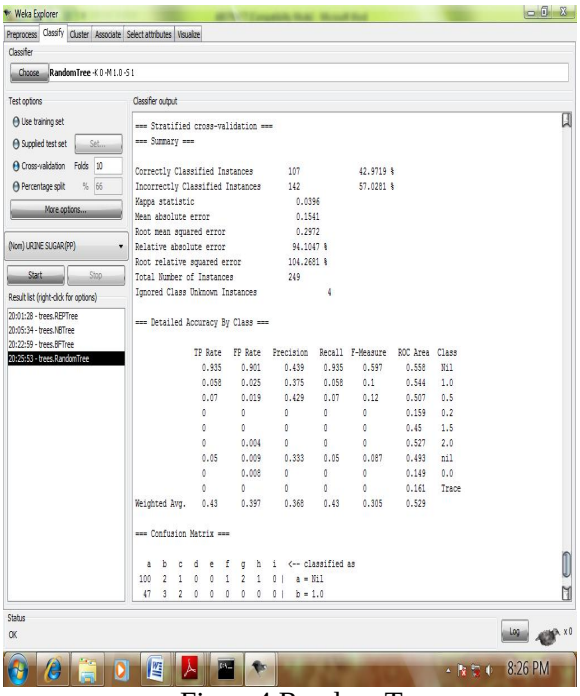

Figure4 Random Tree

3.1.2 Association Rules

One of the reasons behind maintaining any database is to enable the user to find interesting patterns and trends in the data. For example, in a supermarket, the user can figure out which items are being sold most frequently. But this is not the only type of `trend' which one can possibly think of. The goal of database mining is to automate this process of finding interesting patterns and trends. Once this information is available, we can perhaps get rid of the original database. The output of the data-mining process should be a "summary" of the database. This goal is difficult to achieve due to the vagueness associated with the term `interesting'. The solution is to define various types of trends and to look for only those trends in the database.

Apriori

Apriori is a module implementing an Apriori-type algorithm. Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules. It is adapted as explained in the second reference. Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association

rules in data having no transactions. As is common in association rule mining, given a set of itemsets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found (Agrawal et al. 1994).

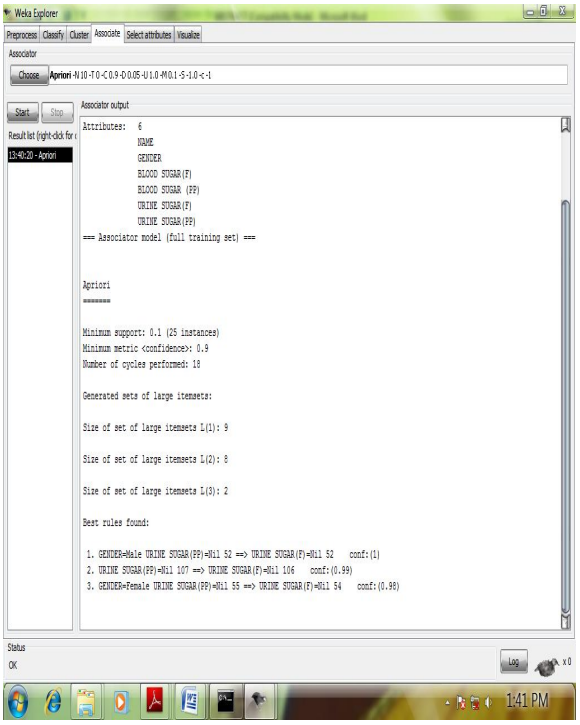

Figure5 Apriori

4. RESULTS & DISCUSSIONS

The main purpose of the system is to guide diabetic patients during the disease. Diabetic patients could benefit from the diabetes expert system by entering their daily glucoses rate and insulin dosages; producing a graph from insulin history; consulting their insulin dosage for next day. The diabetes expert system is not only for diabetic patient, but also for the people who suspect if they are diabetic. It's also tried to determine an estimation method to predict glucose rate in blood which indicates diabetes risk. In our Paper, the target class consists of 249 men and women who are actually affected with diabetes. However, it is unknown that in what stage the disease is. Hence this study will not only help in estimating the maximum number of people suffering from diabetes with specific characteristics but also helps in identifying the state of the disease.

REFERENCES

[1] Mats Jontell, Oral medicine, Sahlgrenska Academy, Göteborg University (1998) "A Computerised Teaching Aid in Oral Medicine and Oral Pathology. " Olof Torgersson, department of Computing Science, Chalmers University of Technology, Göteborg.

[2] T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning (1997) the McGraw-Hill Companies, Inc., pp. 52-78.

[3] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

[4] Tsumoto S., (1997)"Automated Discovery of Plausible Rules Based on Rough Sets and Rough Inclusion," Proceedings of the Third Pacific-Asia Conference (PAKDD), Beijing, China, pp 210-219.

[5] Liu B., Hsu W., (1996) "Post-analysis of learned rules," AAAI, pp. 828-834.

[6] Liu B., Hsu W., and Chen S., (1997) "Using general impressions to analyze discovered classification rules," Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[7] Stutz J., P. Cheeseman. (1996) Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

[8] Witten Ian H., E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Ch. 8, © 2000 Morgan Kaufmann Publishers

[9] http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.

[10] http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_T rees.html, accessed

[11] Wikipedia, ID3-algorithm (accessed 2007/12/09) (URL: http://en.wikipedia.org/wiki/ID3_algorithm)

[12] Srikant,R.,Vu,Q.andAgrawal,R.,(1997), "Mining association rules with item constraints," Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, USA, pp 67-73.

AUTHOR PROFILE:

P.Yasodha Mphil Research Scholar in the Department of Computer Science & Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Enathur, Kanchipuram –631 561. She received the degree in Master of Computer Applications from SCSVMV University in 2010. Her research interest lies in the area of Data Mining and Artificial Intelligence.

M.Kannan has been working as a Assistant Professor and Ph.D Research Scholar in the Department of Computer Science & Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, Enathur, Kanchipuram –631 561. He received the degree in Master of Computer Applications from Bharathidasan University in 2001 and M.Phil(Computer Science) from Madurai Kamaraj University in 2005. His research interest includes Software Engineering & Data Mining.