Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Improving Performance of Diagnosis System for Diabetes Using Data Mining Techniques

D.Sheila Freeda
Reseach Scholar, Bharathiyar University

Dr. Lilly Florence,
Professor, Adhiyamman College of Engineering.

Abstract *- Traditionally Diabetics has been diagnosed using overall cholesterol testing and a detailed lab test for individualcholesterols (HDL/LDL/ VLDL) etc and usually doctors asses the risk and recommend cholesterol test based on age, hereditary, High blood pressure , heart disease, stroke etc and it is usually predicted using these factors as well as environment and life style. However, there could be more factors and relationships both in the way to discover diabetics as well as analysis of effective cure. For that datamining techniques and tools can be used to bring out all the relationships and changing patterns. Datamining is a effective way to analyses and predict structured and unstructured data using techniques such as clustering, association, classification prediction and visualization. In tis paper we are analyzing the previous researches and relationships and the various opportunity that exists in improving the diagnosis as well as effective treatment using data mining tools and techniques.*

Keywords— *Data Mining, Cholesterol, Heart Disease, Diabetes.*

I. INTRODUCTION

Diabetes is a condition that prevents the body from properly using energy from food. It occurs when the pancreas does not produce insulin, or when the pancreas produces insulin, but it is resisted by the body.

Diabetes mellitus is classified into four broad categories: type 1, type 2, gestational diabetes, and "other specific types".[5] The "other specific types" are a collection of a few dozen individual causes.[5] The term "diabetes", without qualification, usually refers to diabetes mellitus.

A. Types of Diabetes

Type 1 diabetes can develop at any age but usually appears before the age of 40, and especially in childhood. It is the most common type of diabetes found in childhood.Type 2 diabetes happens If there is not enough insulin or the insulin is there but not working properly. Type 2 diabetes develops when the insulin-producing cells in the body are unable to produce enough insulin, or when the insulin that is produced does not work properly (known as insulin resistance). So the cells are only partially unlocked and glucose builds up in the blood.

Type 2 diabetes usually appears in people over the age of 40, though in South Asian people, who are at greater risk, it often appears from the age of 25. It is also increasingly becoming more common in children, adolescents and young people of all ethnicities. Type 2 diabetes accounts for between 85 and 95 per cent of all people with diabetes and is treated with a healthy diet and increased physical activity.

Type 3 Gestational diabetes, is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood sugar level. Prevention and treatment involve a healthy diet, physical exercise, not using tobacco and being a normal body weight. Blood pressure control and proper **FOOT** care are also important for people with the disease. Type 1 diabetes must be managed with insulininjections.[5] Type 2 diabetes may be treated with medications with or without insulin. Insulin and some oral medications can cause low blood sugar. Weight loss surgery in those with obesity is sometimes an effective measure in those with type 2 DM. Type 3 Gestational diabetes usually resolves after the birth of the baby.

B. Diagnosis of Diabetes

A physician will recognize the classic symptoms of type 1 diabetes quite easily and order the proper blood, urine and insulin tests[6]. These tests are very simple and quite painless and will traditionally provide a complete analysis and diagnosis in a very short period of time.

Laboratory tests — Several blood tests are used to measure blood glucose levels, the primary test for diagnosing diabetes.

●Random blood sugar test – For a random blood sugar test, you can have blood drawn at any time throughout the day, regardless of when you last ate. If your blood sugar is 200 mg/dL (11.1 mmol/L) or higher and you have symptoms of high blood sugar (see'Symptoms' above), it is likely that you have diabetes.

●Fasting blood sugar test – A fasting blood sugar test is a blood test done after not eating or drinking for 8 to 12 hours (usually overnight). A normal fasting blood sugar level is less than 100 mg/dL (5.55 mmol/L).

●Hemoglobin A1C test – The "A1C" blood test measures your average blood sugar level over the past two to three months. Normal values for A1C are 4 to 5.6 percent. The A1C test can be done at any time of day (before or after eating).

●Oral glucose tolerance test – Oral glucose tolerance testing (OGTT) is a test that involves drinking a special glucose solution (usually orange or cola flavored). Your blood sugar level is tested before you drink the solution, and then again one and two hours after drinking it.

C. Risk Factors for Diabetes

Everyone should have their first screening test by age 35 for men, and age 45 for women. Some guidelines recommend starting at age 20.[6]You should have a test done at an earlier age if you have Cholesterol, Heart disease, Stroke, High blood pressure, A strong family history of heart disease

From this we derive that there are lot of relationship between diabetes, age, demography, lifestyle, cholesterol, heart disease etc. There are few other studies using datamining have establish relationship between diabetes and BMI. In this paper we examine previous studies using datamining to discover the various relationships and examine ways to improve and find additional relationship that may exist.

D. DataMining

Datamining is “The Process of Discovering meaningful, new correlation patterns and trends by shifting through large amount of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques”.Datamining is expected to lead to

- Discovering unknown associations
- Sequences where one event lead to another event
- Recognizing patterns
- Finding out facts previously not known (Clustering)
- Forecasting or prediction
- Determining the significant changes in key measures (Deviation Detection)

II. RELATED WORK

K. Rajesh et al in their research work applied datamining techniques for diabetes diagnosis and applied various comparison classification algorithm and analyzed their performance in one of the algorithm using Decision Tree induction learning technique it gave 91% accuracy of classification rate. RupaBagdi et al their paper used OLAP and datamining integration for diagnosis of diabetes to predict patient who might be diagnosis diabetes using IT3 and C4.5 decision tree algorithms to classify probability high, low or medium for a patient to be diagnosis diabetes. V.V. Jaya Ramakrishna et al used Duo Mining approach for predicting diabetes using BMI index and calculating Fat Test however in vast amount of structured and unstructured biological data there may be many associations and sequencing and patterns may remain to be detected. So in this paper we are proposing some of the tools and techniques that can be used on the data set to find out facts that previously may not be known or confirm facts which doctors use based on experience as a fact. We have analyzed following tools for further studies.

A. Datamining Tools

1. Rapid Miner

Rapid Miner is free open source datamining tool which has got classification discovery, cluster discovery, association discovery, regression discovery, Text mining, outlier discovery and visualization.

2. R

R is free open source dataminingtool which has got classification discovery, cluster discovery, association discovery, regression discovery, Text mining, outlier discovery and data visualization, discovery visualization, sequence analysis, social network analysis.

3. WEKA

WEKA is free open source datamining tool which has got classification discovery, cluster discovery, association discovery, regression discovery, outlier discovery and data visualization, discovery . visualization.

III. TASK OF DATA MINING

The first three tasks - classification, estimation and prediction rules are examples of directed data mining or supervised learning. The next three tasks – association rules, clustering and description are examples of undirected data mining.

A. Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class[1]. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

• Classification of credit applicants as low, medium or high risk
• Classification of mushrooms as edible or poisonous

• Determination of which home telephone lines are used for internet access

B. Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance[1]. Some examples of estimation tasks include:

• Estimating the number of children in a family from the input data of mothers’ education

• Estimating total household income of a family from the data of vehicles in the family

• Estimating the value of a piece of a real estate from the data on proximity of that land from a major business center of the city.

C. Prediction

Any prediction can be thought of as classification or estimation. Our classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the efforts, it is possible to check[1]. Predictive tasks feel different because the records are classified according to some predicted future behavior or way to check theaccuracy of the classification is to wait and see. Examples of prediction tasks include:

• Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer

• Predicting which customers will leave within next six months

• Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail.

D. Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database[1]. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

E. Clustering

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a pre-processing step for other data mining algorithms operating on the detected clusters[1]. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods .Further data set can be numeric or categorical. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self similarity. Clustering is often done as a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster[6].

IV. TOOLS

4.1. WEKA :[4]

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Make file-based system for running machine learning experiments. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- 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, more specifically,data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).

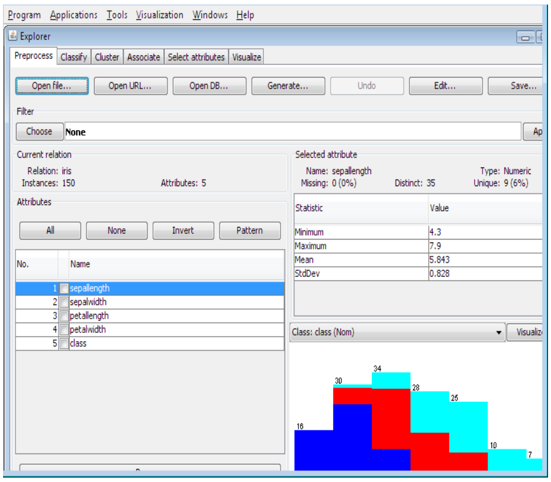
The *Explorer* interface features several panels providing access to the main components of the workbench:

- The *Preprocess* panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for preprocessing this data using a so-called *filtering*algorithm. The *Classify* panel enables applying classification and regression algorithms (indiscriminately called *classifiers* in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).

- The *Associate* panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.

- The *Cluster* panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The *Select attributes* panel provides algorithms for identifying the most predictive attributes in a dataset.

- The *Visualize* panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.



Filter: None

Current relation
Relation: iris
Instances: 150
Attributes: 5

Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Attributes

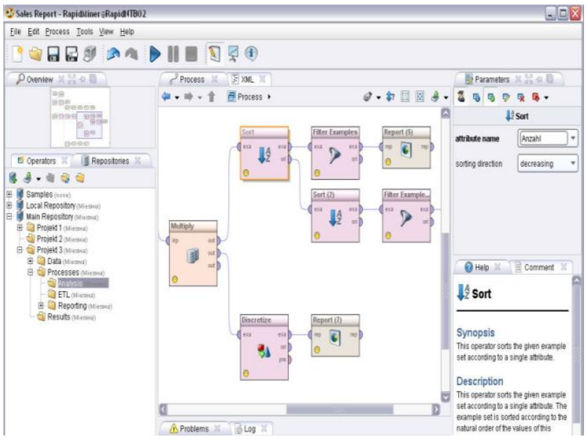
No.		Name
1	□	sepallength
2	□	sepalwidth
3	□	petallength
4	□	petalwidth
5	□	class

Class: class (Nom)

4.2. Rapid Miner :[4]

RapidMiner is a software platform developed by the company of the same name that provides an integrated environment formachine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization. RapidMiner is developed on a business source model which means only the previous version of the software is available under an OSI-certified open source license on Sourceforge.

RapidMiner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures. RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical workflows. Those workflows are called “Process” in RapidMiner and they consist of multiple “Operators”. Each operator is performing a single task within the process and the output of each operator forms the input of the next one. Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. RapidMiner provides learning schemes and models and algorithms from Weka and Rscripts that can be used through extensions.



Sort

attribute name: Anzahl

sorting direction: decreasing

Sort

Synopsis

This operator sorts the given example set according to a single attribute.

Description

This operator sorts the given example set according to a single attribute. The example set is sorted according to the natural order of the values of this

4.3 R :[4]

R is primarily written in C and Fortran. And a lot of its modules are written in R itself. It’s a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R’s popularity substantially in recent years.

Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

Using various Datamining Tools we can find out the knowledge. We can apply Diabetes Data Set on any one of the datamining tools and apply various data mining techniques finally find out the useful pattern.

CONCLUSION

As the amount of unstructured data in our world continues to increase, text mining tools that allow us to sift through this information with ease will become more and more valuable. Lot of biomedical data available and lot of research has also happened on that data using datamining techniques like decision trees to predict diabetes. However all datamining techniques have not been applied structured and unstructured data. Hence lot of opportunities exists in analyzing the data using data mining tools and techniques to improve diagnosis of diabetes.

References

[1]. Donald Michie,Data Mining DiscoveringInteresting Relationships in Large Data Sets,Retrieved from http://www.aaai.org /aitopics/pmwiki/ pmwiki.php/AITopics/DataMining

[2]. Wikipedia, free encyclopedia, Data mining,Retrieved from http://en.wikipedia.org/ wiki/Data_mining

[3]. Mining Biomedical Literature Using Information Extraction ,Ronen Feldman, YizharRegev, Michal Finkelstein-Landau, EyalHurvitz& Boris KoganClearforest Corp, USA & Israel

[4]. Wikipedia, free encyclopedia, web mining, Retrieved from http://en.wikipedia.org/ wiki/ Web_mininghttp://en.wikipedia.org/wiki/Diabetes_mellitus

[5]. Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, Lawlor DA, Selvin E, Stampfer M, Stehouwer CD, Lewington S, Pennells L, Thompson A, Sattar N, White IR, Ray KK, Danesh J (2010). "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies". *The Lancet* **375** (9733): 2215–22. doi:10.1016/S0140-6736(10)60484-9. PMC 2904878. PMID 20609967.

[6]. V.V.Jaya Rama Krishnaiah1, D.V.Chandra Sekhar2, Dr. K.Ramchand H Rao3, Dr. R Satya Prasad4 “Predicting the Diabetes Using Duo Mining Approach”, ***International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012***

[7]. BangaruVeera Balaji1, VedulaVenkateswara Rao2, “Improved Classification Based Association Rule Mining”, ***International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013.***