Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MEDICAL DIAGNOSIS SYSTEM FOR THE DIABETES MELLITUS BY USING BACK PROPAGATION-APRIORI ALGORITHMS.

[1]K. SRIDAR M.E., [2]Dr. D. SHANTHI M.E. Ph.D.,

[1]Research Scholar, Department of CSE, Odaiyappa College of Engg & Tech., Theni, Tamilnadu, India.

[2]Professor and Head, Department of CSE, PSNA College of Engg & Tech., Dindigul, Tamilnadu, India.

E-mail: [1]sridar.krishnan1@gmail.com, [2]dshan71@gmail.com

ABSTRACT

Diabetes is a chronic disease and a major public health challenge worldwide. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 430 million by 2030. Furthermore, 3.8 million deaths are attributable to diabetes complications each year. It has been shown that 80% of type 2 diabetes complications can be prevented or delayed by early identification of people at risk. Early detection of diabetes would be of great value given the fact that at least 50% and 80% in some countries, of all people with diabetes are unaware of their condition and will remain unaware until complications appear. Several data mining and machine learning methods have been used for the diagnosis, prognosis, and management of diabetes. In my research work, clinical data are collected based on the attributes downloaded from Pima Indians Diabetes Database. Real time input have been given to the system from the Glucometer (ie Glucose level for the patient before breakfast and Glucose level for the patient two hours after breakfast) then some on the input attributes given manually to the system. All the inputs have been given to the BP Algorithm (ANN) and Apriori Algorithm (ARM) for diagnosing diabetes. The Diagnosis system can be implemented in Java, Dotnet. The output of the system shows how much percentage of the patient suffers from the diabetes (low, medium or high risks)? In proposed work it is decided to implement the system in Online. The main objective is that the patients can know about their risk in diabetes without the help of doctors. The patient can just login into the website and they can give their attributes (ie) data collected from labs as input and they can diagnosis their own result without doctors.

Keywords: *Artificial Neural Network (ANN), Back Propagation Algorithm, Association Rule Mining (ARM), Apriori Algorithm, Disease Diagnosis.*

1. INTRODUCTION

Diabetes is a condition primarily defined by the level of hyperglycaemia giving rise to risk of microvascular damage (retinopathy, nephropathy and neuropathy). It is associated with reduced life expectancy, significant morbidity due to specific diabetes related microvascular complications, increased risk of macrovascular complications (ischaemic heart disease, stroke and peripheral vascular disease), and diminished quality of life. The criteria for the diagnosis of diabetes and Impaired Glucose Tolerance (IGT) remained unchanged, the ADA recommended lowering the threshold for Impaired Fasting Glucose (IFG) from 6.1mmol/l (110mg/dl) to 5.6mmol/l (100mg/dl)4 [4]. In view of these developments WHO and the International Diabetes Federation (IDF) decided that it was timely to review its existing guidelines for the definition and diagnosis of diabetes and intermediate hyperglycaemia.

There are two types of Diabetes (i.e) Type 1 Diabetes and Type 2 Diabetes. One reason for hyperglycaemia is insulin deficiency, where beta cells in the pancreas fail to produce enough insulin. This is known as type 1 diabetes. The other and most common type of diabetes is recognized as type 2, where the body cannot effectively use the insulin produced.[5] In this context, several data mining and machine learning methods have been proposed for the diagnosis and management of diabetes. In this study it is decided to fuse Artificial Neural Network (ANN) with Association Rule Mining (ARM) for better accuracy. The proposed study is to Design and Develop a Web based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Artificial Neural Network with Association Rule Mining for the diagnosis of Diabetic Disease.

The main objective is that the patients can know about their risk in diabetes without the help of doctors. The patient can just login into the website and they can give their attributes (ie) data collected from labs as input and they can diagnosis their own result without doctors. Artificial Neural Networks consists of three layers: input, hidden and output units (variables). Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.[6].

2. RESEARCH OBJECTIVE

The main objective of this research is to Design and Develop a Web based Artificial Neural Network with Association Rule Mining for the diagnosis of Diabetic Disease. It can answer complex queries for diagnosing Diabetes with greater accuracy and low cost.

3. EXISTING WORK

Hypoglycemia or low blood glucose is dangerous and can result in unconsciousness, seizures, and even death. It is a common and serious side effect of insulin therapy in patients with diabetes. Hypoglycemic monitor is a noninvasive monitor that measures some physiological parameters continuously to provide detection of hypoglycemic episodes in type 1 diabetes mellitus patients (T1DM). Based on heart rate (HR), corrected QT interval of the ECG signal, change of HR, and the change of corrected QT interval, we develop a genetic algorithm (GA)-based multiple regression with fuzzy inference system (FIS) to classify the presence of hypoglycemic episodes. GA is used to find the optimal fuzzy rules and membership functions of FIS and the model parameters of regression method. From a clinical study of 16 children with T1DM, natural occurrence of nocturnal hypoglycemic episodes is associated with HRs and corrected QT intervals. The overall data were organized into a training set (eight patients) and a testing set (another eight patients) randomly selected.[7][8].

4. PROPOSED STUDY

4.1 Data Source

The dataset is downloaded from Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm. Number of Instances in dataset is 768.

The level of glucose present in the patient body is measured with the help of glucometer, the reading which came from the glucometer is taken as the real time input for our system. In addition to this some of the input parameters like Age, Sex, Family history of diabetes, Body mass index (BMI), Waist circumference in centimetre (waist), Hip circumference, Systolic blood pressure, Diastolic blood pressure (BPDIAS), Cholesterol, Fasting blood sugar (FBS) can be given manually.

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

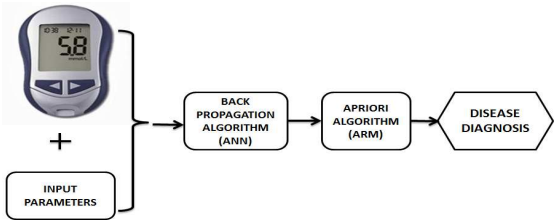

Fig. 1. Block Diagram

The block diagram in Fig. 1 shows the working mechanism of the research. Blood glucose level can be found with the help of glucometer with that reading some of the input parameters listed above can be given as inputs to the Back propagation algorithm and then Apriori algorithm is fused with back propagation for disease diagnosis.

4.2 Artificial Neural Network

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.[9].

The commonest type of artificial neural network consists of three groups, or layers, of units: a layer of "input" units is connected to a layer of "hidden" units, which is connected to a layer of "output" units. Figure.2. displays the three layers viz., input, hidden, output layer.[9].

- The activity of the input units represents the raw information that is fed into the network.
- The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.
- The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

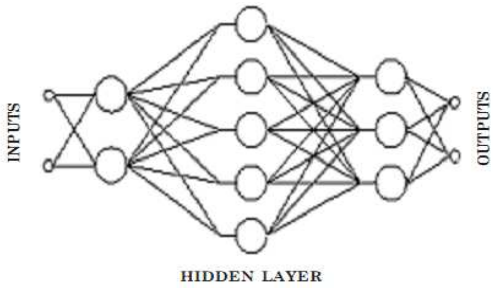

Fig.2. A Three Layer Back Propagation ANN Network

4.2.1 Back propogation algorithm

Back propagation, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. The back propagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN learns the training data.[9].

Actual Algorithm:

1. Initialize the weights in the network (often randomly)

2. repeat

* for each example e in the training set do

1. O = neural-net-output (network, e);

forward pass

2. T = teacher output for e

3. Calculate error (T - O) at the output units

4. Compute delta_wi for all weights from hidden layer to output layer ;

backward pass

5. Compute delta_wi for all weights from input layer to hidden layer ;

backward pass continued

6. Update the weights in the network

* end

3. until all examples classified correctly or stopping criterion satisfied

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. return (network)

4.2.2. Advantages on Artificial Neural Network

1. High Accuracy: Neural networks are able to approximate complex non-linear mappings.

2. Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing and noisy data.

3. Independence from prior assumptions: Neural networks do not make a priori assumptions about the distribution of the data, or the form of interactions between factors.

4. Ease of maintenance: Neural networks can be updated with fresh data, making them useful for dynamic environments.

5. Neural networks can be implemented in parallel hardware.

6. When an element of the neural network fails, it can continue without any problem by their parallel nature.

4.3 Association Rule Mining

A well-known approach to Knowledge Discovery in Databases involves the identification of association rules linking database attributes. Extracting all possible association rules from a database, however, is a computationally intractable problem, because of the combinatorial explosion in the number of sets of attributes for which incidence-counts must be computed. Algorithms for using this structure to complete the count summations are discussed, and a method is described, derived from the well-known Apriori algorithm.[10]

An association rule is a probabilistic relationship, of the form $A \rightarrow B$, between sets of database attributes, which is inferred empirically from examination of records in the database. In the simplest case, the attributes are boolean, and the database takes the form of a set of records each of which reports the presence or absence of each of the attributes in that record. An association rule R is of the form $A \rightarrow B$, where A, B are disjoint subsets of the attribute set I . The support for the rule R is the number of database records which contain $A \cup B$ (often expressed as a proportion of the total number of records). The confidence in the rule R is the ratio:

$$\frac{\text{support for R}}{\text{support for A}}$$

These two properties, support and confidence, provide the empirical basis for derivation of the inference expressed in the rule, and a measure of the interest in the rule. The support for a rule expresses the number of records within which the association may be observed, while the confidence expresses this as a proportion of the instances of the antecedent of the rule. In practical investigations, it is usual to regard these rules as "interesting" only if the support and confidence exceed some threshold values. Hence the problem may be formulated as a search for all association rules within the database for which the required support and confidence levels are attained. Note that the confidence in a rule can be determined immediately once the relevant support values for the rule and its antecedent are computed. Thus the problem essentially resolves to a search for all subsets of I for which the support exceeds the required threshold. Such subsets are referred to as "large", "frequent", or "interesting" sets.

4.3.1 Algorithm for calculating support

It is instructive to begin by examining methods for computing the support for all subsets of the attribute set I . While such "brute force" algorithms are infeasible in most practical cases, they provide a benchmark against which to measure alternatives. The simplest such algorithm takes the form:

Algorithm BF:

∀ records in database do

begin ∀ subsets of record do

add 1 to support—count for that subset;

end

Three properties of this algorithm are of interest in considering its performance:

1. The number of database accesses required.

2. The number of computation steps involved in counting subsets of records.

3. The memory requirement.

For a database of m records, the algorithm involves a single database pass requiring m database accesses to retrieve the records. Assuming the obvious binary encoding of attributes in a record, the enumeration of subsets is straightforward, and will involve a total of up to $m \times 2n$ computation steps, the actual number depending on the number of attributes present in each record. For each subset, the corresponding

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

binary encoding may be used to reference a simple array of support-counts, so the incrementation is trivial, but will require an array of size 2n.[11].

For small values of n, this algorithm is simple and efficient. However, in applications such as shopping-basket analysis values of n in excess of 1000 are likely, making brute-force analysis computationally infeasible. Even in cases in which the actual number of attributes present in a record is small, making exhaustive enumeration practicable, the size of the support-count array is a limiting factor.

For these reasons, practicable algorithms for determining association rules proceed in general by attempting to compute support-counts only for those sets which are identified as potentially interesting, rather than for all subsets of I .To assist in identifying these candidates, it is helpful to observe that the subsets of the attribute set may be represented as a lattice. One form of this is shown as figure 3, for a set of four attributes, {A, B, C, D}.

For any set of attributes to be interesting, i.e. to have support exceeding the required threshold level, it is necessary for all its subsets also to be interesting. For example, a necessary (although not sufficient) condition for ABC to be an interesting set is that AB, AC and BC are all interesting, which in turn requires that each of A, B and C are supported at the required level. This observation, originally made in Mannila et al. (1994) and Agrawal and Srikant (1994), provides a basis for pruning the lattice of subsets to reduce the search space; if, for example, it is known that D is not supported, then it is no longer necessary to consider AD, BD, CD, ABD, ACD, BCD or ABCD. Algorithms which proceed on this basis reduce their requirement for storage and computation by eliminating candidates for support as soon as they are able to do so. The tradeoff for this is usually a greater requirement for access to the database in a series of passes.

4.3.2 Apriori Algorithm

Apriori Algorithm is used to find associations between different sets of data. Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.[12].

The associations that Apriori finds are called Association rules. An association rule has two parts. The Antecedent is a subset of items found in sets of data. The Consequent is an item that is found in combination with the antecedent. Two terms describe the significance of the association rule. The Confidence is a percentage of data sets that contain the antecedent. The Support is a percentage of the data sets with the antecedent that also contain the consequent.

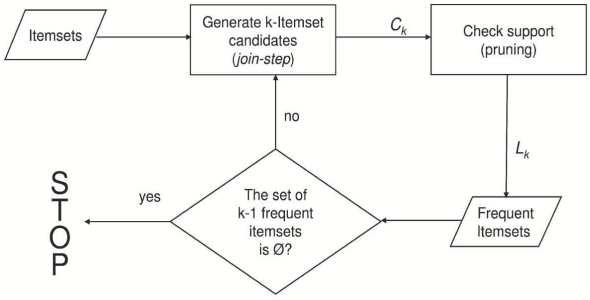

Fig. 4. Apriori Algorithm workflow

Steps

1. Load the general dataset
2. Give the input parameters (i.e MinSupp, No. of Transactions)
3. Creation of temporary dataset with respect to input parameters
4. Extraction of frequent itemsets using Apriori
5. Split the dataset with respect to user input
6. Extract frequent itemsets from each of the split dataset
7. Clustering Process

Merge the frequent items generated from general dataset and actionable dataset

8. Rule Generation for the merged dataset.

Pseudo-code:

C_k : Candidate itemset of size k

L_k : Frequent itemset of size k

L1 ={ frequent items };

for (k = 1; L_k != Ø; K++) do begin

C_{k+1} = Candidates generated by L_k;

For each transaction t in database do

Increment the count of all candidates C_{k+1} that are contained in t

Lk+1 = Candidate in Ck+1 with min_support

End

Return $\cup_k L_k$

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. EXPERIMENTAL RESULT

I have undergone my research work with Java Net beans IDE and SQL 2005 as backend to store the database. The following result has been obtained as result for 786 instances with 9 attributes. The last one is consider as class attribute which clearly shows whether the patient is having high, medium, and low risks.

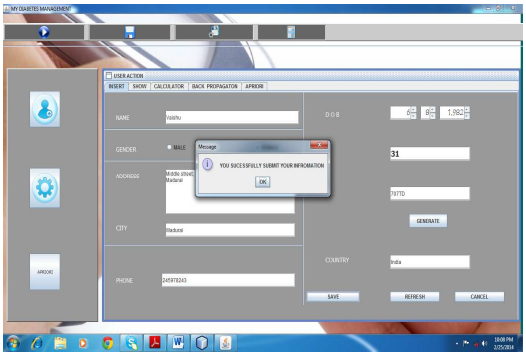

Fig 5.1 Patient Login

In the Patient Login(Fig 5.1) all the patient will give their personal details such as Name, Address, Sex, Age, Phone Number, Date of Birth etc., After entering their own data Unique Patient ID will be generated randomly for each patient.

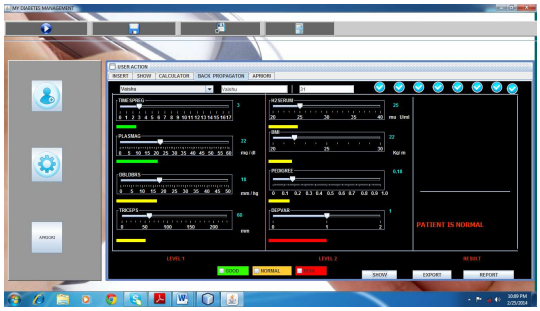

Fig 5.2 Input for Back Propagation Algorithm

As I told earlier there are 8 input parameters to be used as input. (Fig 5.2) For Back propagation algorithm Layer 1 consist of 4 input parameter such as No of times pregnant, Plasma glucose level, Diastolic blood pressure, Triceps skin fold thickness. Layer 2 consists of 4 input parameters such as 2-Hour serum insulin, Body mass index, Diabetes pedigree function, Age. Back propagation algorithm process these 2 input layers and produces the result that the patient is suffering from High diabetic, Normal diabetic, and No diabetic.

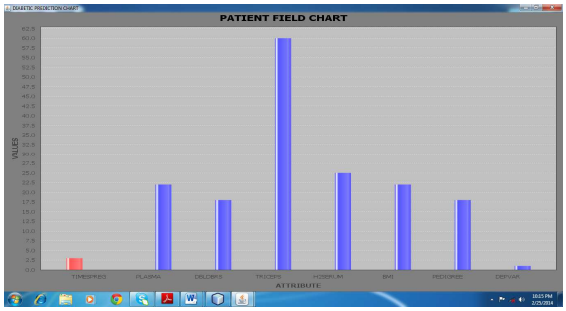

Fig 5.3 Patient field Chart

Patient field chart is drawn based on the Input attributes and values for each attributes. (Fig 5.3)

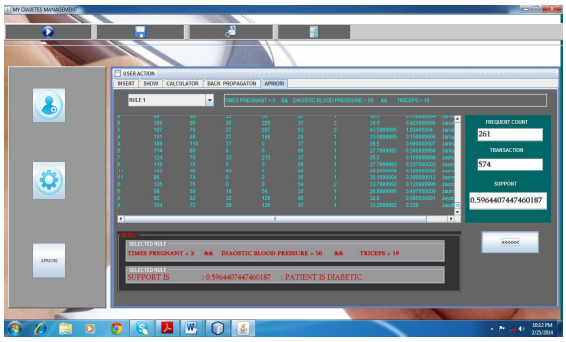

Fig 5.4 Output of Rule 1 Apriori Algorithm

According to Rule 1. 3 input parameters are taken No of times pregnant <3, Diastolic blood pressure>60, Triceps skin fold thickness>23. According to Rule 1 Frequent Count is 291, No of transaction is 574, Minimum support is 0.596. The output of the Rule 1 shows the patient is Diabetic.

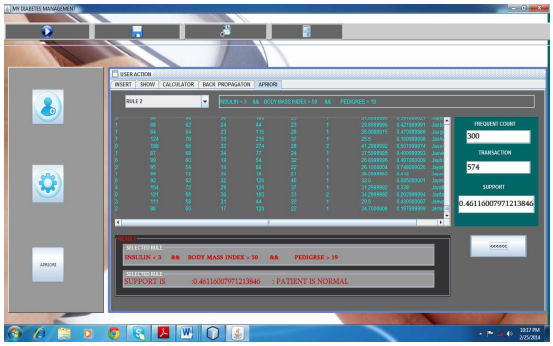

Fig 5.5 Output of Rule 2 Apriori Algorithm

According to Rule 2. 3 input parameters are taken 2-Hour serum insulin>150, Body mass index>18, Diabetes pedigree function>0.5. According to Rule 1 Frequent Count is 300, No of transaction is 574, Minimum support is 0.461. The output of the Rule 2 shows the patient is Normal.

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

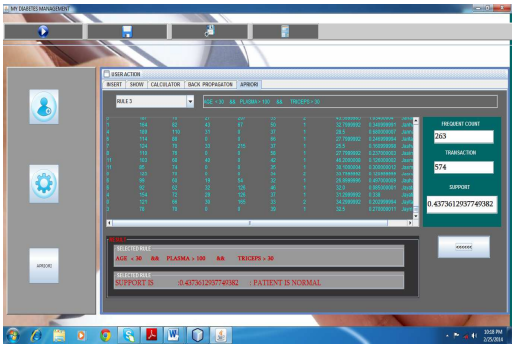

Fig 5.6 Output of Rule 3 Apriori Algorithm

According to Rule 3. 3 input parameters are taken Age<25, Plasma glucose level >100, Triceps skin fold thickness >30. According to Rule 3 Frequent Count is 263, No of transaction is 574, Minimum support is 0.437. The output of the Rule 3 shows the patient is Normal.

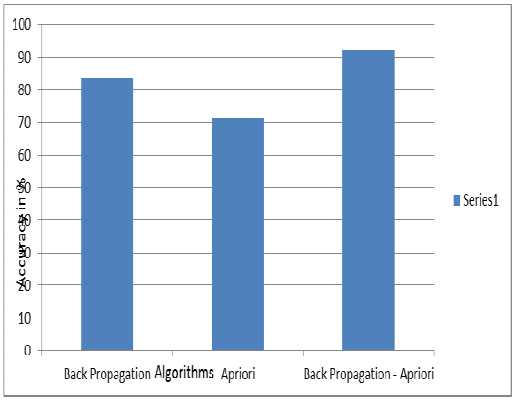

Fig 5.7 Comparision Graph

Comparision graph shows that Back propagation algorithm gives 83.5% accuracy Apriori gives 71.2% and combined Back propagation and Apriori gives highest percentage of 91.2%.

6. LIMITATIONS AND ASSUMPTIONS

Some of the limitations and assumptions of Back Propagation and Apriori algorithms are as follows:

Back Propagation algorithm is generally slow because it requies small learning rates for stable learning.

Back Propagation learning doesnot requires normalization could improve performance.

Apriori algorithm may produce large number of candidate item sets in this process.

Due to the above mentioned limitations both back propagation and apriori algorithms gives less accuracy compared to newly developed algorithm by fusing both back propagation and apriori algorithm.

7. CONCLUSION

In this paper, Back Propagation and Apriori algorithms, combination of both are used for the diagnosis of Diabetes Mellitus fully based on real time input. Back propagation algorithm gives 83.5% accuracy Apriori gives 71.2% and combined Back propagation and Apriori gives highest percentage of 91.2%. A newly developed algorithm gives the greater accuracy, more reliable and ease of use and maintenance. In future this can be implemented based on ONLINE.

REFERENCES:

[1] Blake, C.L., Mertz, C.J.: "UCI Machine LearningDatabases",http://mlearn.ics.uci.edu/databases/diabetes/, 2004

[2] International Diabetes Federation, Diabetes Atlas, 3 rd ed. Brussels, Belgium: *International Diabetes Federation*, 2007.

[3] WHO/IDF 2006 (2007, Jan.). Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia,WorldHealthOrganization[Online].Available:http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf. [4] International Diabetes Federation, Diabetes Atlas, 3 rd ed. Brussels, Belgium: *International Diabetes Federation*, 2007.

[4] Nahla H. Barakat, Andrew P. Bradley , Senior Member, IEEE , a nd Mohamed Nabil H. Barakat "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus" *IEEE Transactions On Information Technology In Biomedicine*, vol. 14, no. 4, July 2010.

[5] "The World's Children with Type 1 Diabetes" *Diabetes and society*.

[6] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model Construction on type 2 diabetic patient's data," *in Lecture Notes Artificial Intelligence*, vol. 3275, P.Perner, Ed. Berlin, Germany: Springer-Verlag, 2004,pp. 153–162.

[7] "Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia", World Health Organization, International Diabetes federation.

[8] WHO/IDF 2006. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia, World Health Organization [Online].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[9] Niti, Anil Dahiya,Navin Rajpal" Decision support system for Heart Disease diagnosis using Neural Network", *Delhi Business Review* X Vol. 8, No. 1 (January - June 2007).

[10] Frans Coenen, Graham Goulbourne, Paul Leng "Tree Structures for Mining Association Rules", Kluwer Academic Publishers, *Data Mining and Knowledge Discovery*, 8, 25–51, 2004.

[11] Sotiris Kotsiantis, Dimitris Kanellopoulos "Association Rules Mining: A Recent Overview", *GESTS International Transactions on Computer Science and Engineering*, Vol.32 (1), 2006, pp. 71-82.

[12] Professor Anita Wasilewska "Lecture notes on Apriori Algorithm.