Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
5(3), 2017, 113-122

ISSN 2347 – 7393

Review Article

DIABETES: A REVIEW OF DATA MINING TECHNIQUE FOR HEALTH CARE

Sami Ahmad[1], Prof. Kailash Patidar[2], Mr. Rishi Kushwah[3], Mr. Jitendra Rai[4]

PG Student, Dept. of CSE, SSSUTMS, Sehore, M.P., India[1]
Professor& Head, Dept. of CSE, SSSUTMS, Sehore, M.P., India[2]
Assistant Professor, Dept. of CSE, SSSUTMS, Sehore, M.P., India[3]
Assistant Professor, Dept. of CSE, SSSUTMS, Sehore, M.P., India[4]

Abstract: Data mining now-a-days assumes a critical part in expectation of sicknesses in medicinal services industry. Data mining is the procedure of choosing, investigating, and displaying a lot of information to find obscure examples or connections valuable to the data analyst. Medicinal data mining has risen faultless with potential for investigating concealed examples from the data collections of medicinal area. These examples can be used for quick and better clinical basic leadership for preventive and suggestive solution. However crude medicinal information is accessible generally circulated, heterogeneous in nature and voluminous for common preparing. Data mining and Statistics can on the whole work better towards finding concealed examples and structures in Data. A variety of data mining techniques help to envisage the disease such as SVM, PCA, GA etc. In this paper, we present the literature review of the earlier work done by the researchers to diagnose the diabetes also discuss the advantages and disadvantage of data mining techniques.

Keywords: Data Mining, SVM, KNN–GA, Diabetes

Introduction: Today the trendy expression is "Health Care" everywhere throughout the world. Early Prediction of maladies can decrease the deadly rate of human. There are substantial and gigantic information accessible in clinics and restorative related foundations. Data Technology assumes a key part in Health Care. Diabetes is a chronic disease with the possibility to bring about an overall Health Care emergency. As per International Diabetes Federation 382 million individuals are living with diabetes around the world. By 2035, this will be served as 592 million. Early forecast of diabetes is very testing undertaking for medicinal experts because of complex reliance on different elements. Diabetes influences human organs, for example, kidney, eye, heart, nerves, and foot and so on. Data mining is a procedure to extricate helpful data from substantial database. It is a

For Correspondence:
ahmad.sami89@gmail.com
Received on: June 2017
Accepted after revision: July 2017

multidisciplinary area of software engineering which comprises computational process, statistical techniques, classification, clustering, machine learning, and discovering designs.

Knowledge discovery in databases is very much characterized prepare comprising of a few unmistakable steps. In Fig: 1 demonstrates the design of Knowledge Discovery in Database. Data mining is the core step, which brings about the disclosure of concealed yet valuable learning from enormous databases. A formal meaning of Knowledge disclosure in databases is given as takes after: ―Data mining is the non-minor extraction of verifiable beforehand obscure and conceivably valuable data about information. Data mining innovation gives a client situated way to deal with novel and shrouded designs in the information. The initiated information can be utilized by the medicinal services chairmen to augment the nature of organization. The found learning can likewise be utilized by the medicinal specialists to decrease the quantity of antagonistic medication impact, to recommend more affordable remedially identical choices. Reckoning patient's future conduct on the given history is one of the critical utilizations of information mining methods that can be utilized as a part of medicinal services administration. Social insurance associations must have capacity to break down information. Treatment records of a large number of patients can be put away and modernized and information mining systems may help in noting a few vital and basic inquiries identified with health care[1].

Background: Diabetes is a whole life chronic condition, which may build the sugar level in the body. It might prompt different complexities. The nourishment we eat is changed over to glucose, which is utilized for vitality. The pancreas secretes insulin which produces vitality for impeccable working of the body. At the point when the patients have diabetes, body either doesn't make enough insulin or doesn't utilize insulin in legitimate way [2].

General symptoms of diabetes:

1. Increased thirst
2. Recurrent urination
3. Loss of body weight
4. Recurrent hunger
5. Slow healing infection
6. Blurred vision
7. Recurrent vomiting

Make a diagnosis test

1. Urine test
2. Fasting blood glucose level
3. Haphazard blood glucose level
4. Oral glucose tolerance test
5. Glycosylated hemoglobin.

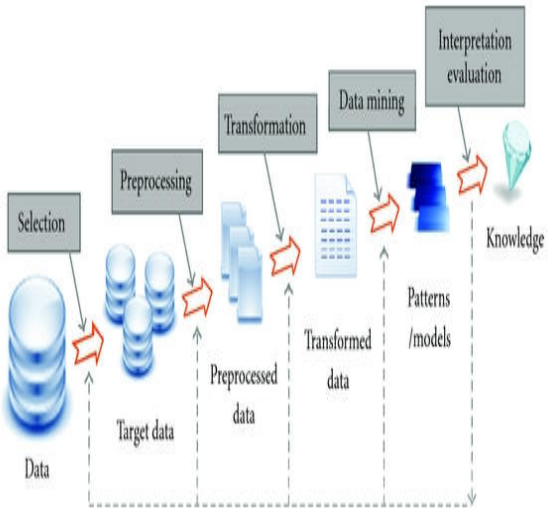

Fig.1 various steps involved in the process of data mining

Classification and Diagnostic Criteria for Diabetes: The Diabetes is classified into categories and follows the various criteria to diagnosis the diabetes. [2]

Classification Diabetes

Former classifications

The primary extensively accepted classification of diabetes mellitus was published by WHO in 1980 (1) and, in customized form, in 1985 (3). The 1980 and 1985 classifications of diabetes mellitus and allied categories of glucose fanaticism incorporated clinical classes and two statistical risk classes. The 1980 Expert Committee projected two foremost classes of diabetes mellitus and named them, IDDM or Type 1, and NIDDM or Type 2. In the 1985 Study Group Report the terms Type 1 and Type 2 were misplaced, but the classes IDDM and NIDDM were preserved, and a class of Malnutrition-related Diabetes Mellitus (MRDM) was initiated. In cooperation the 1980 and 1985

reports other classes of diabetes incorporated Other Types and Impaired Glucose Tolerance (IGT) in addition to Gestational Diabetes Mellitus (GDM). These were reflected in the successive International Nomenclature of Diseases (IND) in 1991, and the tenth revision of the International Classification of Diseases (ICD-10) in 1992. The 1985 categorization was broadly accepted and is used internationally. It represented a negotiated among clinical and anetiological classification and allowed categorization of individual subjects and patients in a clinically constructive manner even when the explicit cause or anetiology was unknown. The recommended categorizations comprise both staging of diabetes mellitus based on clinical descriptive criterion and a corresponding aetiological classification.

Revised classification: The characterization envelops both clinical stages and anetiological types of diabetes mellitus and different classes of hyperglycemia, as recommended by Kuzuya and Matsuda (15).The clinical organizing reflect that diabetes, paying little mind to it's etiology, advances through a few clinical stages amid its normal history. Also, singular subjects may move from stage to arrange in either heading. People who have, or who are creating, diabetes mellitus can be classified by organize as indicated by the clinical attributes, even without data concerning the basic etiology. The order by aetiological type comes about because of enhanced comprehension of the reasons for diabetes mellitus.

Application of the new classification: The new arrangement contains stages which mirror the different degrees of hyperglycemia in singular subjects with any of the malady forms which may prompt diabetes mellitus.

All subjects with diabetes mellitus can be ordered by clinical stage, and this is achievable in all conditions. The phase of glycaemia may change after some time contingent upon the degree of the basic malady forms (Figure 2). The illness procedure might be available however might not have advanced sufficiently far to cause hyperglycaemia. The aetiological grouping mirrors the way that the deformity or process which may prompt diabetes might be identifiable at any phase in the improvement of diabetes - even at the phase of normoglycaemia. Hence the nearness of islet cell antibodies in a normoglycaemic singular makes it likely that that individual has the Type 1 immune system prepare. Sadly there are couple of touchy or profoundly particular pointers of the Type 2 handle at show, despite the fact that these are probably going to be uncovered as etiology is all the more obviously characterized. A similar ailment procedure can cause weakened fasting glycaemia and additionally impeded glucose resistance without satisfying the criteria for the conclusion of diabetes mellitus. In a few people with diabetes, satisfactory glycaemic control can be accomplished with weight decrease, practice and additionally oral specialists. These people, hence, don't require insulin and may even return to IGT or normoglycaemia. Different people require insulin for satisfactory glycaemic control however can get by without it. These people, by definition, have some leftover insulin emission. People with broad beta-cell annihilation, and in this way no lingering insulin discharge, require insulin for survival. The seriousness of the metabolic variation from the norm can either relapse (e.g. with weight decrease), advance (e.g. with weight pick up), or remain the same.

Table 1: Attributes of Diabetes Dataset [3]

Attribute No.	**Attribute**	
1	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2	Pressure	Diastolic blood pressure(mmHg)
3	Skin	Triceps skin fold thickness(mm)
4	Insulin	2-Hour serum insulin (mu U/ml)
5	Pregnancy	Number of times pregnant
6	Mass	Body Mass Index(BMI)
7	Pedigree	Diabetes Pedigree function
8	Age	Age(in years)
9	Class	Class variable(0 or 1)

Related Work: ***Ojugo et al. [5]*** presented a hybrid fuzzy, genetic algorithm trained neural network model as a decision support system for diabetes classification. Adopted data is split

into: training, cross validation and testing to aid model validation with appropriate weights and biases set for each variables. Results indicate that age, obesity and family relations (in first and second degree), environmental conditions are critical factors to be watched; While in gestational diabetes, mothers with or without a previous case of GDM is confirmed if there is: (a) history of babies with weight > 4.5kg at birth, (b) resistant to insulin showing polycystic ovary syndrome, and (c) have abnormal tolerance to insulin.

Songthung and Sripanidkulchai [5] utilized the grouping to mine a broad dataset accumulated from 12 doctor's facilities in Thailand amid 2011-2012 with 22,094 records of screened populace who are females age 15 years or more seasoned. We utilize Rapid Miner Studio 7.0 with Naive Bayes and CHAID (Chi-squared Automatic Interaction Detector) Decision Tree classifiers to anticipate high hazard people and contrasted our outcomes with existing hand-processed diabetes chance scoring instruments. We characterize the objective of hazard forecast as scope which is the capacity to utilize screening information to distinguish people that are in the end determined to have diabetes. Their outcomes demonstrated that the utilization of arrangement presented in this paper rather than hand-processed scoring can enhance the expectation execution with an expansion in scope.

Durgadevi and Kalpana[6] utilized three component determination techniques to be specific, HS, MS and TS are concocted to acquire the profitable subset of significant elements for diminishing the dimensionality of numerous properties. This work proposed a changed subterranean insect digger calculation to remove the grouping rules from the information. Three seat stamped datasets (Cleveland Pima and Wisconsin) from the UCI machine learning storehouse were utilized to investigate adequacy of the proposed show. The acquired outcomes obviously demonstrates that the changed subterranean insect excavator beats the other top information mining grouping calculations like the CN2,RBF,Adaboost and Bagging regarding exactness. Along these lines the proposed display is equipped for creating great outcomes with less components and fills in as a reasonable device for evoking and speaking to the master's choice guidelines with a compelling backing for taking care of ailment expectation issue.

Chaudhari et al. [7] furthermore anticipate the ailment in view of hazard variables, for example, tobacco smoking, liquor consumption, age, family history, diabetes, hypertension, elevated cholesterol, physical inertia, heftiness. Analysts have been utilizing a few information mining procedures to enable wellbeing to mind experts in the analysis of coronary illness. K-Nearest-Neighbor (KNN) is one of the effective information mining methods utilized as a part of order issues. As of late, analysts are demonstrating that joining distinctive classifiers through voting is beating other single classifiers. This paper explored applying KNN to help medicinal services experts in the determination of infection exceptionally coronary heart disease. It additionally explores if incorporating voting with KNN can improve its precision in the finding of coronary illness patients. The outcomes demonstrate that applying KNN could accomplish higher exactness than neural system outfit in the analysis of coronary illness patients. The outcomes additionally demonstrate that applying voting couldn't upgrade the KNN precision in the finding of coronary illness..

Durairaj et al. [8] Neural Networks are one of the soft computing techniques that can be used to make predictions on medical data. Neural Networks are known as the Universal predictors. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The Artificial Neural Networks (ANNs) based system can effectively applied for high blood pressure risk prediction. This improved model separates the dataset into either one of the two groups. The earlier detection using soft computing techniques help the physicians to reduce the probability of getting severe of the disease. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases. In this paper, a detailed survey is conducted on the

application of different soft computing techniques for the prediction of diabetes. This survey is aimed to identify and propose an effective technique for earlier prediction of the disease.

Bagdi and Patil [9] introduced a choice emotionally supportive network which consolidates the qualities of both OLAP and information mining. The framework will anticipate the future state and produce helpful data for powerful basic leadership. With information mining, specialists can anticipate patients who may be determined to have diabetes. OLAP gives an engaged answer utilizing chronicled information of concerned patients. The framework likewise looks at the consequence of the ID3 and C4.5 choice tree calculations.

Patil et al. [10] acquainted another approach with produce affiliation runs on numeric information. They utilized pre-preparing to enhance the nature of information by dealing with the missing esteems and connected equivalent interim binning with inexact esteems in view of therapeutic master's recommendation to Pima Indian diabetes information. In conclusion from the earlier affiliation run calculation is connected to produce the guidelines. Just sort 2 diabetic patients the individuals who are pregnant lady underneath 21years are incorporated into their examination. It demonstrated that the outcomes got are exceptionally encouraging.

Nuwangi et al. [11] utilized progressed and solid information mining systems to distinguish diverse hazard calculates behind the diabetes and the connection between the diabetes and alternate infections. Utilizing affiliation control era, the connection amongst edema and diabetes and wheezes and diabetes has been recognized. The outcome appears, the females matured between 39 – 75 years with typical BMI run, systolic BP run and diastolic BP range and having wheezes will have a high risk towards growing high FBS (fasting blood sugar) level .l.

Kavitha and Sarojamma [12] presents a way to deal with outlining a stage to improve viability and productivity of wellbeing observing utilizing information digging for early discovery of any declining in a patient's condition. The utilization of information mining has additionally been set up in lessening rates of drug blunders Diabetes seriousness evaluation offering elements of information mining in view of the arrangement and relapse tree strategy. Truck is a powerful information mining and information investigation instrument that looks for imperative examples and connections and rapidly reveal concealed structure even in profoundly complex information. This investigation distinguished effortlessly appropriate demonstrative calculations utilizing early clinical test outcomes gotten from the lab tests. The arrangement of tenets given by this investigation is effortlessly justifiable by specialists.

Ganji et al. [13] utilized (FCS-ANTMINER) on open diabetes informational collection (Pima Indians Diabetes informational collection [18]). They acquired an exactness of 84%. Huang et al. [19] utilized three information mining calculations that were Naive Bayes, IB1 and C4.5 to foresee diabetes on information accumulated from Ulster Community and Hospitals Trust (UCHT) in the vicinity of 2000 and 2004. They could accomplish a precision of 98%.

Data Mining Techniques for Diagnosis of Diabetes: Diabetes Mellitus has become a common health dilemma nowadays, which would distress people and lead to different complications resembling visual mutilation, cardio vascular disease, leg amputation and renal malfunction if diagnosis is not done in the right time. In this converse the two classifier techniques with principal component analysis component analysis are implemented for the forecasting of Diabetes and accomplished with best forecasting techniques which has a maximum precision. These are given below:

Decision Trees: Decision tree [15] is a tree structure, which is in the form of a flowchart. It is used as a technique for classification and forecast with representation using nodes and internodes. The root and internal nodes are the test cases that are used to detach the instances with diverse features. Internal nodes themselves are the result of attribute test cases. Leaf nodes

denote the class variable. Figure 2 illustrates a sample decision tree structure.

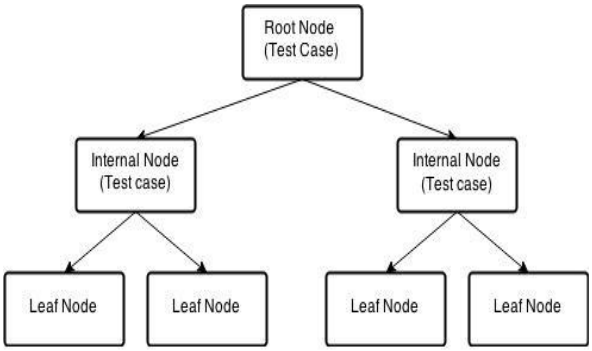

Fig.2. Sample Decision Tree Structure

Decision tree provides a powerful technique for classification and prediction in Diabetes diagnosis problem. Various decision tree algorithms are available to classify the data, includingID3, C4.5, C5, J48, CART and CHAID. In this paper, J48 decision tree algorithm [17] has been chosen to establish the model. Each node for the decision tree is found by calculating the highest information gain for all attributes and if a specific attribute gives an unambiguous end product(explicit classification of class attribute), the branch of this attribute is terminated and target value is assigned to it.

Advantages of Decision Tree

1. Decision tree a high speed
2. Low cost learning and prediction

Disadvantages of Decision Tree

1. Produce less accurate results
2. Less efficient than SVM

Naïve Bayes

The Naïve Bayes Algorithm is a probabilistic algorithm that is sequential in nature, following steps of execution, classification, estimation and prediction. For finding relations between the diseases, symptoms and medications, there are various data mining existing solution, but these algorithms have their own limitations; numerous iterations, binning of the continuous arguments, high computational time, etc. Naïve Bayes overcomes various limitations including omission of complex iterative estimations of the parameter and can be applied on a large dataset in real time. The algorithm works on the simple Naïve Bayes formula given below.

Advantages

1. It enhances the classification performance by eliminating the unrelated features.
2. Its performance is good. It takes less computational time.
3. Easy to implement

Disadvantages

1. This algorithm needs large amount of data to attain good outcomes.
2. It is lazy as they store entire the training examples
3. Practically, dependency exit among variables

Principal Component Analysis: Principal component analysis (PCA) is a standard tool in modern data analysis. It is a simple non parametric method for extracting relevant information from confusing data sets. Principal components analysis method is used for achieving the simplification and generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. The procedure can be followed in many ways i.e. a) Using singular value decomposition method (SVD) b) using the covariance matrix method. In this work we have used MATLAB software for deriving the principal components [18].

Advantages

1. PCA selects and transforms the original biomarkers to a reduced and/or transformed new series of uncorrelated variables.
2. PCA avoids the influence of regression edge in MLR and is easy to be operated.
3. This method generates the uncorrelated variables and provides the information of underlying structure of variables.

Disadvantages

1. The final step of the computation resembles the MLR method,
2. Some of the statistical deficiencies of MLR cannot be totally avoided [7].

Support Vector Machine:

SVM is an arrangement of related regulated learning technique utilized as a part of

therapeutic finding for order and relapse [19, 20]. SVM at the same time limit the observational characterization mistake and augment the geometric edge. So SVM is called Maximum Margin Classifiers. SVM is a general calculation in view of ensured hazard limits of factual learning hypothesis i.e. the alleged basic hazard minimization rule. SVMs can effectively perform non-straight order utilizing what is known as the portion trap, certainly mapping their contributions to high-dimensional component spaces. The bit trap permits developing the classifier without unequivocally knowing the component space. A SVM display is a portrayal of the cases as focuses in space, mapped so that the cases of the different classes are isolated by an unmistakable crevice that is as wide as conceivable [19, 21]. For instance given an arrangement of focuses having a place with both of the two classes, a SVM finds a hyperplane having the biggest conceivable portion of purposes of a similar class on a similar plane. This isolating hyperplane is known as the ideal isolating hyperplane (OSH) that expands the separation between the two parallel hyper planes and can limit the danger of misclassifying cases of the test dataset. Given named preparing information as information purposes of the frame: where=1/ - 1, a consistent that signifies the class to which that point has a place. n=number of information test. Each is a p-dimensional genuine vector. The SVM classifier initially maps the info vectors into a choice esteem, and after that plays out the arrangement utilizing a suitable edge esteem. To see the preparation information, we gap (or independent) the hyperplane, which can be depicted as:

$$\text{Mapping: } W^T.x + b = 0$$

Where w is a p-dimensional weight vector and b is a scalar. The vector w points perpendicular to the separating hyperplane. The offset parameter b allows increasing the margin. When the training data are linearly separable, we select these hyperplanes so that there are no points between them and then try on maximizing the distance between the hyperplane. We have found out the distance between the hyperplane as 2/|w|. To minimize |w|, we need to ensure that for all either

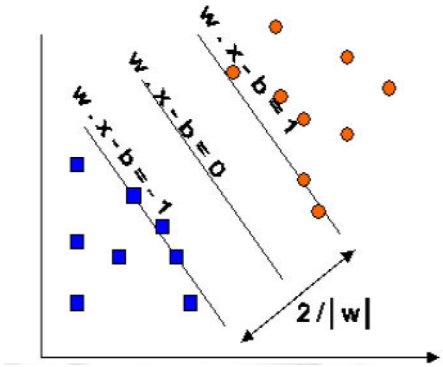

Fig. 3 Maximum margin hyperplane for SVM trained with sample from two classes

Advantages of SVM

1. It is much efficient to predict the disease
2. It give more accurate results than other techniques
3. Construct correct classifiers and fewer over fitting, robust tonoise

Disadvantage of SVM

1. It is less fast to produce the results
2. And more costly than the decision tree
3. It is a binary classifier. For the classification of multi-class, it can use pair wise classification

K-Nearest Neighbour (KNN)

The idea in k-Nearest Neighbor methods [22] is to identify ‘k‘samples in the training set whose independent variables 'm' are similar to 'n', and to use these 'k' samples to classify this new sample into a class, v. Assume that ‘f' is a smooth function, an idea is to seem for samples in our training data that are near it and next to calculate ‘v‘ from the values of ‘y‘ for these samples. When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can calculate among samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance is say, Euclidean distance. The initial training stage for kNN [23] consists of storing all known occurrences and their class labels. Either a tabular representation or a specialized design such as a kd-tree can be used. If we want to amend the value of 'k', an interchange method of n-fold cross-validation on

the training data set can be used. The k-NN algorithm for continuous-valued functions

- Determine the mean value of the k nearest neighbors Distance-weighted nearest neighbor algorithm
- Weight the involvement of k neighbors according to their distance to the query point x_q

– giving greater weight to closer neighbors

– Correspondingly, for real-valued target functions

Advantages

1. Good choice when there is no previous knowledge of data distribution

Disadvantages

1. Low efficiency
2. Dependency on the assortment of good values for k.
3. Requires meticulous tuning to optimally fit the real world data

Genetic Algorithm

Genetic algorithm [24] is a subset of evolutionary algorithm developed from Darwin's theory of gradual evolution and fundamental ideas. Process of optimization is based on a random trend in genetic algorithm. Before the genetic algorithm can be implemented, we must first find encoding system for the intended problem. The most common way to show chromosomes in the genetic algorithms is in binary form. In this case, chromosome is a bit string, the length of which is determined by some existing parameters. In other words, each parameter is related to a bit in a string. In this algorithm, for a fixed number called population, a set of target parameters is produced randomly. The genetic algorithm applies the rule of surviving the best to get the better solutions and then it assigns the number representing the fitting of that set to the member of the population. This process is repeated for every single member. With the retrieval of genetic algorithm operators such as selection, Mutation and crossover imitated from natural genetics, better approximations can be obtained from final solution and this procedure continue to get the convergence criterion. A selection operator chooses some chromosomes among the obtainable chromosomes in a population for reproduction. The techniques of selection are selection of the elite, the roulette wheel, tournament, Boltzmann, ranking, etc. The crossover operator is a random merging which some parts of chromosomes are exchanged. This issue causes that the children are not exactly like their parents and have had a combination of characteristics of their parents. After the merging, Mutation operator is applied on chromosome. This operator chooses a gene from a chromosome randomly and changes the content of that gene. There are three criterions for algorithm termination: 1-The number of generations in algorithm, 2-the population does not become better, 3-classification precision of element with the best fitness does not exceed the threshold level.

Neural Network

The most widely used neural-network learning method is the feed forward back Propagation algorithm. Learning in a Neural Network entails modifying the weights and biases of the network in order to reduce a cost function [10][14]. Our Neural Network based data mining approach consists of three major phases:

Network Designing and Training: The inputs to the Neural Network are initialized to a small random number. Each unit has a bias associated with it. The biases are also initialized to small number. In the figure fig1.,i1, i2, i3 … iN are the input nodes, t1 and t2 is the target node and a(n) is the activation function. A target node is one were the connection of input node ends.

The values of the input nodes are multiplied with the weight w, which is and summed with constant bias value of the node. This is the activation function of the node. This activation function is given as input to the target nodes.

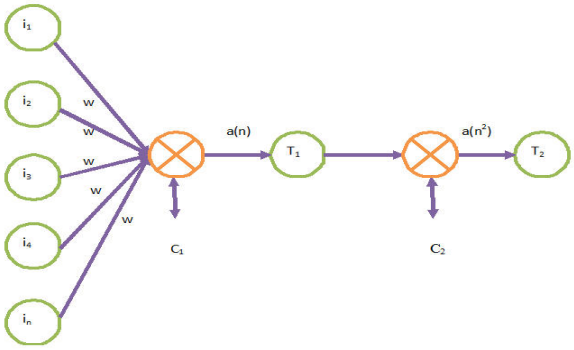

Fig 2. Neural network

The activation function was initially selected to be a relay function, but for mathematical convenience a hyperbolic tangent (tanh) or a sigmoid function are most frequently used. Hyperbolic tangent is defined as:

The input to the next layers will be the activation function multiplied by the preceding activation functions [10][14][9]

Advantages of NN

1. Adaptive Learning, Self-Organization, Real Time Operation Fault Tolerance via Redundant Information Coding.

Disadvantages

1. Less over fitting needs great computational effort. Sample Size must be large.
2. It's time consuming. Engineering decision does not extend the relations between input and output variables so that the model behaves like a black box.

Conclusion: Nowadays the trendy expression is "Health Care" everywhere throughout the world. Early Prediction of disease can lessen the deadly rate of human. Various data mining techniques play a significant role in predicting an efficient dataset from huge database. In this paper, we present the comprehensive review of literature of the various data mining techniques to cure the diabetes. Among these data mining techniques some are very efficient and accurate to unhidden the useful information but they are high costly and less speed. So in future work, design such algorithm which can efficiently determine the useful information with less cost and fast in execution.

References:

[1] Aqueel Ahmed, Shaikh Abdul Hannan, " Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September 2012

[2] S. Sapna, Dr. A. Tamilarasi and pravinkumar: "Implementation of Genetic Algorithm in predicting diabetes" International journal of computer science issues vol9, issues 1, no3, Jan 2013

[3] M. Renuka Devi, J. Maria Shyla "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus"International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730

[4] A.A. Ojugo., A.O. Eboka., R.E. Yoro., M.O. Yerokun and F.N. Efozia "Hybrid Model for Early Diabetes Diagnosis", Second International Conference on Mathematics and Computers in Sciences and in Industry, 2015. In proceeding of IEEE.

[5] Phattharat Songthung and Kunwadee Sripanidkulchai "Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification", 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). In proceeding of IEEE.

[6] M. Durgadevi, Dr. R. Kalpana "Medical Distress Prediction Based on Classification Rule Discovery Using Ant-Miner Algorithm" In proceeding of IEEE, 2017

[7] Anand A. Chaudhari, Prof. S. P. Akarte, " Fuzzy and Data Mining based Disease Prediction using K-NN Algorithm", International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014.

[8] M. Durairaj, G. Kalaiselvi, " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March 2015

[9] Rupa Bagdi, Prof. Pramod Patil "Diagnosis of Diabetes Using OLAP and Data Mining Integration", International Journal of Computer Science & Communication Networks, Vol. 2(3), 314-322,

[10] B. M. Patil, R. C. Joshi, DurgaToshniwal, "Association rule for classification of type -2 diabetic patients", Proc. of the Second International Conference on Machine

Learning and Computing, pp 330-334, 2010.

[11] S. M. Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna& H. A. Caldera, "Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes", Proc of the 6th International Conference on Advanced Information Management and Service(IMS), pp 372-377, 2010.

[12] Kavitha K ,Sarojamma R M, " Monitoring of Diabetes with Data Mining via CART Method", International Journal of Emerging Technology and Advanced Engineering, Website: www.ijetae.com ISSN 2250- 2459, Volume 2, Issue 11, November 2012.

[13] M. F. Ganji, M. S. Abadeh "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis", Expert Systems with Applications 38 (12) (2011) 14650 – 14659.

[14] Mark Hudson Beale, Martin T. Hagan and Howard B. Demuth, "Neural Network Toolbox™ User's Guide".

[15] Application of Artificial Neural Network in Detection of Probing Attacks Iftikhar Ahmad, Azween B Abdullah Department, Abdullah S Alghamdi, 2009 IEEE Symposium on Industrial Electronics and Applications (ISIEA 2009), October 4-6, 2009, Kuala Lumpur, Malaysia.

[16] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2001.

[17] Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining" Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[18] Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, " Diabetes Mellitus Forecast Using Different Data Mining Techniques", International conference on computer and Communication Technology .

[19] Cortes, C., Vapnik, V., "Support-vector networks", Machine Learning, 20(2),pp. 273-297, 1995.

[20] V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995.

[21] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery", Springer, 2(2), pp.121-167, 1998.

[22] Keller, J.M. ; Gray, M.R. ; Givens, J.A.,A fuzzy K-nearest neighbor algorithm, Systems, Man and Cybernetics, IEEE Transactions on (Volume:SMC-15 , Issue: 4). July-Aug. 1985, ISSN: 0018-9472 .

[23] G.Visalatchi et al, A Survey on Data Mining Methods and Techniques for Diabetes Mellitus, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February-2014, pg. 100-105 ISSN: 2321-8363.

[24] S. Bahramian and A. Nikravanshalmani "Hybrid algorithm based on K-nearest-neighbor algorithm and Adaboost with selection of feature by genetic algorithms for the diagnosis of diabetes", IJMEC, Vol. 6(21), Jul. 2016, PP. 2977-2986.