Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus

Veena Vijayan V.
Dept.of Computer Science
Mar Baselios College of Engineering & Technology
Trivandrum, India

Aswathy Ravikumar
Dept.of Computer Science
Mar Baselios College of Engineering & Technology
Trivandrum, India

ABSTRACT
Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various available traditional methods for diagnosing diabetes are based on physical and chemical tests. These methods can have errors due to different uncertainties. A number of Data mining algorithms were designed to overcome these uncertainties. Among these algorithms, amalgam KNN and ANFIS provides higher classification accuracy than the existing approaches. The main data mining algorithms discussed in this paper are EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm and ANFIS algorithm. EM algorithm is the expectation-maximization algorithm used for sampling, to determine and maximize the expectation in successive iteration cycles. KNN algorithm is used for classifying the objects and used to predict the labels based on some closest training examples in the feature space. K means algorithm follows partitioning methods based on some input parameters on the datasets of n objects. Amalgam combines both the features of KNN and K means with some additional processing. ANFIS is the Adaptive Neuro Fuzzy Inference System which combines the features of adaptive neural network and Fuzzy Inference System. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases**.**

Keywords
Data mining, Diabetes, EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm, ANFIS algorithm

1. INTRODUCTION
Data mining is one of the most attractive interdisciplinary subfield of computer science. It involves some computational process, statistical techniques, classification, clustering and discovering patterns in large data sets. The overall goal of the data mining technique is to extract the useful information from the large data set and to transform it into an understandable format so that it can be used for the future use. Diabetes Mellitus is a group of metabolic disease in which the amount of sugar content cannot be regulated. There are mainly four types of Diabetes Mellitus. They are Type1, Type2, Gestational diabetes, Congenital diabetes. Type 1 also called as "Insulin dependent Diabetes Mellitus" or "Juvenile Onset Diabetes Mellitus" occurs when the human body failures to produce insulin. They are characterized by the loss of insulin producing beta cells. Type 2 is also called as "Non Insulin dependent Diabetes Mellitus" or "Adult onset diabetes". Non Insulin dependent Diabetes Mellitus is characterized by the insulin resistance and is found in person above age 40 i.e., human body cannot effectively use the insulin that is produced. Diet, exercise and blood sugar level regulation should be maintained to regulate the Type 2 diabetes. Gestational diabetes mellitus (GDM) are temporary diabetes that resembles Type 2 diabetes in several aspects. It is the condition in which the pregnant women, without previously diagnosed diabetes exhibit an increase level of glucose in the blood. GDM is treatable under careful medical supervision and resolves completely once the baby is born. The Congenital diabetes is caused due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids. Diabetes mellitus causes serious complications such as heart disease, stroke, blindness, kidney failure and cancer. According to World Health Organization (WHO), 37 crores of people are suffering from diabetes around the world and it doubles before the year 2030[1]. Around 48 lakhs of people were died in the year 2012. Most of them belong to lower and middle class families [1].

Diabetes can be controlled by using different measures like insulin and diet. For this it should be identified as early as possible and subsequently provide appropriate treatment. Most of the classifying, identifying and diagnosing treatments are based on chemical and physical tests. Based on the inference obtained from these results, a particular disease can be predicted. Prediction may have errors. This is due to different uncertainty of various parameters used for testing [2]. Such uncertainties make the predictions wrong and prevent the chances of curing the disease. The computing facility has been progressed with great advancements. These advancements provided by information technology, helps to classify the data, predict the outcomes and diagnosis of many diseases more accurately. The main advantage of information technology is that a huge data storage of past patient's records are maintained and monitored by hospitals continuously for various references [2]. These medical data helps the doctors to examine different patterns in the data set. The patterns found in data sets may be used for classification, prediction and diagnosis of the diseases [2].

2. BACKGROUND
Diabetes is a lifelong chronic condition that affects the human body by reducing the insulin which carries glucose into the blood cells. This increases the sugar level in the body leading to different complications like stroke, heart disease, blindness, kidney failure and death. Diabetic patients generally have the following symptoms.

- Increased thirst
- Frequent urination
- Weight loss
- Increased hunger
- Slow-healing infections
- Blurred vision
- Nausea and Vomiting

The following medical tests are used to diagnose the diabetic mellitus [3]

- Urine test
- Fasting blood glucose level
- Random blood glucose level
- Oral glucose tolerance test
- Glycosylated hemoglobin(HbAlc)

2.1 DATASET

The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases. The patients under consideration are the Pima Indian population living in Arizona, USA. More than 50% Pima Indian Population is suffering from diabetes and 95% of them are due to the overweight. Number of research has been done on these populations proved that obesity is the main cause for the diabetes. The data set mainly contain 9 attributes and 768 number of instances [4]. The 8 such attributes along with symbols are listed in Table 1.

1) Total no of times pregnant

2) Glucose tolerance test to find the plasma glucose level concentration in saliva.

3) Diastolic blood pressure measured in mmHg for pressure level(BP)

4) Body Mass Index (BMI)

   BMI= Patients weight in kg/ (Patients height in meter)^2

5) Skin rashes and thickness fold in mm (Triceps)

6) 2- hour Serum Insulin in mu U/ml (INSULIN)

7) Age in years

8) Diabetes pedigree function

9) Diabetes Class Variable (1 indicates diabetic test is positive (presence) and 0 indicates test is negative (absence))

Table1. List of attributes from data sets for simulation tests

Attribute no	Attributes to be tested	Symbols
1	Diastolic blood pressure measured in mmHg	pres
2	Glucose tolerance test to find the plasma glucose level concentration in saliva.	plas
3	Body Mass Index (BMI)	mass
4	Diabetes pedigree function	pedi
5	Skin rashes and thickness fold in mm(Triceps)	skin
6	Age	years
7	2- hour Serum Insulin in mu U/ml(INSULIN)	insu
8	Diabetes Class Variable	binvar

3. LITERATURE REVIEW

Number of data mining algorithms has been proposed to classify, predict and diagnose diabetes. For this, data preprocessing should be done. It is a technique that involves transforming raw data into understandable format. This helps to fill the missing values in between the data. By analyzing the data using the values, it is possible for an expert to find values that are unexpected and erroneous.

3.1 The Expectation Maximization Algorithm [2]

This EM algorithm consists of two steps. The first step is determination of expectation and the second step is to maximization expectation in successive iteration cycles. The expectation involves choosing of a model and then it estimates missing labels. The maximization step involves choosing labels and then mapping of suitable models to labels so that it maximizes the expected log-likelihood of the data. The execution sequence may be listed in 3 steps [2].

Step 1: The expectation step that determines mean value, denoted by μ and infers the values of x and y such that x= [(0.5) / (0.5 + μ) * h] and y= [(μ/ 0.5 + μ) * h] with conditions of x / y = (0.5 / μ) and h = (x + y).

Step 2: The maximization step that determines fractions of x and y and then computes the maximum likelihood of μ at first.

Step 3: Repeats steps 1 and 2 for next cycle. The clusters were determined by cross validation of mean and standard deviation for 7 attributes. The class was then tested for positive and negative conditions of diabetes presence or absence respectively. For result analysis, binary response variables are represented by 1 which means that the diabetes test is positive (presence) and 0 which means that the test is negative (absence) for diabetes But the EM algorithm is not

very accurate for higher dimensional data sets due to numerical imprecision [2].

3.2 K-Nearest Neighbor Algorithm

KNN is a type of lazy learning used for classification. It is simplest algorithm in machine learning. This method can be used to predict labels of any type.

- An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors where k is a positive integer.
- It is analytically tractable and highly adaptive to local information. KNN algorithm uses the closest data points for estimation.
- Implementation can also be done in parallel. Because it is instance based, for each data point, the algorithm check against the training table for the k nearest neighbor. Since each data point is independent of the others, the execution of search and score can be conducted in parallel [4].

There will a number of samples for training. These samples are stored in an n –dimensional space. When an unknown test label is given, the k-nearest neighbor classifier searches these samples which are closest to the unknown sample. Closeness is usually defined in terms of Euclidean distance [4]. The Euclidean distance is between two points P (p1, p2… Pn) and Q (q1, q2… qn) given by equation (1).

$$d(P,Q)=\sum_{i=0}^{n}(P_i-Q_i)^2 \qquad (1)$$

KNN algorithm:-

Step 1: Each of the new instances is checked with the already available cases, based on distance assignment and classified using k value.

Step2:.The distance will be less, if the instances are more similar and vice versa.

Step 3: Observe the distance, k -value and instance. Based on these observations instances are assigned to a specific class.

Step4: The prediction is based on the k-value. So KNN classifier is k-dependent. Here k represents the number of nearest neighbors and for different values of k, outcome may not be the same [4].

Step 5: Determine the value of k for Pima Indian Diabetic Dataset (PIDD) for classification accuracy.

Drawbacks of KNN are:-

- It is necessary to compute distance of each instance to all other training samples. Hence the computation cost will be high.
- Large amount of data set are necessary for training. To incorporate this high data set memory space required will be large.
- Shows poor accuracy rate in multidimensional datasets.
- No thumb rule is available to determine the value of parameter k in which k represents number of nearest neighbors.

3.3 K-means algorithm

Unsupervised algorithms are those algorithms that operate on unlabelled samples. That means the output is unknown even if the input is known. K means algorithm is one among the unsupervised learning algorithm. They take input parameter, number of clusters and n object data set partition into k clusters. Algorithm select k objects randomly. Based on the closeness of each object with corresponding cluster, each object is assigned to one cluster. Next step is to find the points that are closest to each other. To assign the object to the closest center, Euclidean distance is preferred. Once the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters respectively. The process is repeated till there is no change in k cluster centers. K-means algorithm aims at minimizing an objective function, namely sum of squared error (SSE) [4]. SSE is defined as

$$\text{argmin}_C \quad \sum_{i=1}^{k}\sum_{p\in Ci}|p-m_i|^2 \qquad (2)$$

Here E represents the sum of the square error of objects with cluster means for k cluster and p is the object belong to a cluster C_i and m_i is the mean of cluster Ci. Here 'k' is number of clusters and n is the total number of records in dataset.

Input is k- is the number of clusters [4],

D is input -data set.

Output is k clusters.

Step 1: Initialize cluster centers as D.

Step 2: Randomly choose k objects from D.

Step 3: Repeat the following steps until no change in cluster means/ min error E is reached.

Step 4: Consider each of the k clusters. Compare the mean value of the objects in the clusters for initialization.

Step 5: Initialize the object with most similar value from D to one of k clusters.

Step 6: Take the mean value of the objects for each of k cluster.

Step 7: Update the cluster means with respect to object value.

3.4 Amalgam KNN [3]

This comprises both the feature of K- means algorithm and KNN algorithm. A combination of these two will improve the accuracy even for large number of data set. The great observation of KNN is that, it can be combined with any other algorithms for better accuracy. The k value will play a significant role for this. If the k value is very small, it provides less accuracy and if the value is large then the accuracy gets improved. The computational complexity and overhead will be high if we increased k beyond a certain limit. Preprocessing techniques are involved for transforming raw data into understandable format and to avoid noisy data. The required data set is obtained from Pima diabetes dataset and the following steps are followed.

Step1: The inconsistent values are removed through preprocessing.

Step2: Identify and eradicate erroneously classified instances using K means clustering. This helps to decrease the computational cost of KNN.

Step3: The missing values are reinstated by mean and median values.

Step 4: Take the precise clustered instance for KNN with preprocessed subsets as inputs.

Step5: Finally modify and adjust the classification using KNN

Step6: Then the model is tested for distinct values of k.

3.5 Adaptive Neuro Fuzzy Inference system (ANFIS)

ANFIS incorporates the features of fuzzy systems and neural networks. Efficiency of ANFIS can be improved by using adaptive based KNN. The two rules based on ANFIS are [1]

Rule 1:

If (a is A_1) and (b is B_1)

then

$f_1=p_1a+q_1b+r_1$

Rule 2:

If (a is A_2) and (b is B_2)

then

$f_2=p_2a+q_2b+r_2$

In fuzzy region

a and b -- Inputs

A_i and B_i -- Fuzzy sets

f_i --Outputs

p_i, q_i and r_i -- Design parameters

Adaptive group is verified by using the number of training group during i^{th} data processed, categorization result of i^{th} document by j^{th} group and the average value of different categories calculated by feature distance in groups [1].

4. PERFOMANCE COMPARISONS OF ALGORITHMS

Classification accuracy (ACU) is the most common method used for evaluation of performance. Calculation of accuracy is performed by taking ratio of truly classified samples (true negative, true positive) to the total number of samples.

Accuracy = Truly classified samples/total samples (3)

Another evaluation methods used for measuring performance are Sensitivity and Specificity. Sensitivity is calculated by dividing the true positive (TP) samples to the sum of true positive (TP) and false negative (FN) samples.

$$\text{Sensitivity} = TP/(TP + FN) \quad (4)$$

Specificity is calculated by dividing the true negative (TN) samples to the sum of true negative and false positive (FP) samples.

$$\text{Specificity} = TN/(TN + FP) \quad (5)$$

KNN have the sensitivity value 75% and specificity value 58%. Amalgam-KNN has the highest sensitivity and specificity. The Table 2 shows the performance accuracy of different algorithms. EM algorithm is having the least accuracy rate and it exhibit inaccuracies when high dimensional data sets are given as input. By analyzing the table it is clear that amalgam KNN and ANFIS algorithm with adaptive KNN shows the maximum accuracy compared to other algorithms. To achieve the better accuracy, results were analyzed for different k values.

Table 2. Performance analysis

Method	Accuracy
EM algorithm	<70%
KNN algorithm	73.17%
K means algorithm	66-77%
Amalgam KNN algorithm	>80%
ANFIS algorithm with adaptive KNN	80%

5. RECOMMENDATIONS

ANFIS and Adaptive based KNN algorithms perform the classification with a higher efficiency and reduced complexity. Amalgam KNN extracts both the feature of KNN and K means algorithms. Since the current accuracy of both ANFIS with adaptive based KNN and amalgam KNN is greater than 80 %, they can be combined to produce a better accuracy algorithm than the exiting one. Finally Co- active ANFIS which combines both the features of adaptive neural networks and fuzzy systems which is termed as the successor of ANFIS can also be used to improve the performance of the present algorithm. Their conjoint dependence provides astonishing abilities for learning. CANFIS provides non liner rules for classification, prediction and diagnosis among the input output pairs. Results should be compared and tested with the increased values of k. Greater the k value, more will be accuracy rate.

6. CONCLUSION

Data mining and machine learning algorithms in the medical field extracts different hidden patterns from the medical data. They can be used for the analysis of important clinical parameters, prediction of various diseases, forecasting tasks in medicine, extraction of medical knowledge, therapy planning support and patient management. A number of algorithms were proposed for the prediction and diagnosis of diabetes. These algorithms provide more accuracy than the available traditional systems. This paper includes algorithms like Expectation Maximization Algorithm, K Nearest Neighbor algorithm, K-means algorithm, Amalgam KNN algorithm and Adaptive Neuro Fuzzy Inference System algorithm. From the observation EM possess the least classification accuracy and amalgam KNN and ANFIS provide the better classification accuracy results. Amalgam KNN comprises both the feature of KNN and K means. ANFIS in cooperates both the features of adaptive neural network and Fuzzy inference system. The performance and accuracy of the algorithm can be improved if a combination of both ANFIS and amalgam KNN is used. Co active ANFIS was extended with some capabilities of its predecessor ANFIS

to provide better classification and prediction accuracy. Classification shows better accuracy when the k value is increased to a large value.

7. REFERENCES

[1] C.kalaiselvi,G.m.Nasira,2014.”A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS”,IEEE Computing and Communicating Technologies,pp 188-190

[2] Velu C.M, K.R.Kashwan,2013.”Visual Data Mining Techniques forClassification of Diabetic Patients”, IEEE International Advance Computing Conference (IACC),pp-1070-1075.

[3] Sapna. S,Tamilarasi. A and Pravin Kumar.M, 2012 “Implementation of genetic algorithm in predicting diabetes”, IJCSI, International Journal of Computer Science Issues, Vol. 9, Issue 2, No 4, pp. 393-398

[4] Nirmala Devi M.,Appavu alias Balamurugan S.,Swathi U.V., 2013.”,An amalgam KNN to predict Diabetes Mellitus”, IEEE International Conference on Emerging Trends in Computing ,Communication and Nanotechnology(ICECCN),pp 691-695

[5] Asha Gowda Karegowda and Jayaram. A. M., 2009“Cascading GA & CFS for feature subset selection in medical data mining”, IEEE International Advance Computing Conference, Patiyala, India

[6] Krzysztof J.Cios, G.William Moore (2002) ‘Uniqueness of Medical Data Mining’, Artificial Intelligence in Medicine Journal pp 1-19.

[7] Asha Gowda Karegowda , A.S. Manjunath , M.A. Jayaram (2011) “Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes”, International Journal on Soft Computing (IJSC), Vol.2, No.2

[8] Krzysztof J.Cios, G.William Moore (2002) ‘Uniqueness of Medical Data Mining’ Artificial Intelligence in Medicine Journal pp 1-19

[9] Siti Farhanah Bt Jaafar and Dannawaty Mohd Ali,“Diabetes mellitus forecast using artificial neural networks”, Asian conference of paramedical research proceedings, 5-7,September, 2005, Kuala Lumpur, Malaysia.

[10] T.Jayalakshmi and Dr.A.Santhakumaran, “A novel classificationmethod for classification of diabetes mellitus using artificial neural networks”. 2010 International Conference on Data Storage and Data engineering

[11] Edgar Teufel1, Marco Kletting1, Werner G.Teich2,Hans-Jorg Pfleiderer1, and Cristina Tarin-Sauer3,sept.2003“Modelling the Glucose Metabolism with Backpropagation Through Time Trained Elman Nets”, IEEE 13th Workshop on Neural Networks for SignalProcessing, NNSP'03, pp.789 – 798

[12] Fuluf helo V Nelwamondo, Shakir Mohammed andTshilidzi Mawala, 2007 “Missing Data: A comparison ofneural network and expectation maximization techniques”, Current Science, Vol 93, No 11

[13] J. Prather,1997 et al., "Medical data mining: knowledgediscovery in a clinical data warehouse.," Proc AMIAAnnu Fall Symp, pp. 101–105.

[14] R. Bellazzi and B. Zupan, 2008. "Predictive data mining inclinical medicine: Current issues and guidelines,"International Journal of Medical Informatics, vol.77, pp. 81-97

[15] UCI machine learning repository and archive.ics.uci.edu/ml/datasets.html