Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus

Dr. K. Thangadurai,
Assistant Professor and Head,
PG and Research Department,
Department of Computer Science,
Govt., Arts College (Autonomous)
Karur 636 005

N.Nandhini
Research Scholar,
Department of Computer Science,
Periyar University,
Salem 11

Abstract

The aim of data mining is to extract hidden knowledge from huge amount of data set and generate clear and easy understandable patterns. Diabetes is a group of metabolic disease caused by increased level of blood glucose. Different data mining algorithms are applied in medical research in order to diagnosis large amount of medical dataset. Various data mining algorithms were designed for diagnosing diabetes based on physical and chemical tests. The main data mining algorithms discussed in this paper are EM algorithm, K means, C4.5 algorithm, Genetic algorithm and SVM. EM is the expectation maximization used for sampling, to determine and maximize the expectation in successive iterative cycles. C4.5 is a decision tree induction technique that has been successfully applied for medical data. Genetic algorithm is population based model that uses selection and recombination operators to generate new sample points. Support Vector Machine are set of supervised learning method whose training tech permit to represent complex non linear function. K means is a unsupervised which objects are moved among sets of cluster until the desired set is reached. This paper studies the comparison of various data mining algorithms for prediction of diabetes disease.

Keywords: Data mining, diabetes, EM algorithm, K means, C4.5, SVM, Genetic algorithm

1. Introduction

Diabetes is a group of metabolic disease due to high blood sugar levels. It may produce the symptoms of frequent urination, increased thirst and hunger. It may also cause many complications such as heart disease, stroke, kidney failure, foot ulcers and damages to the eyes. There are three types of diabetes.

Type1: The body failure to produce enough insulin. It is also referred as “Insulin dependent diabetes”.

Type2: Cells may fail to respond to insulin properly. It is also referred as “Non insulin dependent diabetes”

Gestational Diabetes: It may occur when pregnant women without a previous history of diabetes develop a high blood glucose level [1].

An estimated 40 million Indian suffers from diabetes and the problems seems to be growing at an alarming rate. By 60 million is expected to double and reach epidemic proportion [2][3].

Data mining is one of the most important techniques to extract hidden information in the medical data set. It is used to improve the quality of health care of diabetes patients. It may involve some computational process, statistic technique, classification, cluster and discovering pattern in large data set. In this paper we discuss different algorithm for the prediction and diagnosis of diabetes. The algorithms like EM, K means, C4.5, SVM and Genetic algorithm [4].

2. Background

Diabetes is a life long chronic condition, which may increase the sugar level in the body. It may lead to various complications. The food we eat is converted to glucose, which is used for energy. The pancreas secretes insulin which produces energy for perfect functioning of the body. When the patients have diabetes, body either doesn’t make enough insulin or doesn’t use insulin in proper way [3].

General symptoms of diabetes:

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

Diagnose test

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

3. Dataset

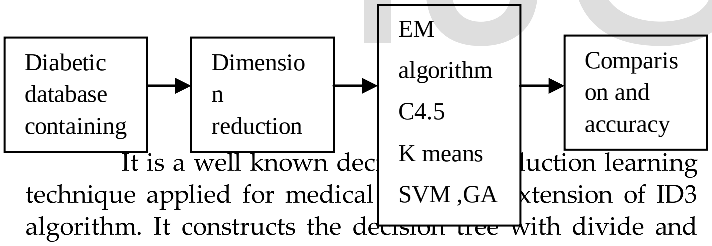
The dataset chosen for classification and experimentation simulation is based on Pima Indian diabetic from university of California. Irvine(UCI) Repository of Machine learning databases. More than 70% Pima Indian population is suffering from diabetes. The dataset mainly contain 9 attributes of 768 number of instance.

S.no	Name of the Attributes	Description
1	Preg	Number of time pregnant
2	Plas	Plasma glucose concentration a 2 hours in a oral glucose tolerance test
3	Pres	Diastolic blood pressure
4	Skin	Triceps skin fold thickness
5	Insu	2 hours serum insulin
6	Mass	Body mass index
7	Pedi	Diabetes pedigree function
8	Age	Age
9	Class	Class variable (0 or 1)

The dataset is based on the numeric and nominal data type [5].

4. Literature review

Different data mining algorithm has been proposed for classify, predict and diagnose diabetes. Data preprocessing is most important task to transforming raw data into understandable format. It may helps to solve the missing values in between the data.



It is a well known dec luction learning technique applied for medical xtension of ID3 algorithm. It constructs the decision tree with divide and conquers technique [6].

Why C4.5 is better than ID3?

It represents supervised learning model with a known output used for comparison of the model output.

1. Handle both continuous and discrete attributes.
2. Handle training data with missing attributes.
3. Handle attributes with different costs
4. Pruning trees after creation

Remove branches that do not help by replacing them with leaf nodes.

Pseudo code:

```
1. Check for base cases
2. For each attribute a, find the normalized information from splitting on a.
3. Let a best be the attribute with the highest normalized information gain.
4. Create a decision mode that splits on a best.
5. Recurs on the sub lists obtained by splitting on a best and add those nodes as children of node [6].
```

At the beginning only the root is present and associated with the whole training set and with the entire case weight equal to 1.0. At each node the divide and conquer algorithm is executed, typing to exploit the locally best choice, with no back tracking allowed.

K Means

It is a unsupervised learning and iterative clustering algorithm in which objects are moved among set of clusters until the desired set of reached. Within cluster a centroid denotes a cluster that is mean point [7].

The main goal of K means is to subset n observation into K cluster in which each observation belonging to the cluster with the nearest mean. It support numerical attributes [4].

Pseudo code

```
1. Initialize cluster centers as D
2. Randomly choose K object from D.
3. Repeat the following steps until no change is cluster means.
4. Consider each of the K cluster. Compare the mean value of the object in the cluster for initialization.
5. Initialize the object with most similar value from D to one of K cluster.
6. Take the mean value of the object for each of K clusters.
7. Update the cluster means with respect to object value.
```

The aim is to minimizing an objective function.

EM Algorithm (Expectation Maximization) [8]

It may consist 2 steps: 1. Determination of expectation. 2. It 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 choose labels and then mapping of suitable models to labels so that it maximizes the expected log likelihood of the data.

Pseudo code

```
1. The expectation step that determines mean value, denoted by μ and infers the value of x and y.
       $x/y=(0.5/ \mu)$
       $h=(x+y)$
       $x=0.5/(0.5+\mu)*h$
       $y=( \mu/0.5+x)*h$
2. The maximization step that determines fraction of x and y and then computes the maximum likelihood of μ.
3. Repeat 1 and 2 for next cycle.
```

EM is not very accurate for higher dimensional data set due to numerical impression.

SVM Support vector Machine

It is simple linear form. It is a representation of the examples as points in space mapped. So that the example of the separate categories are divided by the clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on in addition performing linear classification.

Syntax: SVMstruct=SVMtrain(data, groups, kernel function, rbf)

Data: Matrix of data points, where each row is one observation and each column is one feature.

Group: column vector with each row corresponding to the value of the corresponding row in data. Groups should have only 2 types of entries. So groups can have logical entries or can be a double vector or cell array with 2 values.

Kernel function: the default value of linear separated the data by a hyperplane. The value rbf uses a Gaussian radial basis function. The resulting struct contains the optimized parameters from the SVM algorithm enable to classify new data.

Pseudo code

```
Require x and y loaded with training labeled data. $\alpha$<=0 or <= partially trained.
Svm
        C <= some value
        Repeat
                for all {$x_i$,$y_i$} do
                        optimize $\alpha_i$ and $\alpha_j$
                end for
        Until no changes in $\alpha$ or other resource constraint criteria met.
Retain only the support vectors ($\alpha_i$ >0)
```

Genetic algorithm

Genetic are subclass of evolutionary algorithm where the elements of the search space are binary strings or arrays of other elementary types. It is solution for optimization of hard problem quickly, reliably and accuracy.

Optimization is the process of modifying the input to obtain minimum or maximum of the output. This heuristic method is routinely used to generate useful solution to optimization and search problem.

Pseudo code

1. Start: generate random population of chromosomes
2. Fitness: evaluate the fitness of each chromosome
3. New population: create a new population by repeating following steps until the new population is complete
    a. Selection: select a 2 parent chromosomes from a population according to their fitness the best fitness to be selected to be the parent
    b. Crossover: crossover the parent to form new offspring. If no crossover was performed, offspring is the exact copy of parent
    c. Mutation: mutate new offspring at each levels
    d. Accepting: place new offspring in the new population.
4. Replace: use new population for a further run of the algorithm.
5. Test: if end task met stop and returns the best solution in current population.
6. Loop go to step2

The commonly used techniques for selection of chromosome are roulette wheel, rank selection and steady state selection.

5. Performance comparison of data mining algorithm

Classification accuracy is the most common method used for evaluation of performance

Accuracy = Truly classified samples/total samples

The final output will be pattern which are used to find out whether the person is affected with diabetic or not and pre diabetic.

Accuracy: (TP+TN) / (TP+TN+FP+FN)

A Confusion matrix is a useful tool for analysis classifier accuracy structure of confusion matrix.

AC\PC	C1	C2
C1	TP	FN
C2	FP	TN

Another evaluation methods used for measuring performance are sensitivity and specificity.

Sensitivity=TP/(TP+FN)

Specificity=TN/(TN+FP)

6. Performance analysis

Methods	Accuracy
EM Algorithm	<70%
C4.5	71.2%
K Means	77%
SVM	68%
Genetic algorithm	78.1%

Conclusion

Data mining technique plays a major role in extracting the hidden knowledge in the medical database. The data preprocessing is used to improve the quality of the data. Various data mining algorithm are applied on Pima datasets. It is found that the genetic algorithm gives a better performance over five data mining algorithm. It believed that the data mining can significantly help in the diabetes mellitus research and improve the quality of health care in diabetics mellitus patient.

Reference:

1. http://en.wikipedia.org/wiki/diabetes_mellitus
2. http://www.medicalnewstoday.com/info/diabetes
3. S.Sapna, Dr. A.Tamilarasi and pravin kumar: “Implementation of Genetic Algorithm in predicting diabetes” International journal of computer science issues vol9, issues 1, no3, jan 2013
4. Veena Vijayan.V, Aswathy Ravikumar “Study of datamining algorithm for prediction and diagnosis of diabetes mellitus”

International journal of computer application, Vol 95, no7, June 2014.

5. UCI Machine Learning repository and archieve, ics.uci.edu/me/datasets.html
6. Rupa Bagdi "Diagnosis of diabetes using OLAP and data mining Integration" International Journal of computer science & Communication Networks Vol2, June 2014
7. P.Tangaraju, B.Deepa, T.Karthikeyan "Comparison of Data mining technique for forecasting Diabetes", International Journal of Advanced Research in computing and communication engineering vol3, issue 8, August 2014.
8. Velu C.M, K.R.Kashwan "Visual data mining technique for classification of diabetic Patients" IEEE International Advanced computing conference June 2013
9. Waheeda dhokley, Tahreem Ansari "New Improved Genetic algorithm for coronary Heart disease prediction", International Journal of computer application vol136, no5, Feb 2016.
10. Rahul malhotra, Narinder singh "Genetic algorithm: concept, design for optimization of process controllers" Computer and information science, vol4 No2, March 2011.