Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 78 (2015) 29677-29680

ISSN: 2229-712X

Diagnosing diabetes using data mining algorithms and artificial intelligence systems

Amir Amiri[1*] and Vahid Rafe[2]

[1]Department of Computer Engineering, Malayer branch, Islamic Azad University, Malayer, Iran.

[2]Department of Computer Engineering, Faculty of Engineering, Arak University, Arak 38156-8-8349, Iran.

ARTICLE INFO

Article history:

Received: 29 November 2014;

Received in revised form:

21 December 2014;

Accepted: 6 January 2015;

Keywords

Data mining,

Diagnostic, Diabetes, C5.0,

Decision trees,

Artificial neural networks,

Feature selection.

ABSTRACT

The purpose of this study is to investigate the role and the scope of the application of data mining predictive science and medicine in a bid to build, evaluate, and the exploitation of data mining models in this regard. In this study to examine the related works in the field of prediction data mining in medicine, who recently published and is trying to highlight important issues and summarize methods and algorithms applied in the form of a series of training. According to the study of used in this study, in most cases to explore knowledge in the medical data from a combination of such as smart algorithms artificial neural network and the decision in the direction of the optimal - the former methods have been used.

Introduction

Today, in medical data are collected in the case of the importance of different diseases. Medical centers with intentions to collect data. A study of the data and to gain the results and useful patterns in connection with the disease is one of the goals of the use of the data. The high volume data and confusion resulting from the problems that impede reaching remarkable results. So of data mining to overcome the problem and to obtain good relations between risk factors, according to the outbreak of the disease and the contribution in mortality humans have be [1].

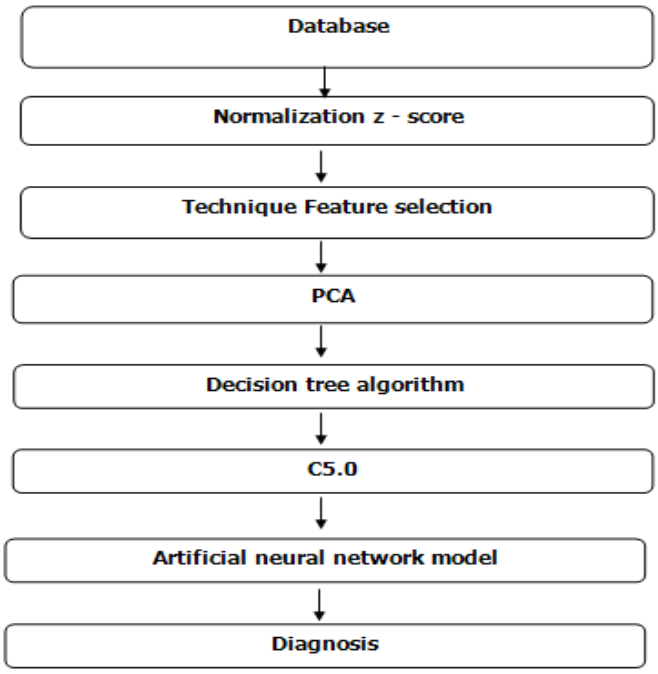

Figure 1: The process of implementing the proposed system

Data mining and knowledge useful patterns relations hidden in a large volume is given. Such studies and searches can be located along the same duration of the Old Creatures and comprehensive statistics. The major difference in the scale of the extent and variety of fields and applications and data size is contemporary approaches to modeling and machine learning the will seek [2].In this thesis is trying to provide a solution for the new data mining diagnosis diabetes, according to a similar cases in this case should offer a solution has been forecast accuracy than investigations and secondly could factors affecting the disease diagnosis.

The proposed algorithm

The proposed algorithm consists of three stages in the following form is also shown. In the first stage of the database, which is about to be fully in the previous section.

Initially the data from the mining database. Then the data explore, by the process makingof decision - tree (D - T). The best the research and the way of the input characteristics of patients (to - output), the relative avoidance rate ill or healthy (by a symbolic depiction of a tree. Artificial Neural Network (neural networks) to estimate, education, adaptability, machine learning, in order to identify and detect the disease has been used.

Knowledge discovery stages in the database 1959 machine learning year term for the first time by Samuel.

1 - data cleansing: " completely different data

2 - integrating data: a combination of multiple sources, dispersed, and heterogeneous data

3- data choice: retrieve data related to discover the knowledge

4 - data: To apply in different ways

5 - data mining: necessary step in the discovery of knowledge and wisdom of data base in which the statistical methods and machine learning proper model is used to extract.

Data mining:

Data mining is adoption or exploitation of knowledge of a very large amounts of data, in other words, data mining process.

Tele:

E-mail addresses: Amiryamir57@gmail.com

With the use of Smart techniques, knowledge of a set of data mining statistical analysis is simply not be able to do it.Data mining very complicated mathematical algorithms to divide data and prediction events is used [3].With the advancement of science and technology and technology tools, the ability to review and storing important data provided with large - scale. Scientific need to search in the data and receive necessary beneficial results.

Data mining, automatic search large data sources, to find patterns and affiliation statistical analysis simply could not do [4].Data mining include information and analysis tools to explore reliable patterns and unknown among a lot of data.Data mining algorithms in various professional methods are used. Data mining, exploitation of knowledge of the data, and learning deductive are called.

Data mining techniques can be in the range of types of data, including style Text, databases, the location information, when information, and other sophisticated data used [5].

Data mining is credible information extraction process, unknown, understandable and reliable of large databases and its use in decision making on major commercial activities. [6]

Data Mining is a process that data mining techniques, intelligent, knowledge of a set of data. [7]

Data Mining i. e., a search in a database to find patterns between the data. [7]

In fact, the discovery of data mining structures, interesting and valuable through a vast collection of data.Data MiningIs an activity that basically with detailed analysis of the data.[6]

- Data mining, extract hidden patterns of relations between the data in a large amount of data and summarize data with innovative ways. [5]
- data mining Recognition is reliable patterns, new, essentially good sense of data.
- Data mining methods in a series of knowledge discovery process can be used to detect patterns and vague relations in the data used.

1 - Data mining operations (categories, clustering, forecast, determining dependence, etc.)

2 - data mining methods (neural networks, the decision - making, the genetic algorithm, etc.)

3-Data - to find a suitable model.

4 - Assessing schemes: best suit pattern.

5-To provide knowledge - knowledge derived using data techniques.

Data mining techniques:

Various data mining techniques can be based on the types of operations are carried out in two predictors and descriptive divide. Predictive techniques to build a model to the database, predicted task unknown cases. While describing techniques, understandable patterns of data for humans. Descriptive techniques include:

- Summarize
- clustering
- community laws
- consecutive patterns

Divination techniques include:

- classification
- then extremism
- time series analysis

Database:

In the relevant variables in this study as a database of PID medical system has been used as a round - of - the - 8 as the property. The variables used in this research - which includes the following characteristics: [8] of the 867, 500 and 267 people healthy people infected with the disease, diabetes. 8 registered feature of individuals based on the definition of the World Health Organization (WHO)May - in this case.

The way it works is the first of a series of included in any field is that includes Beat - to the number of features. In this paper each include 8 Bit could be - that every bit represents one of the features. Being a zero Bit shows a lack of property and it is a sign of the property in [9]. The features were randomly selected from among all existing properties. The suffering of the variables and changes in the table below.

Table 1 - Normal suffering symptoms of laboratory

Attribute number	Mean	Standard Deviation	Unit
times_pregnant	3.8	3.4	Number of times pregnant
glucose_tol	120.9	32.0	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
diastolic_pb	69.1	19.4	Diastolic blood pressure (mm Hg)
triceps	20.5	16.0	Triceps skin fold thickness (mm)
insulin	79.8	115.2	2-Hour serum insulin (mu U/ml)
mass_index	32.0	7.9	Body mass index (weight in kg/(height in m)^2)
pedigree	0.5	0.3	Diabetes pedigree function
age	33.2	11.8	Age (years)

Normalization data

Normalization data scale change is that they are to a narrow range and defined as the distance between the 1 to - 1 map. Normalization of various methods that are described below. Normalization causes large - scale data to divert his side. Several ways to normalization exists in this study of normalization z - score is used. In this way, using data from the mean and standard deviation, are normal feature in the relationship is shown.

$V' = (v - A_v)/\sigma_v$ or$(x–mean(x))/std(x)$ (1)

A_v mean and standard deviation σ_v here.Feature technique selection technical feature selection is to reduce the number of features before applying data mining algorithm is used. Characteristics of irrelevant or sub - may have a negative impact on the task to be expected, or complicated calculations. Feature selection consists of three stages:

- Selection: and cases of non - important records. To do this, there are several techniques and methods that one of these techniques using genetic algorithms to do that.
- Ratings remaining records on the basis of their significance and scored rating.

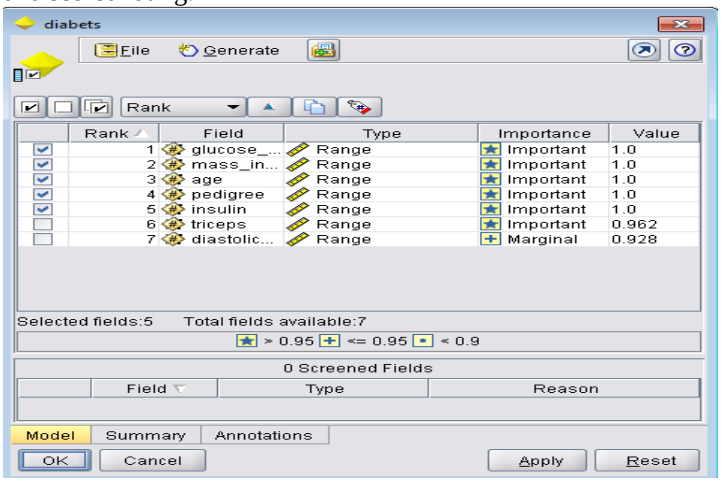

Figure 2. Model Feature Selection

- Selection: a record subsets, to maintain the inputs are more important than all the cases, the rest of the filter. To concentrate on the fields and get recordings that more important processing, faster and easier and less number of computation is reduced significantly.

Model PCA:

PCA to apply the technique is also necessary to reduce the dimensions of the features. This technique must be carefully and that of First of all, the accuracy of the data is not so much diminished when data without this technique are used. When this technique are examined acceptable results and the reduction of the dimensions of the attribute to a good result seems to be necessary. Besides, with regard to reduce by reducing the time and the complexity of the algorithm. From this perspective is important, too. With regard to the aforementioned points only for 5 after new data was considered to be cases mentioned in the figure below.

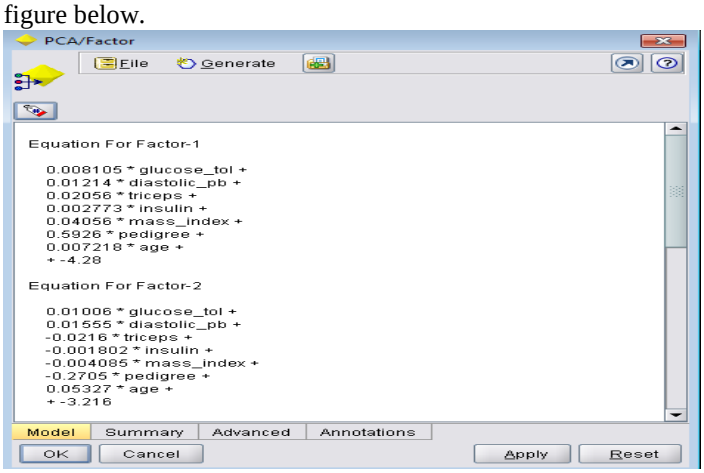

Figure 3 - impose PCA technique on data and the formation of a new components

In high - tech sector, including the warnings, information than any of the variance, the total variance, solidarity, partial solidarity between attributes between attributes and solidarity be simple. In Table resulting output of PCA technique five major factor resulting from the output - observed.

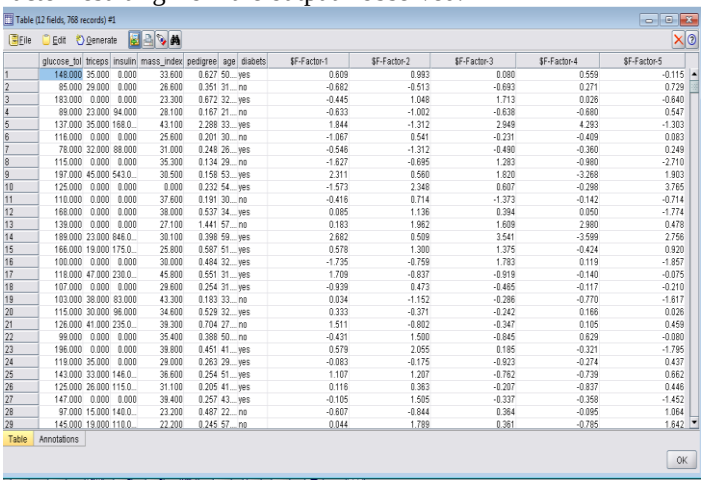

Figure 4 - new components of the execution of the algorithm PCA

Decision tree algorithm:

According to the algorithm, it could be two or more branches. With the survey a tree at the root of the decision to a lower level than a little or no. Each node of the data to decide which split. Decision trees through consecutive separation of data to separate groups are made. The goal in the process of increasing the distance between groups in each. One of the differences between the tree construction methods, the decision is how to measure the distance. Decision trees for predicting a bunch of variables used classification trees. Because these samples in the categories. Every path in the tree usually decided to a leaf is understandable. Of the terms of a tree decision could explain its projections, which is an important advantage.

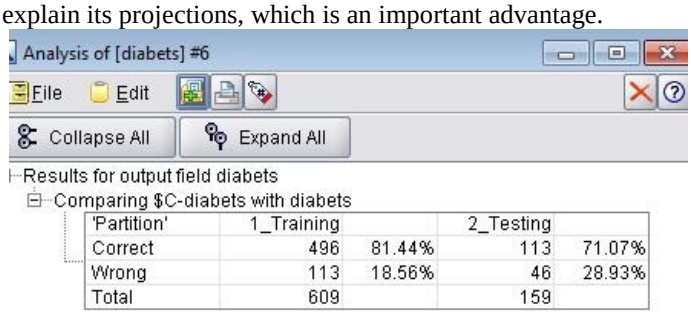

Figure 5. Decision tree [10]

Decision trees frequency of some of the data for each level crossing and tree with variables large anticipated work well. As a result, models quickly, that they are made for the most suitable data sets. If allow unfettered grow spent more time to build a non - intelligent, but it's more important is that fit the data over.

Model C5.0:

C5.0 is the decision to a tree, or a series of laws. The model the Field, the most important information, and classification. Each subsample by the first Turk, or by using different fields, the main categories. This procedure is repeated until other subassemblies fails to other smaller subset or be divided. Finally, the lowest side (where isleaf) test again view was, in fact, that leaves a very important not detected and greed.

The concept of neural networks, related to simulate the learning in humans and the implementation of the computer algorithms. The learning in humans is learning about patterns by iterations. The brain as a system of information processing parallel with the structure of the millions of neuron is formed. That are called nerve tissue of social are neurons that information and messages from one part of the other part of the body. Learning System of neural networks complex, which consists of a series of brain neurons inspired. Although there is a neuron alone is a simple structure, a network of neuron that are connected. Can learn a complex tasks. Artificial neural networks provide the junior level of a nonlinear training are naturally neuron there is real.

Of artificial neural networks would offer to use circuitry hardware and software (algorithms) to build intelligent machines, capabilities of the robots, programs, and so on. This method are able to learn during the process. Neuron biological of four parts:

1 - Dendrite: Information and cells.

2 - Cell body received information.

3-Exxon Information, the cells to another neuron.

4- The confluence of a synapse Exxon from a laptop to dendrite of another neuron say synapse.

In Figure under artificial neuron similarity with biological neuron is shown. Neuron false information processing the smallest units. Neuron inlets by a form of communication in the name of the weight to enter neuron[11].

First, a combined function (typically gathered Σ) is a linear combination of the nodes (x), along with the weight of the imposed related to each node (W) are together and become a scalar value.

$$net_j = \sum_i W_{ij}\, x_{ij} = W_{0j} x_{0j} + W_{1j} x_{1j+\ldots+} W_{Ij} x_{Ij} \quad (2)$$

Table 2. The performance of the algorithm proposed in comparison to other algorithms

data mining technique	Prediction accuracy of diabetes
SVM	77/34
SSVM	76/73
Navies Bayesian	76/30
AD Tree	72/91
Decision Table	71/22
Kstar	69/14
ANN	73/40
FLANN	71/84
MLP	76/89
ANFIS	80.60%
KNN	75.55%
The proposed method	88.02%

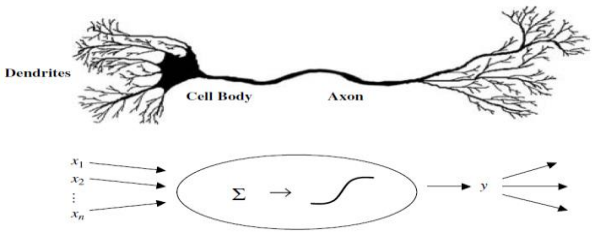

Figure 6. Similarity with artificial neuron biological neurons [12]

To the value of the transfer function is used as input. Neuron in biological, when the combination of neuron to the greatest extent. Signals are sent between the neurons. This is a non - linear behavior. Artificial neural networks. This behavior with regard to the modeling of nonlinear transfer function.

In the table below the results of the implementation of the model with neural network technique is shown. In the model proposed in this paper, we use multiple neural networks. After using neural network in the project, the proposed method of a multilayer perceptron's after the publication of the algorithm for training.

Coincidence Matrix for $N-diabets (rows show actuals)

'Partition' = 1_Training	no	yes
no	351	55
yes	71	132
'Partition' = 2_Testing	no	yes
no	82	12
yes	24	41

Figure 7. The results of the model with neural network multiple techniques

Results for output field diabets

Individual Models

Comparing $C-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	487	79.97%	112	70.44%
Wrong	122	20.03%	47	29.56%
Total	609		159	

Comparing $N-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	474	77.83%	125	78.62%
Wrong	135	22.17%	34	21.38%
Total	609		159	

Agreement between $C-diabets $N-diabets

'Partition'	1_Training		2_Testing	
Agree	512	84.07%	130	81.76%
Disagree	97	15.93%	29	18.24%
Total	609		159	

Comparing Agreement with diabets

'Partition'	1_Training		2_Testing	
Correct	432	84.38%	104	80%
Wrong	80	15.62%	26	20%
Total	512		130	

Figure 8. The results of the implementation of the proposed model

Comparison of the results

For comparison, the proposed method with other existing methods has been included in the table. In which all of the methods discussed in this dissertation with accuracy and classification criteria.

Went on ways to implement the first compared and then each performance compared with each other in our schedules. In the model implemented all specimens were used, and all the default settings were considered software, for example, including input layer with 4 neuron, a hidden layer with the number of 19 and output layer with the number of the cases, including one neuron. Finally, after the implementation of the following results business model that is shown below.

In Figure above the results of the implementation of natural birth for training and test data is shown separately. First is also the performance of the decision and neural network separately, and then to the proposed model.

In Table up as well as you can see the performance of the algorithm proposed in comparison higher accuracy with other techniques. This shows that for reducing the dimensions of feature in diagnosis is better algorithms use playoffs. Or indeed all the way from a combination of functions.

References

[1]. Morgan Kaufmann, "Data Mining - Concepts and Techniques - Second Edition Second Edition [2006], Jiawei Han University of Illinois at Urbana-Champaign "MichelineKamber"

[2]. Ian H. Eibe Frank Mark A. Hall, "Data Mining- Practical Machine Learning Tools and Techniques Third Edition" [2011] Shahrabi J, Shakoorniaz V. Concepts of data mining in; Tehran Shahrabi J, ZolghadrShojaei A. Advanced data mining: Concepts and algorithms. □□□□; Tehran, Jahaddaneshgahi

[5] Kantardzic Mehmed, "*Data Mining: Concepts, Models, Methods, and Algorithms*" Second edition. John Wiley & Sons press,2011.

[6] Larose, Daniel T., "*Discovering knowledge in data An Introduction to Data Mining*", John Wiley & Sons press,2005.

[7] Jiawei Han, Micheline Kamber, Jian Pei , "*Data Mining:Concepts and Techniques*", Third Edition, elsivier,2006.

[8]. ZhaoHui, "Tang" and "Jamie MacLennan" [2005] Data Mining with SQL Server [2005] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics,

[10] Data mining Mhmd Kamtardzyk Translation: AmirSamiraSummer1385Publisher: Computer Science

[11] Ozbakır Lale, Baykasoglu Adil and Kulluk Sinem. "A soft computing-based approach for integrated training and rule extraction from artificial neural networks:DIFACONN-miner", Applied Soft Coputing, 2010 , Vol.10,pp: 304-317.

[12] Martin T.Hagan,Howard B.Demuth,"Neural Network Design" , PWS PublishingCompany,1996.