International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:8, 2013

Analysis of a Population of Diabetic Patients Databases with Classifiers

Murat Koklu and Yavuz Unal

***Abstract*—**Data mining can be called as a technique to extract information from data. It is the process of obtaining hidden information and then turning it into qualified knowledge by statistical and artificial intelligence technique. One of its application areas is medical area to form decision support systems for diagnosis just by inventing meaningful information from given medical data. In this study a decision support system for diagnosis of illness that make use of data mining and three different artificial intelligence classifier algorithms namely Multilayer Perceptron, Naive Bayes Classifier and J.48. Pima Indian dataset of UCI Machine Learning Repository was used. This dataset includes urinary and blood test results of 768 patients. These test results consist of 8 different feature vectors. Obtained classifying results were compared with the previous studies. The suggestions for future studies were presented.

***Keywords*—**Artificial Intelligence, Classifiers, Data Mining, Diabetic Patients.

I. Introduction

HEALTHCARE information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are essential. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytical tools currently available. It has been shown to be a valid, sensitive, and reliable method to discover patterns and relationships. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources [1], [2].

In recent times, the number of people suffering from diabetes is increasing day by day. It is a disease in which body does not produce insulin or use it properly. This increase the risks of developing, kidney disease, blindness, nerve damage, blood vessel damage and contribute to heart disease [3].

There are two types of diabetes; Type-1 diabetes - also called insulin dependent and type-2 diabetes which is with relative insulin deficiency. Patients with type 2 diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose levels [4]. To diagnose diabetes disease at an early stage is quite a challenging task due to complex inter dependence on various factors. There is a critical need to develop medical diagnostic decision support systems which can aid medical practitioners in the diagnostic process. This study deals about the classification of Type II diabetes.

The dataset used in this study is "The Pima Indians Diabetes Data Set" which was taken from the UCI Machine Learning Repository [5]. The original owner of this data set is the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of this dataset from larger database. In particular, all patients selected are females at least 21 years old of Pima Indian heritage.

Weka software package was used throughout this study. Weka software is a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. Weka system is open source software issued under GNU General Public License, where it can be modified by anybody for use [6].

II. Related Work

So many researcher studied about diabetes diagnosis systems. Yasodaet al. classified PIMA diabetes data set with different machine learning algorithms such as Bayes Network classifier, REP tree, Random tree, J48 and appriori [7]. Ming-Yan et al. designed an expert system that can diagnose the diabetes [8].

Han et al. implemented a classifier on PIMA dataset by decision tree that was formed with RapidMiner [9]. Jayalakshmiet al. designed a system that was applied to PIMA dataset for classification aim. The system made us of Artificial Neural Network for classification [2].

Patilet al. produced association rules for PIMA dataset [10]. AlJarullahdesigned a system for diagnosis of diabetes by using PIMA dataset and decision tree that was formed with WEKA software [11].

Aroraet al. used UCI database for both classification and comparison of the classification methods they used. They made use of 5 different dataset (including PIMA) from UCI and applied J48 and Multilayer Perceptron (MLP) for classification and comparison aims [12].

Murat Koklu is with the Selcuk University, Technical Education Faculty, Department of Electronics and Computer Education, Konya, Turkey.(e-mail: mkoklu@selcuk.edu.tr).

Yavuz Unal iswith the Amasya University, Education Faculty, Computer Education and Instructional Technology Department, Amasya, Turkey (e-mail: yavuz.unal@amasya.edu.tr).

III. Methodology

A. Attributes of Dataset

Dataset is composed of 768 instances as seen in Table I. Each patient is characterized in data set by 8 attributes. All attributes are numerical values.

TABLE I
THE PIMA INDIAN DATASET ATTRIBUTES

No	Name	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
3	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hour Serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg/(height in $m)^2$)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric

This attributes are: Diastolic blood pressure, plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function ,age (years), Class variable (0 or 1).

TABLE II
THE STATISTICAL DATUM OF ALL FEATURE

Attribute	Min value	Max value	Mean Value	Standard Dev.
1	0	17	3.845	3.37
2	0	199	120.895	31.973
3	0	122	69.105	19.356
4	0	99	20.536	15.952
5	0	846	79.799	115.244
6	0	67.1	31.993	7.884
7	0.078	2.42	0.472	0,331
8	21	81	33.241	11.76

All statistical datum of each feature vector in dataset are given in Table II.

B. Classification

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. Cluster analysis techniques are used mainly to aggregate objects into groups according to similarity measures. Many studies have been made to compare the many different methods of classification which have been developed.

In this study, three different classifiers (MLP, J48, and Naïve Bayes) have been applied to Pima Indians Diabetes Data Set. These are explained here.

C. Multi Layer Perceptron

The artificial neural network (ANN) or neural network in short, is inspired by simulating the function of a human brain. A neural network can be used to represent a nonlinear mapping between input and output vectors. Neural networks are among the popular signal-processing technologies. In engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters [13], [14]. A general network consists of a layered architecture, an input layer, one or more hidden layers and an output layer [15]. The Multilayer perceptron (MLP) is an example of an artificial neural network that is used extensively to solve a number of different problems, including pattern recognition and interpolation. Each layer is composed of neurons, which are interconnected with each other by weights. In each neuron, a specific mathematical function called the activation function accepts input from previous layers and generates output for the next layer. In the experiment, the activation function used is the hyperbolic tangent sigmoid transfer function [16].

D. J48

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is a software extension and thus improvement of the basic ID3 algorithm designed by Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [10]. For inducing classification rules in the form of Decision Trees from a set of given examples C4.5 algorithm was introduced by Quinlan. C4.5 is an evolution and refinement of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. A set of records are given [17].

E. Naïve Bayes

The Naïve Bayes Classifier technique is mainly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outclass more refined classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease.

Why chosen Naïve Bayes: Naive Bayes or Bayes' Rule is the foundation for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of discovering and considerate data.

Bayes Rule: A conditional probability is the likelihood of specific decision, C, given some evidence/observation, E, where a need relationship occurs between C and E [18].

IV. Conclusion

We have taken dataset values eight features used to classify into 2 classes: (positive for diabetes, negative for diabetes)

TABLE III
TABLE TYPE STYLES

No	*Classification Types*	*Accuracy*	*Time(sec.)*
1	MLP	75.130	1.13
2	J48	73.828	0.08
3	Naïve Bayes	76.302	0.01

Classification results are shown in Table III. With respect to these results, the selected techniques from specialized literature achieved prediction accuracy ratios of 75.13%, 73.828%, and 76.302% for MLP, J48, and Naïve Bayes, respectively. It can be seen that Naïve Bayes outperforms the best performance among the others on Pima Indian dataset.

These classifiers are generally used for the fields of data mining, biomedical engineering, and diagnosing the patients in medicine.

ACKNOWLEDGMENT

This study has been supported by Scientific Research Project of Selcuk University.

REFERENCES

[1] M. Khajehei, F. Etemady, "Data Mining and Medical Research Studies," cimsim, pp.119-122, 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, 2010
[2] T. Jayalakshmi, A. Santhakumaran, , "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks," Data Storage and Data Engineering (DSDE), 2010 International Conference on , vol., no., pp.159-163, 9-10 Feb. 2010
[3] E.I.Mohamed, R.Linderm, G.Perriello,N.Daniele, S.J.Poppl, & A.DeLorenzo. "Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis," Diabetes nutrition metabolism, 15(4),215–221.202.
[4] J.C.Pickup, G. Williams, (Eds.), Textbook of diabetes, Blackwell Science, Oxford.
[5] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[6] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/
[7] P. Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patient Databases in Weka Tool", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011.
[8] Jiang Ming-Yan; Chen Zhi-Jian; , "Diabetes expert system," Intelligent Processing Systems, 1997. ICIPS '97. 1997 IEEE International Conference on , vol.2, no., pp.1076-1077 vol.2, 28-31 Oct 1997
[9] Jianchao Han; Rodriguez, J.C.; Beheshti, M.; , "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner," Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on , vol.3, no., pp.96-99, 13-15 Dec. 2008
[10] B.M Patil, R.C. Joshi, D. Toshniwal, "Association Rule for Classification of Type-2 Diabetic Patients," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.330-334, 9-11 Feb. 2010
[11] A. A. Aljarullah, "Decision tree discovery for the diagnosis of type-II diabetes", International Conference on Innovations in Information Technology, pp. 303-307, 2011.
[12] R. Arora, Suman, "Compatative analysis of classification algorithms on different dataset using WEKA", International Journal of Computer Application, Vol. 54, no:13,pp.21-25, 2012.
[13] D.Marquardt, "An Algorithm for Least Squares Estimation of Non-Linear Parameter", J. Soc. Ind. Appl. Math., pp. 1963.
[14] L.Fausett, "Fundamentals of Neural Networks Architecture.Algorithms and Applications", Pearson Prentice Hall, USA, 1994.
[15] M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail "The Use of a Neural Network Technique for the Prediction of Water Quality Parameters of Axios River in Northern Greece", Journal 0f Operational Research, Springer-Verlag, Jan 2005, pp. 115-125.
[16] L.Khuan, N.Hamzah and R Jailani, "Water Quality Prediction Using LS-SVM with Particle Swarm Optimization",Second International Workshop on Knowledge Discovery and Data Mining, China, 2009, pp. 900-904.
[17] A. K. Sharma, " A comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No:5,pp. 1890-1895, 2011.
[18] J. Kim, D. X. Le and G. R. Thoma, " Naive Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles," DMIN , pp.373-378, 2008.