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

Computer Engineering

Elixir Comp. Engg. 81 (2015) 32040-32043

ISSN: 2229-712X

Hybrid algorithm for detecting diabetes

Amir Amiry
Department of Computer Engineering, Malayer branch, Islamic Azad University, Malayer, Iran.

ARTICLE INFO

Article history:
Received: 2 November 2014;
Received in revised form:
19 April 2015;
Accepted: 29 April 2015;

Keywords
Data Mining,
Diagnostic,
Diabetes,
Genetic Algorithms,
Decision Trees ,
Artificial Neural Networks.

ABSTRACT
The purpose of this research is review of rules and application area of predictive data mining in medical sciences and presenting a frame work for creating, evaluation and exploitation of data mining models in this. In this paper a new method is presented for the diagnosis of diabetes. In this article we want to use the combination of artificial intelligent techniques such as genetic algorithm for attribute reduction from 8 to 4, the artificial neural decision tree to select the best possible conditions Neural Network, and education to estimate adaptability learn Machine, in order to identify the diagnosis of uses. The present study compared to older methods and meta-heuristic model involving a combination of genetic algorithms and neural networks and decision trees in the diagnosis of diabetes is based on the laws and properties of the compound.

Introduction

One of the most common diseases known diabetes or diabetes in the world. According to reports, about 222 million people worldwide suffer from the disease [1,2،3]. Mainly through testing blood sugar Diabetes (Plasma, Glucose, Fasting (which is detected is divided into three categories. Diabetes type I or insulin-dependent diabetes (IDDM) mostly at a young age, children can be seen. Diabetes type II or increasing use of fuzzy sets in medical diagnosis has been observed. been many approaches in terms of several studies which, at best, the neural network RBF [4], neural network MLP [5], KNN near neighborhood [6], evolutionary algorithms [2] algorithm and SVM [7], and fuzzy decision trees [8] and hybrid systems, Fuzzy-BPSO-SVM-NN [9] are introduced and investigated. Firstly, the general procedure described Then according to the contents of (1) to describe the outline of discussion.

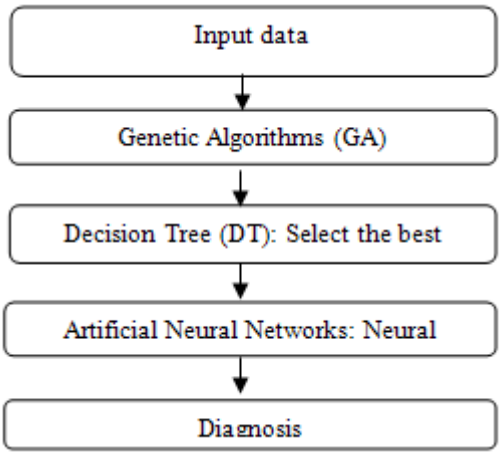


Figure 1. Model for proposed method

Feature selection algorithms identify the features that are relevant but not redundant to the solution. The main task is to rank the relevant features based on their fitness values. There are many algorithms that use a greedy search through the solution space. Decision tree algorithms such as Quinlan's ID3 [10] and C4.5 [11], CART proposed in [12], are some of the most successful supervised learning algorithms. Michalski (1980) proposed the AQ learning algorithm. Narendra and Fukunaga (1977) presented a Branch and Bound algorithm. A well known algorithm that relies on relevance evaluation is RELIEF [13].

Subset search algorithms [14] search and capture the goodness of each subset. There are again many algorithms that are exhaustive, heuristic and random search. Clustering algorithms are also used for feature selection process for which ROCK [15], CACTUS [16] are few of them. Naive Bayes vs Bayes' Rule is the basis for many machine-learning and data mining methods [17]. As for other clinical diagnosis problems, classification systems have been disease diagnosis problem. Among many[18], Tooling, RA obtained 50.00% classification accuracy by using algorithm. [18] WEKA, RA obtained a classification accuracy of 58.50% using Induct algorithm while Tool diag, RA reached to 60.00% with RBF algorithm. [18] Again, WEKA, RA applied FOIL algorithm to the problem and obtained a classification accuracy of 64.00%. [18] MLP+BP algorithm that was used by Tool diag, RA reached to 65.60%[18].

The classification accuracies obtained with T2, 1R, IB1c and K* which were applied by WEKA, RA are 68.10%, 71.40%, 74.00% and 76.70%, respectively. [18] Robert Detrano used logistic regression algorithm and obtained 77.0% classification accuracy.

Cheung utilized C4.5, Naive Bayes, BNND and BNNF algorithms and reached the classification accuracies 81.11%, 81.48%, 81.11% and 80.96%, respectively [18]. Among the various methods given above the proposed method proves to be more efficient and cost-effective.

Proposed Method

First data from the database [19] extracted using a genetic algorithm to select features are more important then the data by the decision tree (DT) have been exploring the best space research and the way the input space unit to patients (output) relative is sick or healthy depicted symbolically by providing a

Tele:
E-mail addresses: Amiryamir57@gmail.com

tree. artificial neural networks to estimate, education, adaptability, machine learning, is used to detect and diagnose the disease. Figure 1 illustrates the entire process of this research

Data Set Description

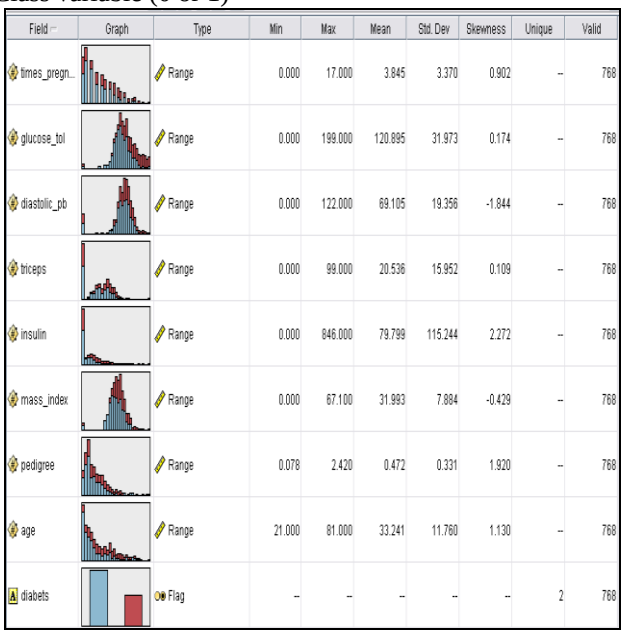
In our work we have used Pima Indian Diabetes data sets[19], for training and testing the neural network model.

Pima Indian Diabetes Dataset

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)



Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
times_pregn...		Range	0.000	17.000	3.845	3.370	0.902	--	768
glucose_tol		Range	0.000	199.000	120.895	31.973	0.174	--	768
diastolic_pb		Range	0.000	122.000	69.105	19.356	-1.844	--	768
triceps		Range	0.000	99.000	20.536	15.952	0.109	--	768
insulin		Range	0.000	846.000	79.799	115.244	2.272	--	768
mass_index		Range	0.000	67.100	31.993	7.884	-0.429	--	768
pedigree		Range	0.078	2.420	0.472	0.331	1.920	--	768
age		Range	21.000	81.000	33.241	11.760	1.130	--	768
diabets		Flag	--	--	--	--	--	2	768

Figure 2. Database of distribution

Extracting Optimal Patterns Using Genetic Algorithm Genetic algorithms, vision of genetics and the theory of evolution Darwin and is based on the survival of the greatest or natural selection. A common use the genetic algorithm to use it as a function is the optimal [20-22]. In the genetic algorithms, how genetic evolution simulation organisms. In each phase of implementation of the genetic algorithm, a group of the search of random processing. so that every point to a series of characters is attributed to the and on the following the application of the genetic operators. Then the tail Decode came to a new points in the search space. In late on the basis that the objective function in any of the value of the points. the likelihood of them in the next stage will be determined [23-27]. may be extracted in all the characteristics of target samples from the goal, or even in some cases may exacerbate the results. to solve this problem in the paper of binary evolutionary algorithm to find the dominant features. We the evaluation to consider 1

$$Fitness=w(s)+(n-s)/n \quad (1)$$

w (s) shows the taxonomy for the accuracy of the following a series of s, s represents the number of selected features and n represents the total number of features. (according to Equation 1) 4 to 8 feature, which includes: blood plasma glucose concentration in two hours, body mass index, diabetes, and is the age.

```
begin
i=0 /* i: number of iteration*/
initialize P(i) /* P(i): population for iteration I */
compute f(P(i)) /* f: fitness function */
perform until (non termination condition)
begin
i=i+1;
choose two parents P1 and P2 from P(i-1)
perform genetic operations
{
crossover;
mutation;
}
reproduce a new P(i)
compute f(P(i))
end-perform
end
```

Figure 3. Schema - a genetic algorithm [13]

The Implementation of The Model of The Neural Network and Tree Decision C & R tree a method based on parts and components of a return to the educational records into pieces and division by greed and other important sectors of any step, and we are divided. You see decision trees in Figure 4.

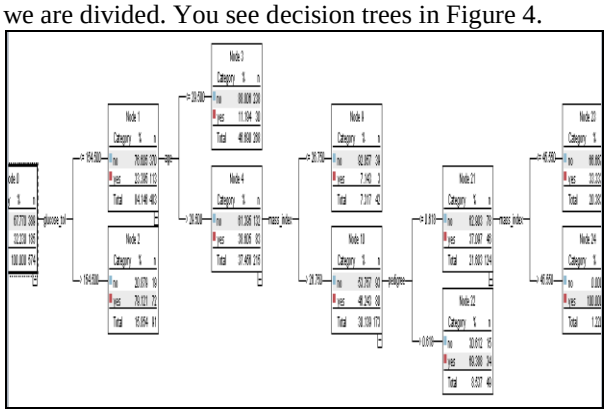


Figure 4. Decision trees you see code decision trees in of the proposed model in Figure 5.

```
glucose_tol <= 127.500 [ Mode: no ]
        age <= 28.500 [ Mode: no ] => no
        age > 28.500 [ Mode: no ]
                mass_index <= 26.350 [ Mode: no ] => no
                mass_index > 26.350 [ Mode: no ]
                        glucose_tol <= 99.500 [ Mode: no ] => no
                        glucose_tol > 99.500 [ Mode: yes ]
                                pedigree <= 0.561 [ Mode: no ] => no
                                pedigree > 0.561 [ Mode: yes ] => yes
glucose_tol > 127.500 [ Mode: yes ]
        mass_index <= 29.950 [ Mode: no ]
        glucose_tol <= 145.500 [ Mode: no ] => no
                glucose_tol > 145.500 [ Mode: yes ] => yes
        mass_index > 29.950 [ Mode: yes ]
                glucose_tol <= 157.500 [ Mode: yes ]
                        age <= 30.500 [ Mode: no ]
                                diastolic_pb <= 61 [ Mode: yes ] => yes
                                diastolic_pb > 61 [ Mode: no ] => no
                        age > 30.500 [ Mode: yes ] => yes
                glucose_tol > 157.500 [ Mode: yes ] => yes
```

Figure 5. Code decision trees in of the proposed model

neural networks, one of the most powerful functions are expected. They to at least to learn mathematics and science. software used by combining several properties tries to the neural networks in bottlenecks and of these traps. Six method of education to make the neural networks are:

Six method of education to make the neural networks

1. **Quick:** these laws to seek and find the characteristics of the data, by a framework (topology).
2. **Dynamic**: this way to create a primary school, but the way it set out on the basis of and / or and undercover units).
3. **Multiple**: this method of several different topologies at once create (the number of networks depends on the training data).
4. **Prune**: This major networks beginning and then with the greed of the weakest units in the mud and mud Education entrance. The way more slowly and slower than other methods, but the results of that and better than others.
5. **RBFN:** the way a technology like the clustering algorithm K-means is about to work on the basis of the value of the target field.
6. **Exhaustive prune**: the way to the network is greed, beginning with a large network, starting with the greed of the weak network of secret layers and network, smaller and more. In this way, training parameters are choosing to seek ways of possible models to get the best models that we can. this method is the slowest neural network and a lot of time spent on education, but most of the best results. the most important characteristic of filling their top decision-making power in crushing a complex issue as a result of the smaller issues and provide a solution is understandable. priority attributes after three neural network mentioned in the Figure 6. As is apparent feature blood plasma glucose concentration in two hours, after which the priority of body mass index properties in the neural networks of all the attributes and characteristics of the frequency of pregnancy and priorities of blood pressure.

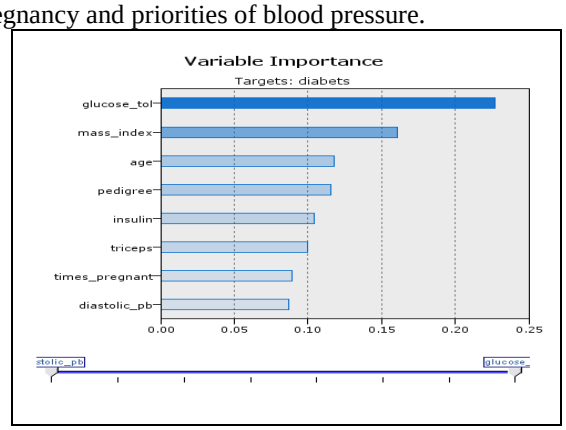


Figure 6. The priority of attributes after three neural network

The evaluation of the proposed method. In the first data from the database and using a binary genetic algorithm randomly initial population that at the beginning of a chromosome 8 Biti would be elected and using fitness function to extract the dominant features of course which led to a decline in the features of 8 4 to. Then, with the use of the decision of the trees that would be able to produce descriptions of the understandable, a collection of data that can be classified as to predict and used [28] production rules. decision-making structure could be in the form of computational techniques to describe, and classification of a set of data also help introduce [29]. After using neural network in this project, the proposed method of a multilayer perception and Quick RBFN and then with the algorithm for training.

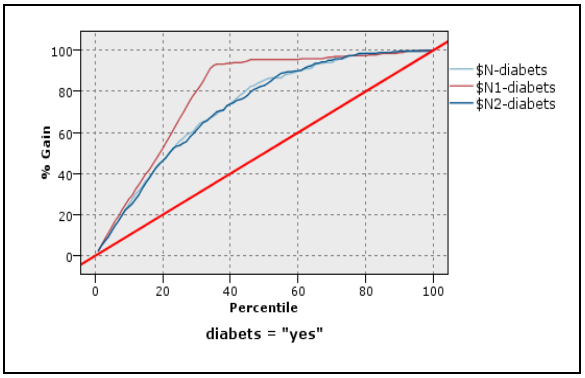


Figure 7. Three neural network with the decision

Confusion matrix

Confusion matrix a tool to show the accuracy of the Conclusions Several different classifications to show that the relationship between the results and the use of the anticipated. that its format, according to the following formulas. in which

Table 1. Table Matrix

	PREDICTED CLASS		
		Class0.0	**Class1.0**
ACTUAL CLASS	Class0.0	a (TP)	b (FN)
	Class1.0	c (FP)	d (TN)

- True Positive : expected the number of the right in the class 0.0
- False Negative: the number of predictions wrong in the class 0.0
- FP: the number of predictions wrong in the class 1.0
- TN: expected the number of the right in the class 1.0 of the following formulas for the evaluation of the models

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Error = \frac{c+d}{a+b+c+d} = \frac{FN+FP}{TP+TN+FP+FN} \quad (3)$$

First, the result of the proposed model in Table 2 we see. So with regard to the results achieved in the diabetic patients Quick neural network of other models better neural network. Finally, with regard to the implementation of the proposed model forecast accuracy educational records 96/16 % forecast accuracy and records 94/49 % even test.

Table 2. The Results of the System

Individual Models

Comparing $R-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	465	76.35%	109	68.55%
Wrong	144	23.65%	50	31.45%
Total	609		159	

Comparing $N-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	481	78.98%	123	77.36%
Wrong	128	21.02%	36	22.64%
Total	609		159	

Comparing $N1-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	593	97.37%	153	96.23%
Wrong	16	2.63%	6	3.77%
Total	609		159	

Comparing $N2-diabets with diabets

'Partition'	1_Training		2_Testing	
Correct	482	79.15%	116	72.96%
Wrong	127	20.85%	43	27.04%
Total	609		159	

Agreement between $R-diabets $N-diabets $N1-diabets $N2-diabets

'Partition'	1_Training		2_Testing	
Agree	415	68.14%	98	61.64%
Disagree	194	31.86%	61	38.36%
Total	609		159	

Comparing Agreement with diabets

'Partition'	1_Training		2_Testing	
Correct	405	97.59%	93	94.9%
Wrong	10	2.41%	5	5.1%
Total	415		98	

Based on the above formulas, the accuracy values are obtained

$$Accuracy = \frac{457 + 236}{457 + 236 + 43 + 32} = 90.23$$

Error value to obtain:

$$Error = \frac{43 + 32}{457 + 236 + 43 + 32} = 9.07$$

Confusion matrix obtained from the following article, we would like to see.

Table 3. Confusion matrix

diabets	no	yes
no	457	43
yes	32	236

Deduction

In this opportunity also to compare the performance of systems relatively similar) as far as the authors are see hybrid system similar Found (was shown by the hybrid system proposed complexity lower results reasonably well in Table 4 compares the method with [30] ANFIS, evolutionary algorithm [30], k the near neighbors [6] and DT [31] and has been

Table 4. The Results of The System

percent forecast accuracy diabetes	data mining technique
77/34	SVM
76/73	SSVM
76/30	Navies Bayesian
72/91	AD Tree
71/22	Decision Table
69/14	Kstar
73/40	ANN
71/84	FLANN
76/89	MLP
75.55%	Amin K near neighbor
90/23%	PROPOSED

References

[1] World Health Organization, Diabetes Center, Fact SheetN2312, www.who.int/mediacentre/fa ctsheets/fs312 /en.

[2] Centers for Disease Control and Prevention, National Diabetes, Fact Sheet 2011 www.cdc.gov /diabetes.

[3] Mohammad Naeem Abadi, Nvshaz Chmachar Amir Ahmadi, E. Thamy and Hossein Rabbani, "Diagnosis of diabetes using SVM" ISCEE 14th 19.

[4] Siti Fahanah Bt Jaafar and Darmawaty Mohd Ali, “Diabetes Mellitus Forecast using artificial neuarl network(ANN)”, IEEE2005

[5] Seyyed Ehsan Thamy, Muhammad Ali Khalil Zadeh, "Intelligent Diagnosis of diabetes using multilayer neural networks," Conference on Biomedical Engineering 28 November 1374 - Iran Conference, Sahand University of Technology 2

[6] Y. Jiang and Z. Zhou, “Editing Training Data for kNN Classifiers with Neural Network Ensemble”, in Proc. ISNN (1),2004, pp.356-361

[7] American Diabetes Association, Diabetes Basics www.diabetes.org/diabetes-basics

[8] Gharekhani Azam Mohammad Fyvzy and J. Hdadnya "offers a new approach for diagnosis of diabetes based on the exact composition of the system (FUZZY), data mining systems (DT) Nerve Fuzzy adaptive systems (ANFIS)"، ikt2212-4th

[9] Muhammad Fyvzy, A. and J. Gharekhani Hdadnya "provide an intelligent hybrid system to detect diabetes" th 20 icee2012

[10] Quinlan, J.R., (1986). Induction of decision trees, Machine Learning 1, 81–106.

[11] Quinlan, J.R., (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco.

[12] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., (1984). Classification and Regression Trees. Wadsworth, Belmont,CA.

[13] H. Liu and H. Motoda, (1998), Feature Selection for Knowledge Discovery and Data Mining,Boston: Kluwer Academic Publishers.

[14] M. Dash, H. Liu and H. Motoda, (2000), “Consistency Based Feature Selection,” Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp. 98-109,Springer-Verlag.

[15] S. Guha, R. Rastogi, and K. Shim, (1999). ROCK, “A Robust Clustering Algorithm for Categorical Attributes” Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, April.

[16] V. Ganti, J. Gehrke, and R. Ramakrishnan, (1999), “CACTUS-Clustering Categorical Data Using Summaries,” Proceedings of the ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA.

[17] Tang, Z. H., MacLennan, J, (2005).: “Data Mining with SQL Server 2005”, Indianapolis: Wiley.

[18] Kemal Polata, & Salih Güne¸sa & Sülayman Tosunb, (2007), Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing, ELSEVIER,PATTERN RECOGNATION.

[19] Newman,D.J.,Hettich, S.,Blake, C.L.S., & Merz,C.J., 1998. UCI “repository of machine learning database”, Irvine, CA: University of California, Dept. of Information and Computer Science.archive.ics.uci.edu/ml/datasets/Pima +Indians+Diabetes

[20] Tseng, L.Y. and Yang, S., “Genetic algorithms for clustering, feature selection and classification” ,IEEE Int. Conference on Neural Networks, pp.1612-1616،1997.

[21] Bala,J.,Huary,J.,Vafaie,H.,De jong, K. and Wechslev,H. “Hybrid learning using genetic algorithms and decision trees for pattern classification” , IJCAI conference , Montreal , August 19-25،1995.

[22] Siedlecki,W. and Sklansky,J., “A note on genetic algorithms for large scale pattern selection”, Pattern Recognition Letters , vol.10،335-347،1989.

[23] M. Mitchell. An Introduction to Genetic Algorithms, MIT Press, Cambridge, MA, 1996.

[24] D. Beasley, D. Bull and R. Martin. An Overview of Genetic Algorithms: Part 1, Fundamentals, University of Cardiff, Cardiff, 1993.

[25] P. Brigger. An Overview of Genetic Algorithms, URL: http://ltswww.epfl.ch /pub_files/brigger/thesis_html/ , 1995.

[26] D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, 1998.

[27] S. Mitra, K. M. Konwar and S. K. Pal, "Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation",Journal of IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol.32, pp.328-339, 2002.

[28] Javad Mahjoub, A.. F.. A.. Martyrs, "the estimated height of wind waves in Neka help decision trees", 8 conference IDMC 20-21 Novomber 2007, Amirkabir University of Technology.

[29] Kantardzic, M, “ Data Mining: Concepts, Models, Methods, and Algorithms”, John Wiley & Sons, 2003

[30] S. E. Thamy, SA. M.. Bamshky, M.. AS. Khalil Zadeh, "type 1 diabetes diagnosis algorithm using ANFIS, GA-NN", First Joint Conference on Fuzzy Systems and Intelligent Systems, Shhyryvr 1386, Ferdowsi University of Mashhad.

[31] Asma A. AlJarullah, King Saud University," Decision Tree Discovery for the Diagnosis of Type II Diabete" 2011 international Conference on Innovation in Information Technology.