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

Computer Engineering

Elixir Comp. Engg. 79 (2015) 30341-30344

ISSN: 2229-712X

A hybrid method based on optimization algorithm of particle motion (PSO) to predict heart disease

Mitra Mohamadi
Department of Computer Engineering, Malayer branch, Islamic Azad University, Malayer, Iran.

ARTICLE INFO

Article history:
Received: 12 January 2015;
Received in revised form:
20 January 2015;
Accepted: 7 February 2015;

Keywords
Heart disease, Data mining algorithm,
C & R K algorithm,
The nearest neighbor, PSO.

ABSTRACT

The main cause of morbidity and mortality in modern society is heart disease. Medical diagnosis is important but complex task that must be carefully and effectively. Although considerable progress has been made in the diagnosis and treatment of heart disease, but research must reach the highest accuracy. Access to a large amount of medical data requires powerful tools for analyzing the resulting data to extract useful knowledge. Data mining is an effective analysis tool for discovering hidden relationships and trends in the data. This paper heart disease through data mining algorithm C & R, the algorithm K nearest neighbor algorithm to move the mass of particles (pso) and improved algorithms, k-nearest neighbor algorithm pso investigated. In this study, we improved the effectiveness of these algorithms see for heart disease. Thus we see that data mining can identify or predict high or low risk of heart disease.

Introduction

In the literature, author of using different algorithms and different characteristics of a heart attack, expected to be smart and efficient use data mining is paid [1]. To predict a heart attack, significantly 15 feature. As a result, the use of data mining techniques and prediction in the same set of data shows that decision tree better than other methods [2] .

In research 313 on in class, and two natural heart patients done[5]. To identify and prediction heart attacks of clustering techniques used data mining. One of the main clustering of data mining is aiming to grouping data to meaningful classes (clusters). So that the resemblance between a bunch of data similarity between the highest and lowest data from two separate cluster[3].

Database:

In this study, the database that used consists of a set of data from the heart of the Imam Ali hospital patients (RA) Kermanshah. It includes 396 is record after prepare and clean - up in the software Sql Server all the records useful was diagnosed with no records and was eliminated. Database include 12 field that includes using them and with the help of the existing prediction models to predict whether these people may be infected with heart disease or not.[4]

Input parameters include:

Age, Blood - Heart Disease, sugar, beat, Cholesterol HDL LDL Smoking, Gender, Blood pressure, PTT.

That the number of parameters studied included the 11 field.

Normalization:

normalization scale change data is so that led them to a narrow range and defined as the distance between the 1 - 1 map. Normalization causes large - scale data to divert his side. In this study of normalization Min Max is used.

$$v = \frac{v - min_a}{max_a - min_a}(new\,max_a - new\,min_a) + new\,min_a$$

It establishes a linear transformation on the main event. min_a that assumption and max_a, respectively, according to a minimum and maximum values.

A normal - - min, then each - a little v of A in the [newmin $_a$, new max_a] map which are normal - - min - tben each of the relationship between the original values.

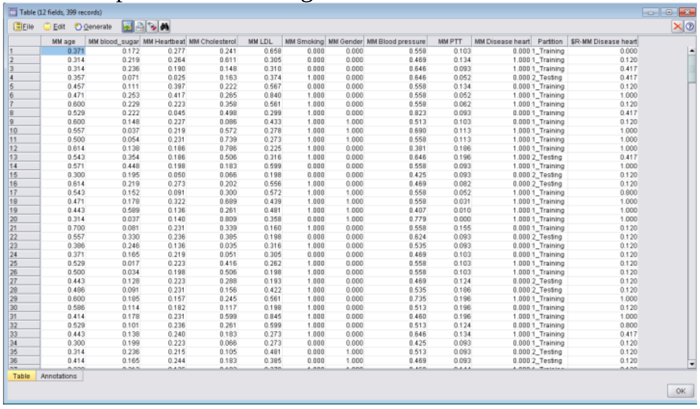
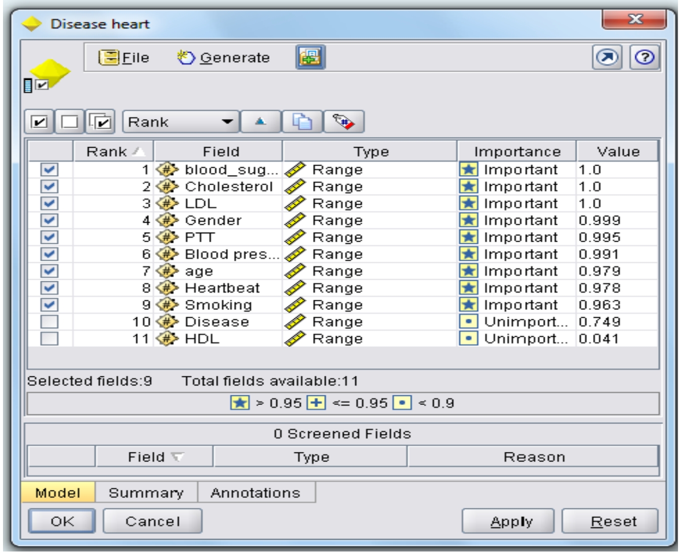


Table (12 fields, 399 records)

	MM age	MM blood_sugar	MM Heartbeat	MM Cholesterol	MM LDL	MM Smoking	MM Gender	MM Blood pressure	MM PTT	MM Disease heart	Partition	$R-MM Disease heart
1	0.371	0.172	0.277	0.241	0.658	0.000	0.000	0.558	0.103	0.000	1_Training	0.000
2	0.314	0.219	0.264	0.611	0.305	0.000	0.000	0.469	0.134	1.000	1_Training	0.120
3	0.314	0.236	0.190	0.148	0.310	0.000	0.000	0.646	0.093	1.000	1_Training	0.417
4	0.357	0.071	0.025	0.163	0.374	1.000	0.000	0.646	0.052	0.000	2_Testing	0.417
5	0.457	0.111	0.397	0.222	0.567	0.000	0.000	0.558	0.134	0.000	1_Training	0.120
6	0.471	0.253	0.417	0.265	0.840	1.000	0.000	0.558	0.052	1.000	1_Training	1.000
7	0.600	0.229	0.223	0.358	0.561	1.000	0.000	0.558	0.062	1.000	1_Training	0.120
8	0.529	0.222	0.045	0.498	0.299	1.000	0.000	0.823	0.093	0.000	1_Training	0.417
9	0.600	0.148	0.227	0.086	0.433	1.000	1.000	0.513	0.103	0.000	1_Training	0.120
10	0.557	0.037	0.219	0.572	0.278	1.000	1.000	0.690	0.113	1.000	1_Training	1.000
11	0.500	0.054	0.231	0.739	0.273	1.000	1.000	0.558	0.113	1.000	1_Training	1.000
12	0.614	0.138	0.186	0.786	0.225	1.000	0.000	0.381	0.186	1.000	1_Training	1.000
13	0.543	0.354	0.186	0.506	0.316	1.000	0.000	0.646	0.196	1.000	2_Testing	0.417
14	0.571	0.448	0.198	0.183	0.599	0.000	0.000	0.558	0.093	1.000	1_Training	1.000
15	0.300	0.195	0.050	0.066	0.198	0.000	0.000	0.425	0.093	0.000	2_Testing	0.120
16	0.614	0.219	0.273	0.202	0.556	1.000	0.000	0.469	0.082	0.000	2_Testing	0.120
17	0.543	0.152	0.091	0.300	0.572	1.000	1.000	0.558	0.052	1.000	1_Training	0.800
18	0.471	0.178	0.322	0.689	0.439	1.000	1.000	0.558	0.031	1.000	1_Training	1.000
19	0.443	0.589	0.136	0.261	0.481	1.000	1.000	0.407	0.010	1.000	1_Training	1.000
20	0.314	0.037	0.140	0.809	0.358	0.000	1.000	0.779	0.000	1.000	1_Training	1.000
21	0.700	0.081	0.231	0.339	0.160	1.000	0.000	0.558	0.155	0.000	1_Training	0.120
22	0.557	0.330	0.236	0.385	0.198	0.000	0.000	0.624	0.093	0.000	2_Testing	0.120
23	0.386	0.246	0.136	0.035	0.316	1.000	0.000	0.535	0.093	0.000	1_Training	0.120
24	0.371	0.165	0.219	0.051	0.305	0.000	0.000	0.469	0.103	0.000	1_Training	0.120
25	0.529	0.017	0.223	0.416	0.262	1.000	0.000	0.558	0.103	1.000	1_Training	0.120
26	0.500	0.034	0.198	0.506	0.198	1.000	0.000	0.558	0.103	1.000	1_Training	0.120
27	0.443	0.128	0.223	0.288	0.193	1.000	0.000	0.469	0.124	0.000	2_Testing	0.120
28	0.486	0.091	0.231	0.156	0.422	1.000	0.000	0.535	0.186	0.000	2_Testing	0.120
29	0.600	0.185	0.157	0.245	0.561	1.000	0.000	0.735	0.196	1.000	1_Training	1.000
30	0.586	0.114	0.182	0.117	0.198	1.000	0.000	0.513	0.196	0.000	1_Training	0.120
31	0.414	0.178	0.231	0.599	0.845	1.000	0.000	0.460	0.196	1.000	1_Training	1.000
32	0.529	0.101	0.236	0.261	0.599	1.000	0.000	0.513	0.124	0.000	1_Training	0.800
33	0.443	0.138	0.240	0.183	0.273	1.000	0.000	0.646	0.134	1.000	1_Training	0.417
34	0.300	0.199	0.223	0.066	0.273	0.000	0.000	0.425	0.093	0.000	1_Training	0.120
35	0.314	0.236	0.215	0.105	0.481	0.000	1.000	0.513	0.093	0.000	2_Testing	0.120
36	0.414	0.165	0.244	0.183	0.385	0.000	1.000	0.469	0.093	0.000	2_Testing	0.120

Figure 1. Impose normalization



Disease heart

	Rank	Field	Type	Importance	Value
☑	1	blood_sug...	Range	Important	1.0
☑	2	Cholesterol	Range	Important	1.0
☑	3	LDL	Range	Important	1.0
☑	4	Gender	Range	Important	0.999
☑	5	PTT	Range	Important	0.995
☑	6	Blood pres...	Range	Important	0.991
☑	7	age	Range	Important	0.979
☑	8	Heartbeat	Range	Important	0.978
☑	9	Smoking	Range	Important	0.963
□	10	Disease	Range	Unimport...	0.749
□	11	HDL	Range	Unimport...	0.041

Selected fields:9 Total fields available:11

> 0.95 <= 0.95 < 0.9

0 Screened Fields

Field	Type	Reason

Figure 2. Feature Selection

Tele:
E-mail addresses: mitramohamadi374@yahoo.com

Feature model selection:

Feature Selection techniques for reducing the number of technical specifications before applying the data - mining algorithm is used [6]. In data mining, some features available at the base of the important and determining role in carrying out a forecast, but others may have, it matters little or no irrelevant, then it should these fields of database to be eliminated, data mining, with the focus on determining Fields successfully and more carefully. The action Feature Selection techniques. This technique percent of the importance of the fields and the importance of using the % can be diagnosed as the field, it is necessary to act in data mining company or not [7,8].

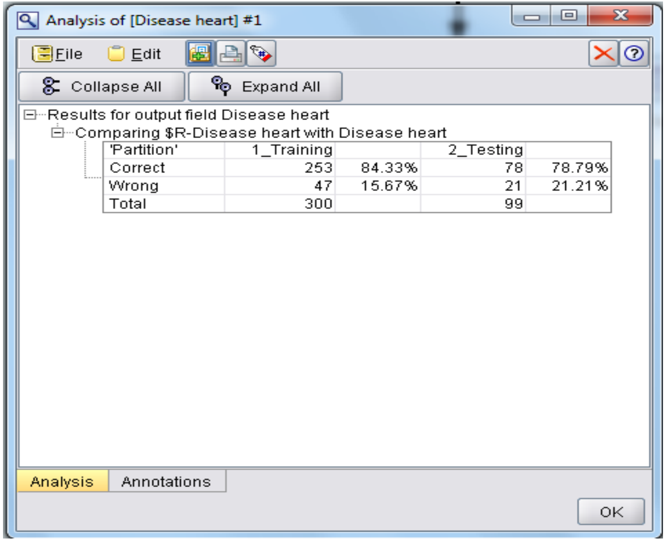
Decision trees:

In classification methods for selecting categories options there is one of the most important and at the same time, the tree in decision - making [9]. Decision tree is a flowchart like that any internal node (90), a non - leaf in tests to determine the quality. Each branch an outcome of the test and each node (bottom) a leaf nodes labeled the class .

If a line is assumed to be given due to the lack of a bunch of X (class), the values qualities tree nodes are tested and a route from the tree roots decision to achieve a leaf nodes during the category .

The use of the decision - making due to their simplicity and speed in the construction and what is common in that category. Generally, the decision - making good accuracy, although the successful use, used to. A structured approach decision trees are generally division and solve the recursive top to bottom, and it is in an attempt to the input variable spaces in the end nodes .

A number of different algorithms, which can be used to build the decision to include: C5. 0, Chaid, Cart, Quest.



Results for output field Disease heart

Comparing $R-Disease heart with Disease heart

'Partition'	1_Training		2_Testing	
Correct	253	84.33%	78	78.79%
Wrong	47	15.67%	21	21.21%
Total	300		99	

Figure 3. The accuracy of the algorithm forecasting model C & R

K algorithm nearest neighbour:

KNN algorithm is one of the most important classification algorithms due to be implemented in many fields, is used. This algorithm for the classification, a record, the gap between the record of all existing lines in a series of training, K similar to the most or the nearest its neighbor's and the record label that is in the majority of the class to new record. Away from the formula for calculating, Euclidean distance [10]. If the rows with n trait to put them into a vector n show next:

$$X=(x_{1,}x_{2},x_{3,\ ....}x_{n}) \qquad (1)$$

$$Y=(y1, y2,y3, \ldots.yn) \qquad (2)$$

$$DIST(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

After interval calculated using the above formula, K to choose the most similar lines and using the label them new data .

Algorithms K data on the nearest neighbor

to develop the model with the algorithm k nearest neighbor, the data sets randomly divided into two parts of education and test fit with the equivalent of 75 % and 25 % - divided. This algorithm with different value of k in Matlab software 2012 was implemented in the end, it was observed that this algorithm with k= 7 compared to other values of k has a better result. The accuracy of the model to the nearest neighbor k in Table 5 - 5 is shown .

Table 2 pointed out that the case - - Mice recognition accuracy of the model 80 \/ 78 % in the training set as well as 25 \/ 71 % in the Test series.

Algorithm PSO

Group - based optimization particles, an optimization technique based on the possibility of laws, which is in the year 1995 by Dr. eberhardt and Dr. Kennedy. The basic idea of this method of collective behavior fish or birds in search of food. Each particle is a fitness value by a fitness function. Whatever little space in search of food in goal (model) movement of birds closer, more worthy also has every particle has a speed that is leading the particle motion. Each particle by following the optimal particles in the current state, to move in the issue continues. In this way every bit of trying to adjust its path and move toward the best personal experience and collective experience, the final solution.

In this study to find the weight of the algorithm mass movement - particles. so that this algorithm is an issue of the optimal - they express - as the weight - to find some way when the method of k the nearest neighbour to use classification. Error classification accuracy of the minimum and maximum category. i. e. The goal in the issue of the optimal - minimize the category - timing error. The data sets randomly divided into two parts of education and test fit with the equivalent of 75 % and 25 % - divided. This algorithm with different value of k in MATLAB software 2012. Finally, it was observed that this algorithm with the values of k = 4 compared to other values of k has a better result.

Table 3 shows that - - to - sample with this model has recognition accuracy 67 \/ 93 % in the training set as well as 75 \/ 78 % in the Test series.

Matrix confusion

According to Table 4 format in which:

* TP: the number of correct predictions in class 0.0

* FN: the number of false predictions in class 0.0

* FP: the number of false predictions in class 1.0

* TN: the number of correct predictions in class 1.0

Table 1 on the basis of the following formula for assessing models.

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

$$Error = \frac{c+d}{a+b+c+d} = \frac{FN+FP}{TP+TN+FP+FN} \qquad (4)$$

Table 1: Features in anticipation of cardiovascular diseases

attributes used	comments
Age	patient age
Blood-sugar	blood sugar
Disease	(except for heart disease)
Heart beat	her heart
Cholesterol	cholesterol levels
HDL	cholesterol full dense
LDL	cholesterol less dense
Smoking	smoking
Gender	gender patients
Blood pressure	blood pressure
PTT	screening test in order to assess their ability In the formation of the blood clot as appropriate

Table 2. The accuracy of the model to the nearest neighbor k

train Performance	٪78/80
test Performance	٪71/25

Table 3. The accuracy of the algorithm k improved nearest neighbor

train Performance	%81/67
test Performance	%75/75

Table 4. Matrix confusion

Matrix confusion	PREDICTED CLASS		
ACTUAL CLASS		Class0.0	Class 1.0
	Class0.0	a (TP)	b (TP)
	Class 1.0	C (FP)	d (TN)

Table 5. Matrix confusion educational complex model to the nearest neighbor k

	K Nearest Neighborhood		
NUM		0.0	1.0
	0.0	187	11
	1.0	56	62

Table 6. Matrix confusion test series model to the nearest neighbor k

	K Nearest Neighborhood		
NUM		0.0	1.0
	0.0	42	2
	1.0	21	15

Table 7. Comparing the results of the models used in the stud

K nearest neighbor		C&R		model
Test	educational	Test	educational	set
%71/25	**%78/80**	**%78/79**	**%81/3**	concentration
%77/2		%70/5		Accuracy
8/22		5/29		Error

To evaluate the results:

After the implementation of the nearest neighbor k model in MATLAB software 2012 matrix turbulence related to training and test data collection, according to the following tables.

Comparison of the results

For comparison, the proposed method with other method of existing - table in which all - discussed with classification accuracy and mentioned.

References:

[1] T. T. Mai Shouman, Rob Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," presented at the Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia

[2] U. A. Jyoti Soni , Dipesh Sharma, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *International Journal of Computer Applications (*0975 – 8887*)*, vol. 17– No.8, pp. 43–48, March 2011 **.**

[3] http://www.salamatnews.com/

[4] B. K Rani, R. K.Srinivas, Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," *(IJCSE) International Journal on Computer Science and Engineering*, vol. 02, pp. 250–255, 2010

[5] Dehghani T, Afshari Saleh M, Khalilzadeh M. A genetic K-means clustering algorithm for heart disease data. 5th Conference of Data Mining of Iran, 2011; Amirkabir University.

[6]Shahrabi J, Shakoorniaz V. Concepts of data mining in Oracle 11.2008; Tehran

[7]Shahrabi J, ZolghadrShojaei A. Advanced data mining: Concepts and algorithms. 2009; Tehran, Jahaddaneshgahi

[8]T. D. Bala Sundar V, N SARAVANAN, "Development of a Data Clustering Algorithm for Predicting Heart," *International Journal* of Computer Applications (0975 – 888),vol. 48– No.7, pp. 8–13, June 2012 **.**

[9]Han.Jand Kamber.M , “Data Mining : Concepts and Techniques”, Second Edition ,Morgan Kaufman Publisher , 2006

[10]T.Larose.D, "Discovery Knowledge indata: An introduction to data mining" ,New jersey,2005