International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.3, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

AN OPTIMIZED APPROACH IN TEXT MINING FOR MEDICAL DIAGNOSIS

SHRUTI MADAN[1] ,RAKESH GARG[2], SUMAN[3]

[1]Student, M.Tech, Computer Science, Hindu College of Engg. Sonipat

[2,3]Assistant Professor, Computer Science, Hindu College of Engg. Sonipat

Article Received: 02/05/2015 | Article Revised on:08/05/2015 | Article Accepted on:12/05/2015

SHRUTI MADAN

ABSTRACT

Data mining methods in medical are used to analyze the medical data information resources. Some important and popular data mining techniques are classification, document clustering, prediction of disease and association rules,. There are varieties of applications of data mining techniques. In medical, data mining plays an important role for predicting diseases. To predict a disease number of tests should be required from the patient..But by using data mining technique the number of tests should be reduced. This plays an important role in time and performance. This research paper analyzes how the data mining techniques are used for predicting the heart diseases. In this paper we will use K-means clustering technique on text data for predicting a disease and then optimize the results by using PSO.

Keywords: Data mining, Text mining, Document Clustering, PSO, K-Means

INTRODUCTION

Medical data mining has high possible & unused quality for exploring the hidden patterns in the medical domain. These patterns can be put to use for medical diagnosis for widely made distribution in the medical data. These data should be in an organized form. Then the organized data is collected and integrated to form a hospital knowledge system. Data Mining is useful for discovering knowledge out of the data and present it in the human understandable form [11]. Large volume of the data that is collected daily is examined in this process. It is a cooperative effort of humans and computers. Humans are experts in describing problems and computers have good search capability, by balancing both we can achieve the better results. There are two primary goals of data mining: *prediction* and *description.*[10] In Data Mining the Prediction of disease plays an important role. There are different types of diseases predicted in data mining namely Cancer, Diabetes, Thyroid disease etc. This paper analyzes the Heart disease.

The rest of this paper is organized as follows. The literature survey is described in section 2. Section 3 describes the prediction of disease by kmeans and results are then optimized by PSO. Section 4 describes results and comparison. Conclusions and References are described in Section 5.

Literature Survey

Jyoti Soni, Ujma Ansari & Dipesh Sharma performed a work," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" [1]. This paper presents the results of a data mining techniques that are applied on the same dataset

that reveals that Decision Tree performed too well and some time Bayesian classification had similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks not performed well. The accuracy of the Decision Tree and Bayesian Classification improved after applying genetic algorithm.

R. Chitra," Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques"[2]. This paper shows that in data mining Neural Networks are one of the analytical tools that can be utilized to make predictions for medical data and Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system.

K.Srinivas, G.Raghavendra Rao and A.Govardhan," Analysis of Attribute Association in Heart Disease Using Data Mining Techniques".[3]. In this paper, a new measure was proposed that finds the association among the various attributes in a dataset. The experiments are conducted on synthetic and real data sets. The measure is applied to both frequent and infrequent item set to the dataset and it was found that the infrequent item set are also having the association among the attributes.

Shweta Kharya," Using data mining techniques for diagnosis and prognosis of cancer disease".[4]. In this paper various data mining approaches that have been utilized for breast cancer diagnosis and prognosis are discussed.

S.Vijiyarani & S.Sudha, "Disease Prediction in Data Mining Technique – A Survey"[5]. This research paper analyzes how data mining techniques are used for predicting different types of diseases i.e heart disease, Diabetes and Breast cancer using Naïve bayes, K-NN, Decision List..

V. Manikantan & S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods"[6]. MAFIA and C4.5 algorithms are used with K-Means clustering algorithm for heart disease prediction and remove the data from the database that is applicable to heart attack.

Mai Shouman, et al. [12] worked on k-means clustering with the decision tree to predict the heart disease. To increase the efficiency they used several centroid selection methods. Data set of heart disease was the collection of 13 input attributes that were taken from Cleveland Clinic Foundation. In diagnosis of heart disease patients k-means clustering and decision tree both together achieve higher accuracy.

Our Approach

We have implemented K-means and PSO on the text database. Database contains some records of patients of heart disease. We took the input from the user and diagnose the disease based on some parameters. The result will show the clusters that contain number of the records and detect whether the patient is at risk of heart disease or not.

Data Set:

Table 1.1 Data Set

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
2	double	String	String	double	double	String	String	double	String	double	String	double	String	String
3	63	male	typ_angin	145	233	t	left_vent	150	no	2.3	down	0	fixed_def	<50
4	67	male	asympt	160	286	f	left_vent	108	yes	1.5	flat	3	normal	>50_1
5	67	male	asympt	120	229	f	left_vent	129	yes	2.6	flat	2	reversibl	>50_1
6	37	male	non_angir	130	250	f	normal	187	no	3.5	down	0	normal	<50
7	41	female	atyp_angi	130	204	f	left_vent	172	no	1.4	up	0	normal	<50
8	56	male	atyp_angi	120	236	f	normal	178	no	0.8	up	0	normal	<50
9	62	female	asympt	140	268	f	left_vent	160	no	3.6	down	2	normal	>50_1
10	57	female	asympt	120	354	f	normal	163	yes	0.6	up	0	normal	<50
11	63	male	asympt	130	254	f	left_vent	147	no	1.4	flat	1	reversibl	>50_1
12	53	male	asympt	140	203	t	left_vent	155	yes	3.1	down	0	reversibl	>50_1
13	57	male	asympt	140	192	f	normal	148	no	0.4	flat	0	fixed_def	<50
14	56	female	atyp_angi	140	294	f	left_vent	153	no	1.3	flat	0	normal	<50
15	56	male	non_angir	130	256	t	left_vent	142	yes	0.6	flat	1	fixed_def	>50_1
16	44	male	atyp_angi	120	263	f	normal	173	no	0	up	0	reversibl	<50
17	52	male	non_angir	172	199	t	normal	162	no	0.5	up	0	reversibl	<50
18	57	male	non_angir	150	168	f	normal	174	no	1.6	up	0	normal	<50
19	48	male	atyp_angi	110	229	f	normal	168	no	1	down	0	reversibl	>50_1
20	54	male	asympt	140	239	f	normal	160	no	1.2	up	0	normal	<50
21	48	female	non_angir	130	275	f	normal	139	no	0.2	up	0	normal	<50
22	49	male	atyp_angi	130	266	f	normal	171	no	0.6	up	0	normal	<50
23	64	male	typ_angin	110	211	f	left_vent	144	yes	1.8	flat	0	normal	<50
24	58	female	typ_angin	150	283	t	left_vent	162	no	1	up	0	normal	<50

K-Means Approach: This technique is a popular traditional partitioning clustering technique. It is one of the easy approaches used for resolving known clustering issues. For document clustering problem, this approach allocates each of the document to one of the K number of clusters. An efficient cluster here will be a sphere where centroid is considered to be its centre of gravity. The aim of this approach is to reduce the approximate mean coverage of document data set corresponding to their cluster midpoint; when centre of the group can be used as the mean of the document set in a group. The centroid $\mu 1$ of the document set in a cluster ω is calculated as mentioned below:

$$\mu 1\ (\omega) = \frac{1}{|\omega|} \sum_{x \in \omega} x$$

P.S.O

This is a populace build searching tool that was first proposed by Eberhart & Kennedy in 1995. Eberhart & Kennedy originally developed this method for optimization of continuous non-linear functions. It is a stochastic tool for optimization that can be used

easily to resolve various optimization issues. For using this approach profitably, an important factor knows how the solution is mapped into the P.S.O object that directly alters its utility and attainment.

A '*swarm*' implies to a grouping of a number of optimal solutions where each optimal result is known as a '*particle*'.

Here, individual objects are stemmed with random positions & velocities then a *fitness function* is calculated. The aim of this approach is to know the particle's position that gives the best results of this function using positional coordinates of objects as input values. The best position of the particle can be reset as follows:

$$p_{id}(a+1)= \begin{cases} p_{id}(t) & if \quad f(x_i(t+1)) \geq f(p_{id}(t)) \\ x_i(t+1) & if \quad f(x_i(t+1)) < f(p_{id}(t)) \end{cases}$$

The best global position is chooses as best of personal best positions as follows:

$$p_{gd}(t) = min\ \{pid0\ (t),\ldots\ldots,p_{idk}(t)\}$$

After finding the above values, each particle updates its position and velocity as per give equations:

$$v_{id} = w*v_{id}+c_{11}*rand_{11}*(p_{id} - x_{id})+c_{12}*rand_{12}*(p_{gd} - x_{id})$$

$$x_{id} = x_{id} + v_{id}$$

Here, p_{id} denotes objects personal experience, p_{gd} is the global experience, $rand_{11}$ and $rand_{12}$ denotes random constants lying in range (0,1) for wide search space exploration, c_{11} and c_{12} denotes constants generally taken as 2 , w is the inertia weight lying in range (0.1,0.9) to control P.S.O convergence.

The PSO generates a global optimum solution. Its disadvantage includes stagnation of population due to loss of population diversity as a result of which PSO might take time to converge to an optimum solution and non-detection of outliers. K-Means has been integrated with PSO to make it work faster and reduce the clustering time required by PSO alone. This hybridization exploits the fast convergence of K-Means and global searching ability of PSO.

Results

The following form is to be filled by the user and the above mentioned techniques are used to detect whether the patient is at risk of heart disease or not.

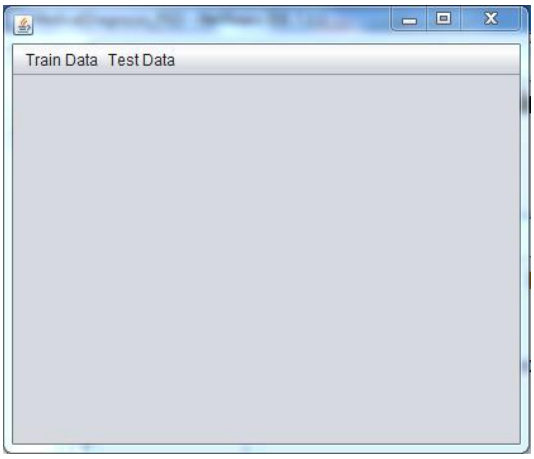

Fig 2 Train data & Test Data

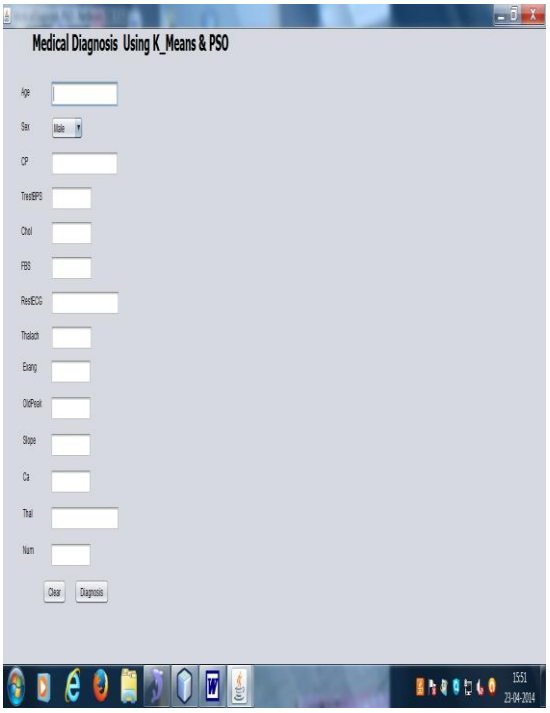

Fig 3 Medical Diagnosis using K_means & PSO

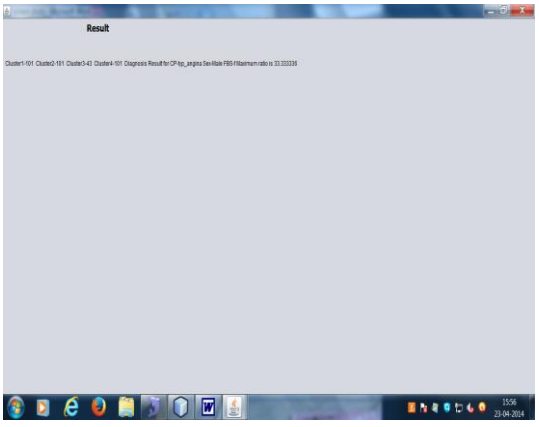

Fig 5 Result

The result is showing the four clusters that contains the number of the records of different age group.Cluster1 is showing the number of the record of age group 30-40, Cluster2 40-50, Cluster3 50-60 and cluster4 60-70. If the person is at risk of heart disease then the Fasting Blood Sugar(Fbs) is showing true otherwise false.

CONCLUSION AND FUTURE WORK

In this present study, we have concluded that the existing K-Means clustering algorithm for the Document clustering is not much effective alone and was improved using PSO based approach. Clustering technique alone is not able to produce the better clusters but if it is optimized then results are pretty good. In future work, we can use these techniques for treatment of the heart disease.

REFERENCES

[1] Jyoti Soni, Ujma Ansari & Dipesh Sharma" Predictive Data Mining for Medical Diagnosis", International Journal of Computer Applications (0975 – 8887),Volume 17– No.8, 2011.

[2] R. Chitra," Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", ICTACT Journal on Soft Computing, Volume: 03, ISSUE: 04, ISSN: 2229-6956, 2013

[3] K.Srinivas, G.Raghavendra Rao and A.Govardhan, " Analysis of Attribute Association in Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,Vol. 2, pp.1680-1683 (2012)

[4] Shweta Kharya," Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, 2012

[5] S.Vijiyaran & S.Sudha , "Disease Prediction in Data Mining Technique – A Survey" International Journal of Computer Applications & Information Technology, ISSN: 2278-7720, 2013

[6] V. Manikantan & S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering (IJACTE), 2319 – 2526, Volume-2, Issue-2, 2013

[9] Joshua Z. Huang, Michael K. Ng, H. Rong, and Z. Li. (2005). Automated dimension weighting in k-means type clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence,27(5), 1–12.

[10] Mannila, H.: Methods and Problems in Data Mining. In: The International Conference on Database Theory, pp. 41–55 (1997)

[11] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[12] Mai Shouman, Tim Turner and Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", Proceedings of the International Conference on Data Mining, 2012.