International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

Data Mining and Clustering of Medical Data from PTB Database Using MATLAB Analysis

Priyanka Sharma[1], Ajay Abrol[2], Sameru Sharma[3]

[1]priyankasharma1808@gmail.com, [2]ajayabrol17569@rediffmail.com, [3]sameru33@gmail.com
Department of Electronics and Communication Engineering Government College of Engineering and Technology, Jammu

Abstract: Medical data base is history of patient recorded in terms of past history and investigation data collected from various tests and reports. Authors collect PTB data base from PhysioNet site and consider ten parameters in order to study and classify data into various clusters. The number of clusters is determined by the difference among the variables and nature of information contained in the parameters. In this study nine parameters are considered from the medical database and the results show the clustering using single linkage procedure.

Keywords: Pattern, Cluster, Single-linkage, Euclidean distance

I. Introduction

Patterns are generated by considering a combination of different attributes. Depending upon the information in the attributes the patterns can be grouped. The grouping of the patterns is clustering. Within a cluster the patterns are similar and resemble with one another. Number of clusters and merging of the cluster can be performed depending upon the nature and type of data set. The Knowledge about the features including statistical or structural feature information is needed to be utilized. The separteability among the patterns is considered superior in non linear transformation compared to linear transformation [3].

Feature selection in pattern recognition refers to identifying effective subset of the original characteristic features that are used in clustering. Features extraction transformations input features to a new salient features [2]. During feature selection a subset of the feature are selected. The features that are filtered are not used in computation. During feature extraction, all of the features are transformed to new dimension and all are used. The new dimensional space is a reduced dimensional space [4].

Data mining is a field that discovers group determines interesting distributions and finds patterns in a given data. Clustering is a process of partitioning data into clusters. Data in clusters are similar to each other compared to data in different clusters [5].

In this work author has collected data from PhysioNet [6][7] as given in table 1.1. Data belongs to a particular illness with the symptoms projected in the form of data set. The data is compiled and subjected to single linkage Cluster approach. Recent research found that supervision of a certain amount in clustering algorithms improves the accuracy. Pair wise and instant level constraints, is a supervision type that is used in clustering applications, “Huang Anna, Milne David Frank Eibe, H.Witten Ian”. Clustering Documents with information is givens as in table 1.1.

II. Method

It is explained that patterns can also be put into groups based on the values of their attributes. Such groups are called clusters, and the process of forming clusters is called clustering. The process has often been found to be useful in the exploratory stages of researching a domain to learn how

Table 1.1: Data of five patients obtained from PhysioNet

Patient No.	Age	Sex	MI	Pulmonary System	Peripheral	Blood Pressure	Cardiac Output	RIVA	RCX	RCA Peripheral
S0263	75	M	✓	110	70	N/A	N/A	NO	100%	70%
S0265	48	M	✓	120	70	11	9,35	50%	NO	100%
S0290	45	M	✓	120	90	10	11,4	70%	NO	NO
S0326	66	M	✓	120	70	N/A	N/A	50%	90%	100%
S0339	54	M	✓	100	70	N/A	N/A	75%	NO	70%

Priyanka Sharma, Ajay Abrol, and Sameru Sharma, “Data Mining and Clustering of Medical Data from PTB Database Using MATLAB Analysis,” *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 73-76, 2019.

patterns in domain can be clustered. Clustering is also known as 'unsupervised learning', that is, there is no training set available to supervise, or guide, the learning [1]. In this paper a study of five patterns X1 to X5 and the values of their numeric attribute A1 and A2 in table 1.2 is explained in detail. It is not necessary to know what the attribute means; only their values need to be considered. Intuitively put the five patterns into two clusters such that the patterns within a cluster are more similar to one another than to patterns in another cluster.

Table 1.2: Set of five patterns to be clustered.

Patient No.	Patterns	Age (A1)	Pulmonary System (A2)
S0263	X1	75	110
S0265	X2	48	120
S0290	X3	45	120
S0326	X4	66	120
S0339	X5	54	100

The steps followed in this research for clustering are as follows:

1. Choose a value number of clusters to be made.

2. Consider each of the patterns to be in separate cluster, that is, each cluster contains different pattern.

3. Calculate the pair wise distance between clusters and merge two similar clusters into one cluster, thus reducing the number of clusters to one. If there is more than one pair of clusters that are closest to each other, then arbitrarily select a pair to merge. This measures the distance between clusters.

Before finding the distance between the clusters, it is required to calculate the distance between the patterns. This is done by the Euclidean distance, which is most often used to measure the distance. The calculations for Euclidean distance between patterns in table 1.2 are explained as follows:

Distance between pattern X1 and X2 is $\sqrt{(48-75)^2+(120-110)^2} = 28.79$. The diagonal values are zero because the distance between a pattern and itself is zero. The table will be symmetric because, for example, the distance between X1 and X2 is equal to the distance between X2 and X1. Therefore, the distance between X2 and X1 is 28.79. Similarly, the distance between X1 and X3 is $\sqrt{(45-75)^2+(120-110)^2} = 31.62$ which is same for distance between X3 and X1. Calculating all the distances between the following:

- X1 and X4 is $\sqrt{(66-75)^2+(120-110)^2} = 13.45$ same between X4 and X1.
- X1 and X5 is $\sqrt{(54-75)^2+(100-110)^2} = 23.25$ same between X5 and X1.
- X2 and X3 is $\sqrt{(45-48)^2+(120-120)^2} = 3$ same between X3 and X1.
- X2 and X4 is $\sqrt{(66-48)^2+(120-120)^2} = 18$ same between X4 and X1.
- X2 and X5 is $\sqrt{(54-48)^2+(100-120)^2} = 20.88$ same between X5 and X2
- X3 and X4 is $\sqrt{(66-45)^2+(120-120)^2} = 21$ same between X4 and X3
- X4 and X5 is $\sqrt{(54-66)^2+(100-120)^2} = 23.32$ same between X5 and X4
- X3 and X5 is $\sqrt{(54-45)^2+(100-120)^2} = 21.93$ same between X5 and X3

Hence the matrix for patterns can be expressed as follows:

Table 1.3: Euclidean distances between the five patterns in table 1.2

Pattern	X1	X2	X3	X4	X5
X1	0	28.79	31.62	13.45	23.25
X2	28.79	0	3	18	20.88
X3	31.62	3	0	21	21.93
X4	13.45	18	21	0	23.32
X5	23.25	20.88	21.93	23.32	0

Now to create two clusters of the five patterns X1 to X5 by single-linkage procedure, considering the off-diagonal values in table 1.3 it is seen that the minimum distance between any two clusters is 3 which is between X2 and X3, which can be merged into one cluster {X2,X3}. Then the distance between four clusters is given as:

Table 1.4: First Cluster formation

Cluster	X1	X2,X3	X4	X5
X1	0	28.79	13.45	23.25
X2,X3	28.79	0	18	20.88
X4	13.25	18	0	3.32
X5	23.25	20.88	23.32	0

Since, we are using single-linkage procedure, the distance between{X2, X3} and say {X1} is minimum(28.79, 31.62) = 28.79. Similarly for next set of cluster formation, the distance

Priyanka Sharma, Ajay Abrol, and Sameru Sharma, "Data Mining and Clustering of Medical Data from PTB Database Using MATLAB Analysis," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 73-76, 2019.

between{X1, X4} and say {X5} is minimum (23.25, 23.32) = 23.25, therefore, the matrix is given as follows:

Table1.5: Second cluster formation

Cluster	X2,X3	X1,X4	X5
X2,X3	0	18	20.88
X1,X4	18	0	23.25
X5	20.88	23.25	0

Table 1.6: Third cluster formation

Cluster	X1,X2,X3,X4	X5
X1,X2,X3,X4	0	20.88
X5	20.88	0

The distance between{X1, X2, X3, X4} and {X5} is minimum (23.25, 20.88, 21.93, 23.32) = 20.88. Hence, the final group of clusters formed are{X1, X2, X3, X4} {X5}.

III. Results

It is seen that the clustering method using single-linkage procedure is successfully applicable in forming different cluster of patterns with similar attributes. Here, the clusters are formed in relation with age and the pulmonary systems of different patients and different types of clustering can be observed taking different set of patients. Similarly, other attributes can also be used to group patients into different clusters. The results of multidimensional data show the following results.

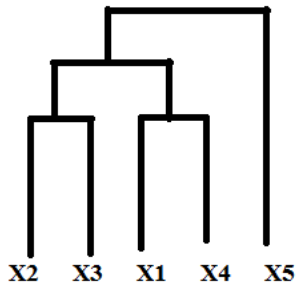

Fig. 1.1 Schematic diagrams of cluster formation for patients in Table 1.1 and Table 1.7

Figure 1.1 shows the schematic diagrams of cluster formation for patients in Table 1.1 and Table 1.7. the results using MATLAB analysis are obtained and are as shown in figure 1.2.

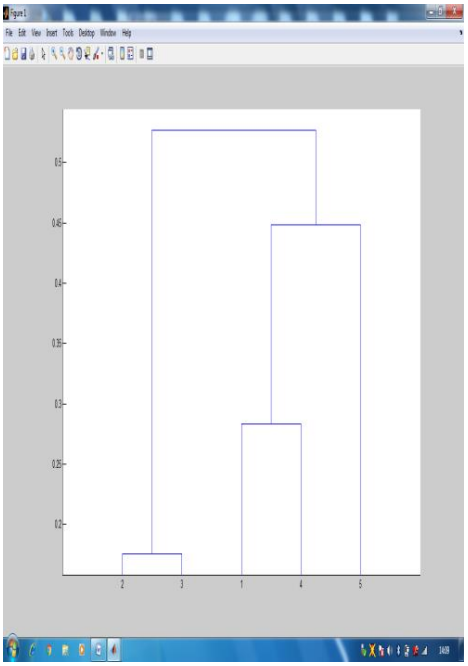

Fig. 1.2 Results as obtained using MATLAB

IV. References

[1] Rajjan Shinghal, " Pattern Recognition Techniques and Application", published by Oxford Universiry Press, 1st edition, pp 223-239.

[2] Jaink. A, Murty N.M., Plymn J.P., "Data clustering: A Review.

[3] PU King-Sun, Rosenfeldazriel, "Pattern Recognition and Image Processing" IEEE Transactions on Computers Vol. C-25, NU12, December, 1976

[4] Fuka Karel, Hanka Rudoff, "Feature Set Reduction for Document Classification Problems"].

[5] Halkidi Maria, Batistakis Yannis, Vazirgiannis Michalis, "ON clustering Validation Techniques" Journal of Intelligent Information System 17, 2/3, PP. 107-145, 2001, Kluwer Academic Publishers Manufactured in the Netherlands].

[6] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

[7] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic

Priyanka Sharma, Ajay Abrol, and Sameru Sharma, "Data Mining and Clustering of Medical Data from PTB Database Using MATLAB Analysis," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 73-76, 2019.

[8] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, published by Morgan Kauffman, 3rd edition.

[9] Mrs. Bharati M. Ramageri, "Data Min-Ing Techniques And Applications" ,Indian Journal of Computer Science and Engineering Vol. 1 No. 4, ISSN : 0976-5166 pg: 301-305.

[10] Ke Jie, Dong Hongbin, Tan Chengyu and Liang Yiwen, "PBWA: A Provenance-Based What-If Analysis Approach for Data Mining Processes" Chinese Journal of Electronics Vol.26, No.5, Sept. 2017

[11] LiHua Wang BeiHang Zijun Zhou, "Congestion Prediction for Urban Areas by Spatiotemporal Data Mining", International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery 978-1-5386-2209-4/17 2017 IEEE

[12] Sagardeep Roy Anchal Garg," Analyzing Performance of Students by Using Data Mining Techniques A Literature Survey" 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, Oct 26-28, 2017, 978-1-5386-3004-4/17

Priyanka Sharma, Ajay Abrol, and Sameru Sharma, "Data Mining and Clustering of Medical Data from PTB Database Using MATLAB Analysis," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 73-76, 2019.