Dengue Disease Detection using K- Means, Hierarchical, Kohonen- SOM Clustering

P.Yogapriya, P.Geetha

Abstract: ***Data Mining is the process of extracting useful information. Data Mining is about finding new information from pre-existing databases. It is the procedure of mining facts from data and deals with the kind of patterns that can be mined. Therefore, this proposed work is to detect and categorize the illness of people who are affected by Dengue through Data Mining techniques mainly as the Clustering method. Clustering is the method of finding related groups of data in a dataset and used to split the related data into a group of sub-classes. So, in this research work clustering method is used to categorize the age group of people those who are affected by mosquito-borne viral infection using K-Means and Hierarchical Clustering algorithm and Kohonen-SOM algorithm has been implemented in Tanagra tool. The scientists use the data mining algorithm for preventing and defending different diseases like Dengue disease. This paper helps to apply the algorithm for clustering of Dengue fever in Tanagra tool to detect the best results from those algorithms.***

IndexTerms**:*****Clustering,DataMining,Dengue,Hierarchical,K-Means,Kohonen-SOM,Tanagra***

I. INTRODUCTION

Data mining is the procedure of extracting valuable information it is also known as (KDD). Data mining is a fundamental concept for creating complex information. In various cases, information is stored, so it can be used later. The data mining is the process of retrieving hidden information from the large sets of data. Data mining is a process of storing gigantic amounts of data stored in the database, data warehouse, or other information repositories.

Dengue is also known as viral fever. Dengue is mosquito-borne viral infectivity which is wide and valued by Aedes girls mosquitoes. Dengue has to rotate into a rigorous physical condition problem occurs repeatedly in the moist and sub-tropical region.

The cluster is a collected work of the data items that are related to one another surrounded by the same cluster and are not related to the items in previous clusters.

Clustering analysis has been a well-known problem in data mining due to its heterogeneity of applications. Hierarchical clustering methods can be classified as either agglomerative or divisive, K-Means clustering algorithm is an unsupervised learning algorithm.

K-Means need not require training data and it cannot work with existing classified or marked data. K term refers to a number of the group. K-Means is also known as the nearest centroid classifier or Rocchio algorithm. Kohonen-SOM algorithm is an unsupervised learning algorithm which leads to the idea of the neighborhood of the clustering unit. During the self-organizing process, the load vectors of the attractive unit and its neighbors are modernized. Those three algorithms are used to compare the results of Dengue affected people by their age groups.

Revised Manuscript Received on August 02, 2019.

P.Yogapriya, Department of Computer Science, Alagappa University/ Dr.Umayal Ramanathan College for Women/ Karaikudi, India.

Dr. P.Geetha, Department of Computer Science, Alagappa University/ Dr.Umayal Ramanathan College for Women/ Karaikudi, India.

II. RELATED WORK

Ritu Chauhan et al expose data analytical tools and data mining techniques to analyze the medical data using hierarchical clustering algorithm [1].

P.Manivannan, P. Isakki et al obtainable a manuscript has been projected four stages to predicting the dengue fever. R 3.3.2 tool is used for preparing and collecting the dengue data set and calculate the result produced from those abovementioned algorithms [2].

Sahanaa C et al calculate the situation of being unhealthy and the Condition of being to passing away of dengue for a stage of five years. The information was sourced from the National Health Profile 2017, review of inferior data was done, The chart was devised to learn the method of the virus which has been implemented [3].

Shoukat et al are accessible to analyzes the bother of dengue fever in the district Jhelum, Pakistan. Dataset obtained from Executive District Officer(EDO) in Jhelum. K-Means, K-Medoids, DBSCAN, Optics are used. Distinguish the result construct from individual algorithms [4].

P.Sathya, A.Sumathi et al planned toward a forecast of dengue illness using clustering techniques. The dataset was gathered from Lotus and 24care hospitals. Approximate the performance of all the techniques separately based on figures and charts be forced upon the dataset [5].

Rao, K., K., N., et al anticipated classification rules using a decision tree. The main objective is creating a prediction model for predicting the probability of occurrences of dengue disease. The decision tree classification model achieved a 97% accuracy [6].

III. METHODOLOGY

EXISTING METHODOLOGY

Many alive, developed, examine dataset contains absent values. They are introduced due to a variety of reasons, such as physical data access, tools errors and inaccurate quantifications. Finding of defective data is easy in most cases, looking for unfounded values in a dataset.

Three types of harms generally associated with missing values such as loss of effectiveness, difficulty in the organization and analyzing the data, excess resulting from differences among lost and full data. In existing the K-Means clustering algorithm and Hierarchical clustering algorithm with Tanagra using some different dataset such as a Car, Bank, Census dataset. In that dataset, It can compare cluster K-means1 and cluster K-Means 2 with discrete and continuous value and then cluster HAC1 and cluster HAC2 for finding mean standard deviation and recall accuracy for that dataset.

PROPOSED METHODOLOGY

Various data mining techniques exist for predicting the severity of the Dengue fever in patients at an early stage. Even then there exist cases, where the predictive results have to be improved to predict the attack of dengue fever. In order to carry out the task, the proposed a new technique. The technique compares the results of three different clustering techniques. K-Means clustering, Hierarchical clustering, and Kohonen – SOM clustering technique. The steps involved in the process are as follows,

(i)Data collection

(ii)Data Preparation

The first step deals with data collection using a dataset from humid and sub-tropical areas. The input data have been collected from metropolitan areas. The second step deals with data preparation using dengue sufferer of data possessed from domestic clustering of data. In data preparation method the attributes of the dataset which contains EPID, Fever, Bleeding, Myalgia, Flu, Fatigue, Results, Age, IgM G test1, IgM G test2, IgM G test 3, Max_fever rate, Temperature, serotype, Days of illness to identify the type of fever. To construct the K-Means clustering algorithm for initializing K clusters into N partitions of the dataset then construct the Hierarchical clustering algorithm for which the data are grouped together in the form of trees and then compare Kohonen- SOM algorithm which leads to the idea of the neighborhood of the clustering unit. Therefore, to compare this algorithm for finding the best results from those algorithms. To overcome the existing method using Tanagra tool to compare the algorithm to detect the attack of dengue fever in the form of the accuracy of a dataset and categorization of the age group.

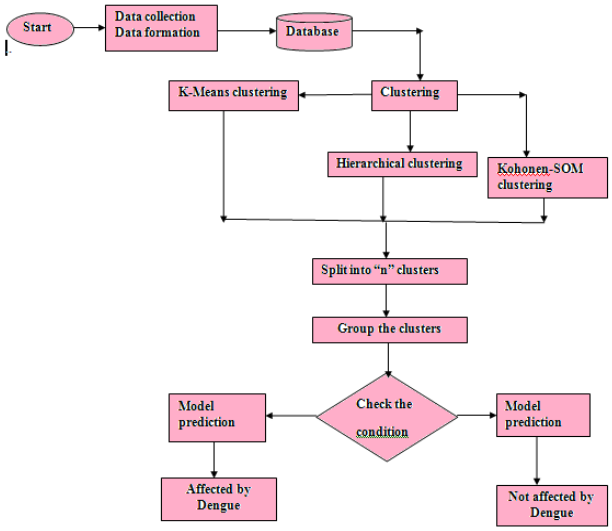

Figure.1 Data Flow Diagram for proposed Methodology

DATASET CONSTRUCTION

(a)TOOL DESCRIPTION

Tanagra tool was written as an aid to learning and investigate data mining urbanized by Ricco Rakotomalala. In the Tanagra tool, the data can be imported and exported in the text files. Tanagra tool which includes the following analysis such as Association Rule Mining, Clustering, and Classification techniques. Tanagra is an open-source software intended mostly for research use. In Tanagra software, the dataset can be imported into different data formats such as text files, ARFF files, CSV files, as well as Microsoft Excel, IBM Visual Warehouse, and Oracle Express formats. In this analysis of Dengue fever, the dataset can be imported into Microsoft Excel file formats.

(b)DATASET DESCRIPTION

In the dataset description, Dengue dataset contains 15 attributes and 115 examples. The attribute includes EPID, Fever, Bleeding, Myalgia, Flu, Fatigue, Results, Age, IgM G test1, IgM G test2, IgM G test 3, Max_fever rate, Temperature, serotype, Days of illness. The attributes categorized into two types of values such as Discrete value, Continue value. In this Dengue dataset, 15 attributes can be classified into Discrete values and the remaining one as continue value.

Table.1 Dataset description

Dataset description

15 attribute(s)
115 example(s)

Attribute	Category	Informations
EPID	Continue	-
Fever	Discrete	3 values
Bleeding	Discrete	2 values
Myalgia	Discrete	2 values
Flu	Discrete	2 values
Fatigue	Discrete	2 values
Results	Discrete	2 values
Age	Continue	-
Igm G test1	Discrete	3 values
Igm G test2	Discrete	3 values
Igm G test3	Discrete	4 values
max_fever_rate	Continue	-
Serotype	Continue	-
Temperature	Discrete	12 values
Days of illness	Continue	-

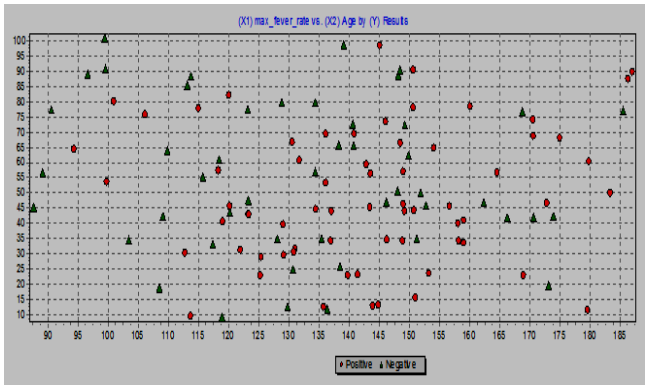

Figure.2 Scatterplot for Kohonen-SOM clustering

In the above Figure, 2 represent the scatterplot for Kohonen – SOM clustering which leads to the idea of the neighborhood of the clustering unit. Axis(X1)Max fever rate versus Axis (X2) Age by Axis(Y) Results where rounded shape indicates positive results of Dengue and Triangle shape indicates Negative results of Dengue.

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6
EPID	75.500000	56.857143	33.368421	66.343750	99.500000	20.277778
Age	47.722222	75.857143	29.578947	38.968750	60.642857	76.222222
max_fever_rate	169.555556	140.500000	126.684211	138.312500	130.785714	126.055556
Serotype	1.500000	0.714286	1.105263	3.531250	2.214286	2.111111
Days of illness	5.333333	4.000000	8.210526	5.000000	11.214286	8.500000

Figure.3 Cluster Centroid results for Kohonen- SOM

In figure.3 represent the clustering centroid results for Kohonen- SOM clustering in that it can be partitioned into 10 clusters which predicts the results of neighboring cluster units.

A.

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6
EPID	104.909091	49.055556	23.157895	63.210526	28.875000	74.285714
Age	63.909091	55.944444	74.473684	36.526316	46.250000	50.857143
max_fever_rate	130.272727	120.666667	143.157895	137.473684	108.750000	168.238095
Serotype	2.363636	0.222222	2.578947	3.236842	1.750000	1.142857
Days of illness	12.181818	5.333333	7.368421	5.263158	12.875000	5.000000

Figure.4 Result of K-Means

In the above Figure. 4 represent the result of clustering K-Means. Initialize the "K" partition into "n" clusters. In this above figure initialize "K" partition into 10 clusters to intimate the cluster centroids with point values.

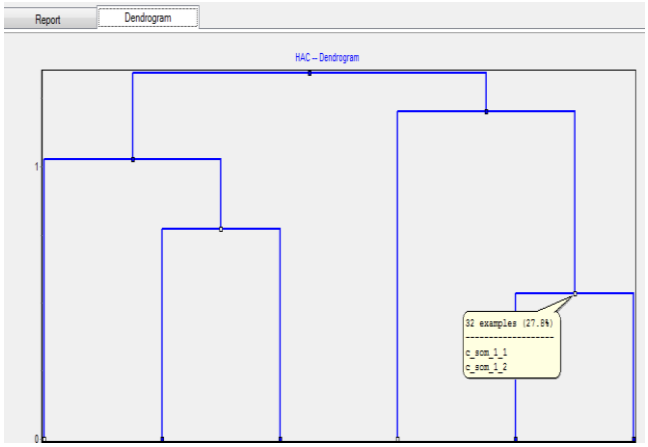

Figure.5 Dendrogram for hierarchical clustering

In the above Figure. 5 represent the Dendrogram for Hierarchical clustering in this clustering can define the parameters for cluster selection with 10 iterations. The white box intimates the cluster_som which is used to grouping the neighborhood clustering unit.

IV. RESULT AND DISCUSSION

Dengue is a break borne viral disease. To detect the Dengue enthusiasm with Dengue-related dataset using K-Means clustering algorithm, Hierarchical clustering algorithm, and Kohonen – SOM clustering. In this comparison of the K-Means clustering algorithm, Hierarchical clustering algorithm, Kohonen – SOM clustering to deduct the best results from those algorithms. To overcome the existing method using Tanagra tool to compare the algorithm to detect the attack of dengue fever in the form of the accuracy of a dataset with their categorization of age group. In these above-mentioned algorithms is used to partition the clusters and to specify the dendrogram for the clusters and then finally grouped the clustering with neighborhood units.

Table. 2 Comparative Results of three different Clustering algorithm

Algorithm Used	Computation Time Taken(ms)	Cluster Centroid for Age groups		Max Fever rate	Best Cluster Selection	Accuracy in %
		Cluster1	Cluster2			
K-Means	0ms	63.90909	55.94444	130.27272	0.5239	61.9%
Kohonen-SOM	0ms	47.72222	75.85714	169.55556	0.4689	65.7%
Hierarchical clustering	15ms	38.96875	60.03125	138.31258	0.3567	49.5%

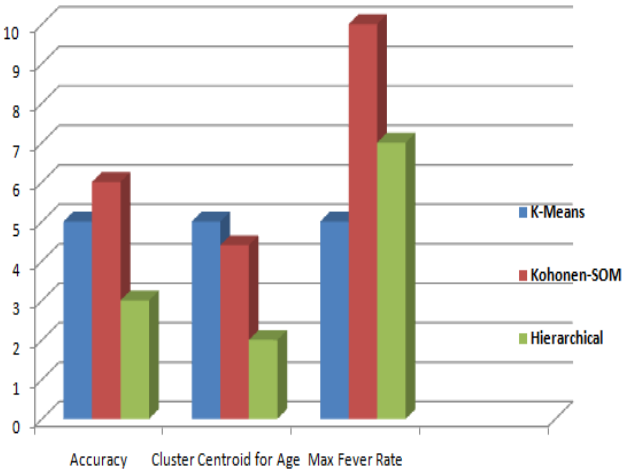

Figure.5 Bar Graph comparison

V. CONCLUSION

Investigative data mining covers a minor road area of research. The various steps involved in the proposed method were Data collection, Data formation, Data preparation. After preparation, clustering of the given dataset was carried out with three different techniques, K-Means clustering algorithm, Hierarchical clustering algorithm, Kohonen –SOM clustering. The result obtained from the three methods was compared based on recall accuracy and age groups of clustering centroids. From the results, it is well evident that Kohonen – SOM clustering and K-Means clustering provided the best results in deducting the data. This analysis mainly focuses on discrete and continuous values of Dengue databases on which clustering algorithms are applied. The clusters of random shapes are formed if the data is constant in nature. As future work, the authors proposed to improve the method by including more number of datasets and to use DBSCAN clustering techniques to predict Dengue fever.

REFERENCES

1. Ritu Chauhan et al., " Data Clustering method for Discovering Clusters in spatial cancer Databases", International Journal of Computer Applications, volume.10, Special Issues. 6, November 2010.
2. P.Manivannan, P. Isakki Devi., "Dengue fever prediction using K-Medoid Clustering Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, volume.5, Special Issue.1, March 2017.
3. Sahanaa C.*, Amit Kumar Mishra, Joy Bazroy., "Trend of Morbidity and Mortality of Dengue in Tamil Nadu and Puducherry, South India", International Journal of Community Medicine and Public Health, volume.5, Special Issue.1, January 2018.
4. Kamran Shaukat et al., "Dengue Fever in Perspective of Clustering Algorithms", International Journal of Data Mining in Genomics and Proteomics, volume.6, Special Issue.3, 2015.
5. P.Sathya, Mrs.A.Sumathi.," Predicting Dengue Fever Using Data Mining Techniques", International Journal of Computer Science and Technology, volume.6, Special Issue.2, March-April 2018.
6. Rao, K, K, N., Varma, S, P, G., and Rao, V, N., "Classification Rules using Decision Tree for Dengue Disease", International Journal of Research in Computer and Communication Technology, volume.3, Special Issue.3, March 2014.
7. S.Muthukumaran, E.Ramaraj., "A Multilayered Backpropagation Algorithm to Predict Significant Attributes of UG Pursuing Students Absenteeism at Rural Educational Institution", International Journal of Computer Science and Engineering, volume.6, Issue.3, December 2018.
8. P.Manivannan, P.Isakki Devi., "Dengue Fever Prediction using K-Means Clustering Algorithm", International Conference Techniques in Control, Optimization and Signal Processing, volume.5, Special Issue.1, March 2017.
9. KR.Sivabalan, E.Ramaraj., "Remote Sensing Satellites and its agricultural development technical aspects", International Journal of Computer Science and Engineering, volume.6, Special Issue.9, September 2018.
10. Mohammed Shahadat Hossain, Ishrat Binteh Habib, Karl Andersson, "A Belief Rule-Based Expert System to Diagnose Dengue Fever under Uncertainty", IEEE Computing Conference, volume.1, Special Issue.5, July 2017.
11. Shamimul Hasan, Sami Faisal Jamdar et al., "Dengue Virus: A global human threat: Review of Literature", Journal of International Society of Preventive and Community Dentistry, volume.6, Special Issue.1, Jan-Feb 2016.
12. N.K. Kameswara Rao, Dr. G.P.Saradhi Varma, Dr. M. Nagabhushana, "Classification rules using Decision Tree for Dengue Disease", International Journal of Research in Computer and Communication Technology, volume.3, Issue.3, March 2014.
13. R.Karthikeyan, Dr. P. Geetha, Dr. E. Ramaraj, " Rule-Based System for Better Prediction Of Diabetes", IEEE Third International Technology Conference on Computing and Communication, volume.1, Issue.1, 2019.

AUTHORS PROFILE

P.Yogapriya is currently pursuing the M.Phil Degree in computer science at the Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi. Her research interest includes Dengue Disease and Data mining.

P.Geetha is working as the Associate Professor in the Department of Computer Science, Dr. Umayal Ramanathan College For Women, Karaikudi. She has the sound knowledge in many research fields especially in Data mining, Big Data, and Analytics. She has published 11 International conferences and 17 international journals. She has 13+ years of experience in teaching.