ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Dengue Fever Prediction using K-Medoid Clustering Algorithm

P.Manivannan, Dr. P. Isakki @ Devi

Research Scholar, Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

Assistant Professor, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

ABSTRACT: Dengue is a threatening disease which is caused by female mosquitoes. It is typically found in hot regions. The dengue diseases occur in 4 serotypes (DENV-1, DENV-2, DENV-3 and DENV-4). A dengue disease ranges from mild febrile disease to severe hemorrhagic fever. Predicting the relationship between dengue serotypes and humans age will definitely help the biotechnologists and bioinformaticians to move one step forward to discover medicines for dengue. This paper has been proposed four stages namely preprocessing, attribute selection, k-medoid clustering and predicting the dengue fever R 3.3.2 tool is used for preprocessing the household of dengue dataset. D win's method has been applied to generate filled dataset by replacing all missing values for nominal and numeric attributes with mode and mean value. Various data mining techniques have been used for predicting dengue virus. The main goal of research work is to predict the people who are affected by dengue depending upon categorization of age using the K-medoid clustering algorithm, which has been implemented.

KEYWORDS: Dengue, Data Mining, Medical Documents, Clustering techniques, K-medoid clustering Algorithm.

I. INTRODUCTION

Dengue fever is divided into two types, i.e., classical dengue fever and dengue hemorrhagic fever, according to world health organization. DHF1, DHF2, DHF3 and DH4 are further four types of dengue hemorrhagic fever. Symptoms of dengue include severe joints pain, headache, rashes, thrombocytopenia, leucopenia and muscle ache. Due to the symptoms of muscle ache and joint pains, this disease is also called the break bone fever. In a very small portion of cases, the disease might develop further into life-threatening dengue hemorrhagic fever (DHF) which results in reduce the number of blood platelets, blood plasma leakage and may result in dengue shock syndrome (DSS) in which the blood pressure might drop to dangerously low levels.

Data mining is the complication data analysis tools to discover valid patterns and relationships in a large data set. These tools can include mathematical algorithms arithmetical models and machine learning methods. Clustering is one of the data mining techniques used in the process of partitioning a set of data or objects into a set of meaningful sub-classes.

Data clustering is a process of placing similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. In the field of data mining, different clustering algorithms are proved for their clustering quality. The main advantage of clustering is that interesting structures and patterns can be found directly from huge datasets with little or none of the background knowledge. The quality of a clustering method depends on:

- The similarity measure used by the method and its fulfillment
- Its capability to discover some or all of the hidden patterns
- The definition and representation of cluster chosen

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

This research work deals with the clustering algorithm, namely, centroid based K-medoids, which is widely used simplest partition based unsupervised learning algorithm that solves the well-known clustering problem. The method follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori [6].

In this research work, dengue fever data mining techniques are described data mining applications, partitioning clustering method can be used to develop the dengue sector.

II. LITERATURE REVIEW

Bhat, A., et al. [1] has provided K-medoids clustering using partitioning around medoids for performing face recognition. It explores a novel technique for face recognition by performing classification of the face images using unsupervised learning approach through K-Medoids clustering. Partitioning Around Medoids algorithm (PAM) has been used for performing k-medoids clustering of the data. The results are suggestive of increased robustness to noise and outliers in comparison to other clustering methods. Therefore the technique can also be used to increase the overall robustness of a face recognition system and thereby increase its invariance and make it a reliably usable biometric modality. Comparing these two algorithms, K Medoids algorithm provides better result.

Marimuthu, T., et al. [2] exposed novel bio-computational model for mining the dengue gene sequences. It proposed a bio-computational model called sequence miner to interpret the relationship among the dengue viruses. It performs the classification, association rules and visualizing the results through the interactive tool. The accuracy of the proposal model is 96.74% which is calculated by giving the 10,735 varying length of the sequences as the input, 10,198 sequences are correctly classified. The relationship between dengue serotypes are predicted via the proposed tool. It helps to the biotechnologies and drug designers for discovering an effective vaccine for dengue.

Rao, K., K., N., et al. [4] proposed classification rules using decision tree. This paper discovered the rules for the disease hit and explores what rule can act in this area for the future prediction. The main objective is creating prediction model for predicting the chances of occurrences of dengue disease. Knowledge extracted from the clustering model which help to identify the significant characteristics of insolvent people. The decision tree classification model achieved 97% accuracy.

Sharma, H., et al. [5] have exposed detecting pattern of disease spread in various states of India using data mining techniques like classification and clustering author finds hidden patterns which give meaningful decision making to epidemic diseases and the impact of diseases in the various states of India. Apriori algorithm and K-Means techniques are performed and finally the authors hope this paper can act as a decision maker in order to identify that what all states are there where need to focus in order to identify the cause of diseases.

Santhanam, T., et al. [8] have proposed a comparative analysis between k-medoids and fuzzy c-means clustering algorithms for statistically distributed data points. This research work deals with, two of the most representative clustering algorithms namely centroid based K-Medoids and representative object based Fuzzy C-Means are described and analyzed based on their basic approach using the distance between two data points. The performance of the algorithms is investigated during different execution of the program for the given input data points. Based on experimental results the algorithms are compared regarding their clustering quality and their performance. The total elapsed time to cluster all the data points and Clustering time for each cluster are also calculated in milliseconds and the results compared with one another. From the result, the performance of the FCM algorithm is relatively better than that of K-Medoids algorithm.

Santhanam, et al. [7] has exposed computational complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. In this research, the most representative algorithms K-Means and K-Medoids were examined and analyzed based on their basic approach. The best algorithmic program in each category

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

was found out based on their performance. The input data points are generated by two ways, one by using normal distribution and another by applying uniform distribution. The algorithms were implemented using JAVA language and the performance was analyzed based on their clustering quality. The execution time for the algorithms in each category was compared to different runs. The accuracy of the algorithm was investigated during different execution of the program on the input data points. The average time taken by K-Means algorithm is greater than the time taken by K-Medoids algorithm.

III. DATA SET DESCRIPTION

A. *DATA COLLECTION*

The input data have been collected from urban Ho Chi Minh City, Vietnam. Initially the data size is 171 attributes and 1910 records.

B. *DATA PREPARATION*

The dengue patient's data are collected from household clustering of dengue. The dataset was gathered between October 2010 to January 2013 in Ho Chi Minh City (HCMC)—the largest city in Vietnam. The Hospital for Tropical Diseases (HTD) is the guideline hospital for epidemic diseases in southern Vietnam, located in central HCMC. Attributes used in this study are described in Table I.

TABLE I. DATA SET INFORMATION

Attributes	Description	Possible Values
c_age	Age of patients	1-100
c_agegroup	Stages of age	>5, 5>=15, 15-35, 35-55, >=55
c_igmresult1	Immunoglobulin test 1	neg, pos, equiv
c_igmresult2	Immunoglobulin test 2	neg, pos, equiv
c_igm_seroconvert	Immunoglobulin serotype convert	neg, pos, equiv
c_iggresult1	Immunoglobulin G test 1	neg, pos, equiv
c_iggresult2	Immunoglobulin G test 2	neg, pos, equiv
c_iggresult3	Immunoglobulin G test 3	neg, pos, equiv
c_ns1result1	Non structural protein1 result 1	Samp_finished, neg, pos, equiv
c_ns1result3	Non structural protein1 result 3	Samp_finished, neg, pos, equiv
c_ns1result_all	All Non structural protein1 result	Samp_finished, neg, pos, equiv
i_serotype	Type of dengue serotype	0, 1, 2, 3, 4
i_finalclass	To identify the fever is dengue or not	Dengue, not dengue

TABLE I represent the attributes description and mentioned the possible values.

C. *Preprocessing and Feature selection*

Many existing, industrial and research datasets contains missing values. They are introduced due to various reasons, such as manual data entry procedures, equipment errors and incorrect quantifications. Hence, it is usual to find missing data in most of the information sources used. The detection of incomplete data is easy in most cases, looking for null values in a dataset. However, this is always not true, since missing values can appear with the form of outliers or even wrong data.

Missing values make it difficult for analysts to effectuate data analysis. Three types of problems are usually associated with missing values.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

- Loss of efficiency
- Complications in handling and analyzing the data
- Bias resulting from differences among missing and complete data

Step 1: Apply D win's method to impute the missing values

	c_igm_seroconvert	c_iggresult1	c_iggresult2	c_iggresult3
219	negative	neg	neg	neg
220	negative	neg	neg	neg
221	negative	neg		neg
222	seroconvert	pos	pos	pos
223	negative	neg	neg	
224	negative	neg	neg	neg
225	negative	neg	neg	neg
226	positive	equiv	neg	pos
227	negative	neg		neg
228	negative	neg	neg	neg
229	negative	equiv	neg	neg
230	negative	neg	neg	neg
231	negative	equiv	pos	pos
232	positive	pos	pos	pos
233	negative	neg	neg	neg
234	equiv/pos	pos	pos	pos
235	negative	pos	pos	pos
236	seroconvert	neg	pos	pos
237	negative	equiv	equiv	equiv

Fig. 1. Raw Data Set

Figure 1 illustrates original data set have been displayed.

	c_igm_seroconvert	c_iggresult1	c_iggresult2	c_iggresult3	c_ns1result1
219	positive	neg	neg	neg	neg
220	negative	neg	neg	neg	neg
221	negative	neg	neg	neg	neg
222	negative	neg	neg	neg	neg
223	negative	neg	neg	neg	neg
224	negative	neg	neg	neg	neg
225	positive	pos	pos	pos	neg
226	negative	pos	pos	pos	neg
227	negative	equiv	neg	neg	neg
228	negative	neg	neg	neg	neg
229	negative	equiv	neg	neg	neg
230	negative	neg	neg	neg	neg
231	negative	neg	equiv	neg	neg
232	seroconvert	equiv	equiv	equiv	neg
233	negative	neg	neg	neg	neg
234	negative	equiv	neg	equiv	neg
235	negative	neg	equiv	neg	neg
236	negative	pos	pos	pos	neg
237	negative	neg	neg	neg	neg

Fig. 2. Impute Missing Values using Dwins Method

Figure 2 represents the imputed missing values using dwins method.

Data: del

	age	agegroup	c_igmresult1	c_igmresult2	c_igm_seroconvert	c_iggresult1
1	47	6	2	2	4	2
2	51	6	2	2	4	2
3	83	3	1	2	4	2
4	8	4	2	2	4	2
5	55	6	2	2	4	2
6	60	6	2	2	4	2
7	38	4	2	2	4	2
8	16	5	2	2	4	2
9	66	4	1	1	3	2
10	58	5	2	2	4	2
11	43	6	3	3	5	3
12	1	6	2	2	4	2
13	69	5	2	2	4	3
14	48	3	2	2	4	2
15	54	6	2	2	4	2
16	32	3	3	2	5	3
17	32	3	2	2	4	2
18	36	5	2	2	4	2
19	52	6	2	2	4	2

Fig. 3. numerical data of dengue dataset

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Figure 3 shows the numerical data. The dengue dataset values are changed to numerical values for further implementation of algorithms.

Step 2: Apply One R technique for identifying weights of the values

One R Method

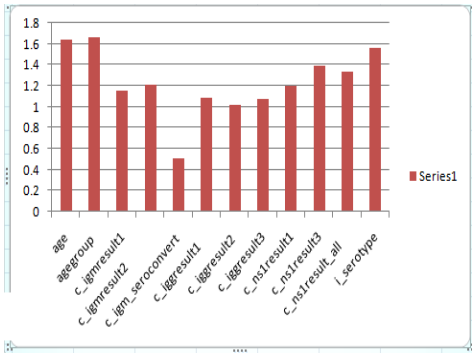

Fig.4 One R Feature Selection Method

Figure 6 represents the weights of each attribute. It decides which weighted attributes are suitable for further clustering process.

IV. METHODOLOGY

CLUSTERING

Clustering is one of the better important research areas in the field of data mining. In simple words, clustering is a division of data into different groups. Data are grouped into clusters in such a way that data of the alike group are similar and those in other groups are dissimilar. Clustering is a method of unsupervised learning. The main advantage of clustering over-classification is flexible to changes and helps single out useful features that distinguish different groups. Clustering methods can be classified into the following types-

- Partitioning Method
- Grid-Based Method
- Density-based Method
- Hierarchical Method
- Density-based Method
- Constraint-based Method
- Model-Based Method

Partitioning Method

Partitioning is one of the clustering methods. Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will sort the data into k groups, which satisfy the following requirements:

- Each group contains at least one object.
- Each object must belong to exactly one group.

K-Medoid Algorithm

The k-medoids algorithm is a clustering algorithm relevant to the k-means algorithm and the medoid shift algorithm. Both the k-means and k-medoids algorithms are partitioned (breaking the dataset up into groups) and both attempt to minify squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids selects data points as centers. K-medoid is a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. An effective tool for determining k is the silhouette. It is more robust to noise and outliers as compared to k-means. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimum i.e. it is a most centrally located point in the given data set [5].

Among many algorithms for K-medoids, Partitioning Around Medoids (PAM) proposed by Kaufman and Rousseeuw (1990) is known to be most powerful. However, PAM also has a drawback that it works inefficiently for large data sets due to its complexity. We are interested in developing a new K-medoids clustering method that should be fast and efficient.

Algorithm for K-Medoids

Input:
(1) Database of O objects
(2) A set of k initial centroid Cm= {C1, C2 . . . Ck}

Output:
A set of k clusters

Steps:
1. Initialize initial medoid which is very close to centroid {C1, C2 . . . Ck} of the n data points
2. Associate each data point to the closest medoid. ("closest" here is defined using any valid distance metric, most generally Euclidean distance, Manhattan distance or Minkowski distance)
3. For each medoid m
4. For each non-medoid data point o
5. Swap m and o and compute the total cost of the configuration
6. Select the configuration with the lowest cost
7. Repeat steps 2 to 5 until there is no change in the medoid.

Complexity of Algorithm

Enhanced algorithm requires a time complexity of *O (n 2)* for finding the initial centroid, as the maximum time required here is for computing the distances between each data point and all other data-points in the set D. Complexity of remaining part of the algorithm is O(k(n-k)2) because it is just like PAM algorithm. So over all complexity of the algorithm is O (n2), since k is much less than n.

Demonstration of PAM

Cluster the following data set of ten objects into two clusters i.e. *k*=2. Consider a data set of ten objects as follows:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

	agegroup	i_serotype
2	6	4
3	6	0
4	3	1
5	4	1
6	6	0
7	6	0
8	4	1
9	5	2
10	4	2
11	5	3
12	6	1
13	6	0
14	5	1
15	3	4
16	6	0
17	3	0
18	3	4

Fig. 1. Data point for distribution of the data

IV. EXPERIMENTAL RESULTS

TABLE II. SELECTED ATTRIBUTE

Table II represents the one R feature selection has been applied on the preprocessed data. Relevant attributes only selected for further process.

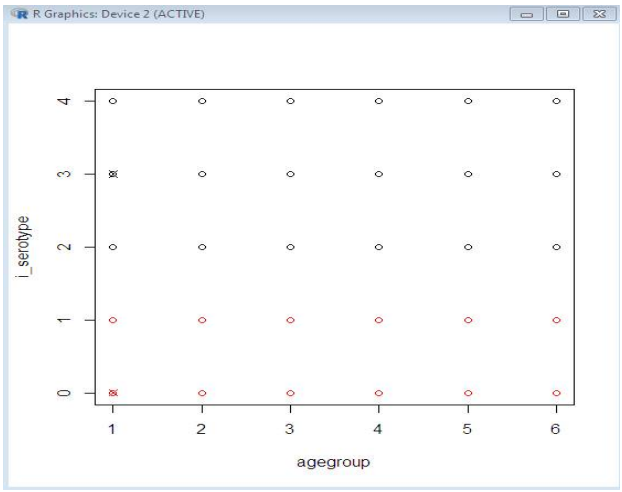

Fig. 3 Distribution of the data

Method	Selected Attributes					
OneR method Attributes	Age	agegroup	c_igm result1	c_igm result2	c_igm_sero convert	c_iggresult1
	c_iggresult2		c_iggresult3		c_nslresult1	
	c_nslresult1		c_nslresult1		i_serotype	

- First, initialize *k* centre
- Let us assume c1 and c2 are selected as medoid.
- Calculate distance so as to associate each data object to its nearest medoid.

TABLE III. K-MEDOIDS CLUSTERED ATTRIBUTES

	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5	Age group 6
cluster 1	572	8	15	29	15	31

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

cluster 2	1040	21	31	51	28	69

Table III represents the clustered age group data depending upon two clusters using k-medoids clustering algorithm.

TABLE IV K-MEDOIDS CLUSTERED ATTRIBUTES

	Serotype 0	Serotype 1	Serotype 2	Serotype 3	Serotype 4
cluster 1	572	8	15	29	15
cluster 2	1040	21	31	51	28

Table IV represents the clustered serotype data depending upon two clusters using k-medoids clustering algorithm.

V. CONCLUSION

The proposed k-medoid algorithm runs just like K-means clustering algorithm. The proposed algorithm is used the systematic method of choosing the initial medoids. The performance of the algorithm may vary according to the method of selecting the initial medoids. K-medoid clustering is increasing the efficiency of the output. This is the most effective technique to predict the dengue patients with serotypes and dengue dataset was fully clustered.

VI. FUTURE WORK

In this research, partitioning cluster method is implemented to find the evolutionary patterns for predicting dengue fever. In future, it will be enhanced that the evolutionary patterns can find out by using hierarchical clustering algorithms for dengue fever prediction.

REFERENCES

[1] Bhat, A., "K-Medoids Clustering using Partitioning Around Medoids for Performing Face Recognition", International Journal of Soft Computing, Mathemetics and Control(IJSCMC), Volume. 3, No. 3, August 2014, pp: 1-12.
Marimuthu, T., and Balamurugan, V., "A Novel Bio-Computational Model for Mining the Dengue Gene Sequences",International Journal of Computer Engineering & Technology, Oct 2015, Volume. 6, Issue. 10, pp: 17-33.

[2] Petel, A., and Singh, P., "New Approach for K-mean and K-medoids Algorithm" International Journal of Computer Applications & Technology and Research, 2013, Volume. 2, Issue. 1, pp: 1-5.
Rao, K, K, N., Dr. Varma, S, P, G., and Dr. Rao, N, M., "Classification Rules Using Decision Tree for Dengue Disease", International Journal of Research in Computer and Communication Technology, March- 2014, Volume.3, Issue.3, pp: 340-343.
Sharma, H., and Sharma, P., "Application of Data Mining In DetectingnPattern of Disease Spread In Various States Of India", International Journal of Advanced Research in Computer Science and Software Engineering, June 2014, Volume. 4, Issue. 6, pp: 291-294.
Shaukat, K., Masood, N., Shafaat, B, A., Jabbar, K, Shabbir, H., and Shabbir, S., "Dengue Fever in Perspective of Clustering Algorithms", Data Mining in Genomics & Proteomics, 2015, Volume. 6, Issue. 3, pp:
1-5.

[3] Velmurugan, T., and Santhanam, T., "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points", Journal of Computer Science, 2010, Volume. 6,Issue. 3, pp: 363-368.

[4] Velmurugan, T., and Santhanam, T**.,** **"**A Comparative Analysis Between K-Medoids and Fuzzy C-Means Clustering Algorithm for Statistically Distributed Data Points", Journal of Theoretical and Applied Information Technology, Volume. 27, No. 1, 2011, pp: 19-30.

[5] https://en.wikipedia.org/wiki/K-medoids.

[6] Enhanceedu.iiit.ac.in/wiki/images/Kmedoids.pdf