Dengue Fever Prediction Using K-Means Clustering Algorithm

P. Manivannan
Research Scholar
Department of Computer Science
Ayya Nadar Janaki Ammal College
Sivakasi, Tamil Nadu, India
manivannan0823@gmail.com

Dr. P. Isakki @ Devi
Assistant Professor
Department of Computer Applications
Ayya Nadar Janaki Ammal College
Sivakasi, Tamil Nadu, India
harish24devi@gmail.com

Abstract— **Dengue fever is a virus infection which is transmitted to humans by mosquitoes that living in tropical and subtropical climates and carries the virus. The dengue viruses occur in 4 serotypes (DENV-1 to DENV-4). A dengue disease ranges from mild febrile disease to severe hemorrhagic fever. Predicting the relationship between the dengue serotypes will surely help the biotechnologists and bioinformaticians to move one step forward to discover antibiotic for dengue. This paper has been focused four stages namely preprocessing, attribute selection, 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 substituting all missing values for nominal and numeric attributes with mode and mean value. Dengue virus can be predicted by applying different data mining techniques. The main goal of research work is to predict the people who are affected by dengue depending upon categorization of age group using K-means clustering algorithm has been implemented.**

Keywords—Dengue, Data Mining, Medical Documents, Clustering techniques, K-means algorithm.

I.INTRODUCTION

Dengue fever is classified into two categories namely type 1 and 2, conceding to world health organization. The first type is a classical dengue which is referred as dengue fever and the other type is known as dengue hemorrhagic fever. DHF 1, DHF 2, DHF 3 and DHF 4 are the four types of dengue hemorrhagic fever [1]. Symptoms of dengue include an acute headache, joints pain, rashes, leucopenia, thrombocytopenia and muscle ache. The break-bone fever holds the symptoms of joints pain and muscle ache. In a very small segment of cases, the disease might develop further into aggressive dengue hemorrhagic fever (DHF) which results in decrease in 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 method of finding knowledge such as patterns, associations, changes, anomalies and significant structures, from huge amount of data stored in database, data warehouses or other information storage place. Due to the wide availability of huge amount of data in electronic forms, the probability require to turn data into useful knowledge for extensive application including market analysis, business management, and decision support. Data mining has attracted an immense deal of attention in information industry in recent years.

Clustering or data grouping is one of the approaches in data mining. This research work includes data mining clustering techniques that can be used to develop the dengue sector. This paper deals with the application of k-means clustering algorithm to predict the dengue fever.

II.LITERATURE REVIEW

Marimuthu, T., et al [1] exposed new bio-computational model for mining the dengue gene sequences and 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 [2] proposed classification rules using decision tree. This paper discovered the rules for the disease hit and explores what role can act in this area for the future prediction. The main objective is creating a prediction model for predicting the chances of occurrences of dengue disease. Knowledge takes out from the clustering model which helps to analyze the significant characteristics of affected people. The decision tree classification model achieved 97% accuracy.

978-1-5090-4778-9/17/$31.00 ©2017 IEEE

Sharma, H., et al [3] have exposed detecting pattern of disease transmission in various states of India using data mining techniques like classification and clustering. In this Paper finds hidden patterns which give purposeful decision making to infectious diseases and the impact of diseases in the various states of India. Apriori algorithm and K-Means techniques are performed and finally the authors belief 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.

Shaukat, K., et al, [4] presented a various classification algorithms to predict the dengue fever and comparing the algorithms by using the Correctly and Incorrectly Classified data, Relative Absolute Error, Tp rate, Fp rate precision measurements. Five Classification techniques are used namely NB, J48, SMO, RT and REP tree. NB and J48 are the better performance classifier techniques by that way, they has achieved an accuracy of 92% and 88%, takes minimum time to run and produce ROC area=0.815, and had minimum error rate.

Bhatia, R., et al [5] exposed predicting the outbreak of dengue based on environmental factors using Support Vector Machine (SVM) based algorithm. The result shows that there is a strong relation between the environmental factors and the high incidence cases of dengue. The results show the effective classification of the dengue cases can be performed by using PCA (Principal Component Analysis) technique for feature selection followed by c-SVM model with Gaussian kernel. This model provides both good accuracy and complexity of the system.

Subitha, N., et al [6] has proposed spatial data mining for diagnosis of dengue fever. It extracts pattern from spatial database by using K-Means algorithm which refers to patterns implicitly stored in spatial databases. This paper has been performed to improve the quality of mining in a pruned data set using association mining, progressive refinement, and fast algorithm. This concept is applied in the area of image segmentation with microscopic blood image as input and signals are filtered using the neural network to predict the well good result about dengue fever. Classification result is carried out using the back propagation network (BPN). This gives 98% accurate result within a short period.

Sudsom, N., et al [7] have described a spatial clustering approach for identifying risk households of dengue virus infection during the period of insecticide spraying-ultra low volume (ULV). Application of spatial analyst tools and spatial statistics tools in ArcGIS 10.1 were used to determine mosquito density and identify risk households using ovitrap index. The prediction maps of Aedes aegypti vector abundance were illustrated by kriging technique. Base on the results, the cluster of Ae. Aegypti populations were detected on four day after the spraying. This finding shows the significant spatial pattern of dengue vector populations may cause high risk areas of dengue virus infection after insecticide treatment. This methodological framework could be used for improving the strategy of dengue vector and outbreak control. The spatial association between dengue vector and the overage of space spraying requires further study.

III.DATA SET EXPLANATION

Data Collection

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

Data Preparation

The dengue patient's data was collected from household clustering of dengue, which has collected between Oct-2010 to Jan-2013 in Ho Chi Minh City in Vietnam. The Hospital for Tropical Diseases (HTD) is the guideline hospital for infectious diseases in southern Vietnam, located in central HCMC. Attributes in dengue dataset are described in Table I.

Table I. Data Set Information

Attributes	Description	Possible Values
c_patientid	Patients identity	Numbers
c_housecode	Family identity code	Numbers
c_typecontact	Contact id	neighbor,
c_age	Age of patients	1-100
c_datevisit1	First visit date	-
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	neg, pos, equiv
c_iggresult2	Immunoglobulin G test	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	Samp_finished, neg, pos, equiv
i_onsetdate	Fever arrival date	-
i_serotype	Type of dengue	0, 1, 2, 3, 4

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

Preprocessing and Feature Selection

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

Step 2: To convert all the values of the attribute into numeric values

Step 3: Apply One R Filter for calculating the weights of the attributes

	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. Replaced Missing Values using Dwins Method

Figure 2 represents the replaced missing values using dwins method have been displayed

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. Changed values into numerical format

Figure 3 shows the numerical data for dengue dataset.

IV. METHODOLOGY

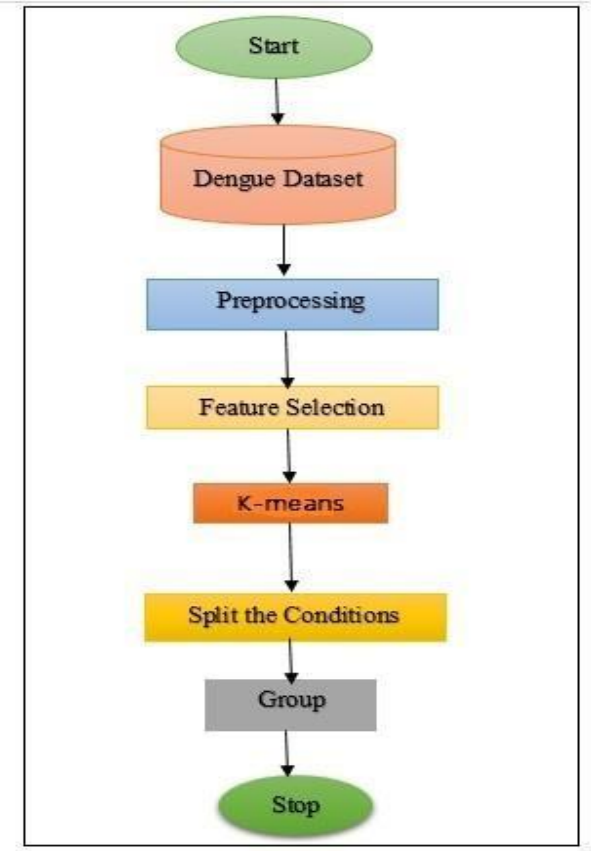

Fig. 4. Methodology Framework

Figure 4 represents the methodology framework for k-means clustering algorithm.

V. K-MEANS CLUSTERING ALGORITHMS

Clustering Methods

Clustering is a data mining technique which is used in the process of dividing a set of data or objects into a set of meaningful sub-classes. In other words, identical objects are grouped in one cluster and non-identical objects are grouped in another cluster. The main benefit of clustering over classification includes that, it is flexible to changes and assists single out helpful features that distinguish among various groups. The types of Clustering methods can be listed as follows:

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

Partitioning Method

The most popular method of clustering is the partitioning based method. Assume we are given a database of 'n' objects and the partitioning method generates 'k' partition of data. Each partition indicates a cluster and k ≤ n. It means that it will sort the data into k groups, which convince the following requirements:

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

K-means Clustering

K-Means is a clustering algorithm used to classify or group the objects based on attributes/features that are partitioned into K number of group where K is positive integer number. In this paper, k-means clustering algorithm can partition the dengue data set into k clusters. The grouping is achieved by minimizing the sum of squares of distances between data and the related cluster centroid.

K-means Algorithm

K-means algorithm sorts out the k-means clustering problem and works as follows,

- Set the N number of cluster for K (K= 0, 1, 2 … N)
- Then, Center of the cluster is initialized
  μi = value , i =1,...,k
- Each data pint marks the closed cluster
  $\mathbf{c}i=\{j:d(\mathbf{x}j,\mu i)\leq d(\mathbf{x}j,\mu l),l\neq i,j=1,...,n\}$
- Set the locus of each cluster to the mean of all data points insertion to that cluster
  $\mu i=1|ci|\sum j\in ci\mathbf{x}j,\forall i$
- Repeat steps 2-3 till line up
- Notation |**c**|= number of elements in **c**

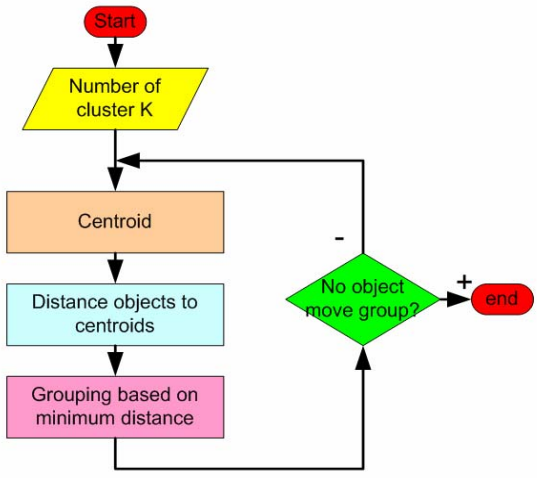

Fig. 5. K-means process

Figure 5 represents the k-means algorithm process.

If the number of data is less than the number of clusters then assign each data as the centroid of the cluster. Cluster number will be assigned to each centroid. If the number of data is bigger than the number of clusters for each data, then calculate the distance to all centroid and find the minimum distance. This data is belonging to the cluster that has a minimum distance from this data. Since it is not sure about the location of the centroid, so it essential to adjust the centroid location based on the current updated data and assign all the data to this new centroid. This process is replicated until no data is moving to another cluster anymore. Mathematically this loop can be tested convergent [10].

VI. EXPERIMENTAL RESULTS

Table II. Selected Attribute

Method	Selected Attributes					
One R Method Attributes	ag e	age group	c_igm result1	c_igm result2	c_igm_sero convert	c_ig gres ult1
	c_iggresult2		c_iggresult3		c_nslresult1	
	c_nslresult1		c_nslresult1		i_serotype	

Table II represents the one R feature selection method applied on the collected data. Preprocessed data is taken for further process.

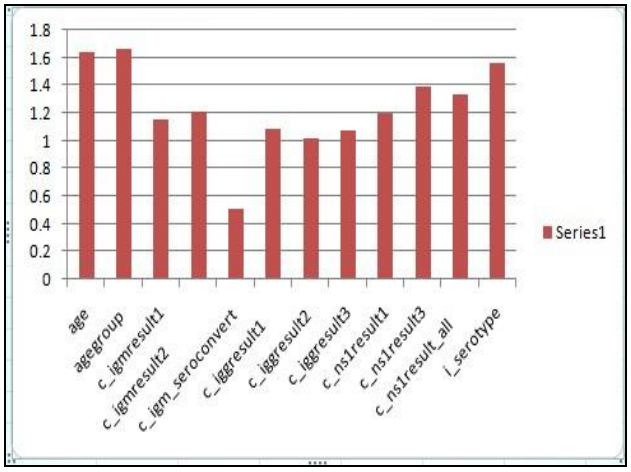

Fig. 6. One R Feature Selection Method

Figure 6 represents the one R feature selection have been displayed

Figure 7 shows k-means clustered form distance between the age group and serotypes. The age group attribute has been divided into three groups namely, age group 1 (0-15)/age group 2 (15-55)/age group 3 (>=55) and the serotype attribute is referred to as 0-4. Serotype 0 is the classic/normal dengue fever and serotypes 1 to 4 are the types of dengue hemorrhagic fever.

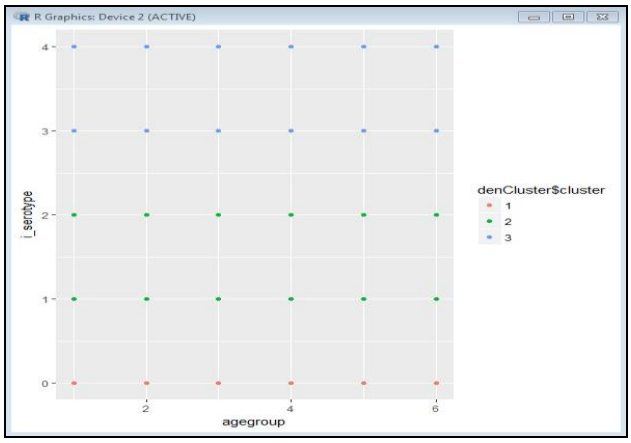

Fig. 7. Distances between the age group and serotype clusters

Table III. K-Means Clustered Attributes

	serotype 0	serotype 1	serotype 2	serotype 3	serotype 4
agegroup 1	913	0	0	0	0
agegroup 2	0	327	323	0	0
agegroup 3	0	0	0	98	249

Table III represents the clustered serotype data depending upon age group using k-means clustering algorithm.

VII. CONCLUSION

Dengue is a viral disease responsible for most of the illness and death in tropical and subtropical regions. Clustering techniques are very good tools for the visualization of diseases. In proposed methodology, household clustering of dengue is to be done by using dengue serotypes depending upon the age group through applying the k-means clustering algorithm. K-Means clustering is increasing the proficiency of the output. This is the most effective technique to predict the dengue patients with serotypes and dengue dataset was fully clustered.

REFERENCES

[1]. 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]. 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

[3]. Sharma, H., and Sharma, P., "Application of Data Mining In Detecting Pattern 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

[4]. 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

[5]. Shweta., Bhatia, R., Jindal, A., and Sood, M., "Prediction of Dengue Outbreak using Environmental Factors", 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences, 2016, pp: 716-719

[6]. Subitha, N., and Padmapriya, A., "Diagnosis for Dengue Fever Using Spatial Data Mining", International Journal of Computer Trends and Technology, Augest 2013, Volume. 4, Issue. 8, pp: 2646-2651

[7]. Sudsom, N., Thammapalo, S., Pengsakul, T., and Techato, K., "A Spatial Clustering Approach to Identify Risk Areas of Dengue Infection After Insecticide Spraying", Journal Technology, Dec 2015, Volume.78, pp:73-77

[8]. https://en.wikipedia.org/wiki/K-means_clustering

[9]. http://www.onmyphd.com/?p=k-means.clustering

[10]. http://people.revoledu.com/kardi/tutorial/kMean/Algorithm.html