Development of GIS based Disease Outbreak Detection System utilizing Open Source Technologies

Vineet Kumar
Geoinformatics Department
IIRS, ISRO
Dehradun, India, vineet1109@gmail.com

Koti Shiva Reddy
Geoinformatics Department
IIRS, ISRO
Dehradun, India, shivareddy@iirs.gov.in

***Abstract*—In the recent times, design and implementation of desktop based disease outbreak detection system using open source technologies has always been a focused and challenging task. Health data has both the spatial and temporal variability in nature which is required to be mapped to have the clusters of the infected cases to understand the spreading of disease in an area. This paper outlines the development of desktop based disease outbreak detection system, a GIS tool by the use of open source applications like Python, Java and PostGIS along with interface design and database development whereas the WMS services of Bhuvan, Google Map and Geoserver has been used for the visualization purpose. Density base clustering approach has been used to map the point data and cluster formation. The data sets used for the study are standard health data generated by different government organization. The developed GUI will lead to detect clusters in the particular areas where the number of cases is much higher and results can be used to do the prospective and retrospective analysis to have the future scenarios generation to make the early decisions.**

Keywords- Disease outbreak detection system, PostGIS, WMS services, density based clustering.

I. INTRODUCTION

Public health weather for a developed nation or a developing nation has become a great concern. During the recent time, the spreading of deadly disease like Ebola and H1N1(bird flu) has shown the importance for the implementations of outbreak detection systems in order to have the minimum casualties and also the further spreading of diseases should be tracked and controlled by timely implementation of control and preventive measures. The best existing way to do the surveillance is spatial clustering of the disease data having the geo-locations of the effected person i.e. the latitude and longitude of the location, type of the disease and population of the area.

Disease outbreak detection systems using clustering algorithms like k-mean, CluStream etc. does not support for continuous evolving data like real time data [1]. Outbreak detected from these systems comes generally with some duration of time gap which create concern, when there is a huge outbreak spreading within a little span of time. To have better and effective response toward controlling epidemic disease, is to have a real time disease outbreak detection system which can perform clustering of the real time stream data directly coming from the ground sources like asha workers, village level community health centers(CHC) and district hospitals.

To do the real time stream data processing, algorithms needs number of resources like high performance computing and much important an online source for stream data, which is difficult to generate as in developing countries till now much of the work is not automated [2]. The data collected until it reaches the last level of management who do the clustering on the basis of their experience is in hard copy format or some word document format. So, to have the knowledge of disease outbreak, evaluation of algorithms having optimum space and time complexity is needed and a system to generate stream data is to be developed to have real time stream data clustering.

II. ALGORITMH IMPLEMENTED

Density based clustering algorithm has played a vital role in finding nonlinear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm [3, 4]. It uses the concept of density reachability and density connectivity.

Algorithmic steps for DBSCAN clustering

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

1) Start with an arbitrary starting point that has not been visited.

2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).

3) If there are sufficient neighborhoods around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).

4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.

5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6) This process continues until all points are marked as visited.

Analysis:

- O(m) Space Complexity
- Using KD Trees the overall Time Complexity reduces to O(m * logm) from O(m^2) [5,6]

Advantages

1) Does not require a-priori specification of number of clusters.
2) Able to identify noise data while clustering.
3) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters [7,8].

III. SYSTEM IMPLEMENTATION

Python PYQt module has been used for development for system interface. Cluster file output has been displayed in webview. A selection panel has been given to show outputs using google map services, bhuvan WMS and geoserver. The browser button connect the code to the database having latitude, longitude and the number of cases in that area.

TABLE I. INPUT DESCRIPTION

Inputs	*Description*
District Name	*Name of the district to be mapped*
Disease Vector	*Name of the diseaseto be clustered*
Database File	*Database file containing lat, long and case information.*

When the system starts, the first step is to select the district for which the case data information is available, than the disease to clustered in entered, after the connection is set up with the database file to get the required values. The clusters generated by algorithm processing can be viewed over google map service or with the help of geoserver or by bhuvan wms service. Downloaded files are the shape files of the clusters.

IV. RESULT AND CONCLUSION

The developed system uses the algorithm taking latitude and longitude as input values and classifying them into the respective clusters. The lat, long point going into the cluster depends on the density and the distance from the center of the circular cluster. Fig.1. Shows the output of the dbscan algorithm implemented in python.

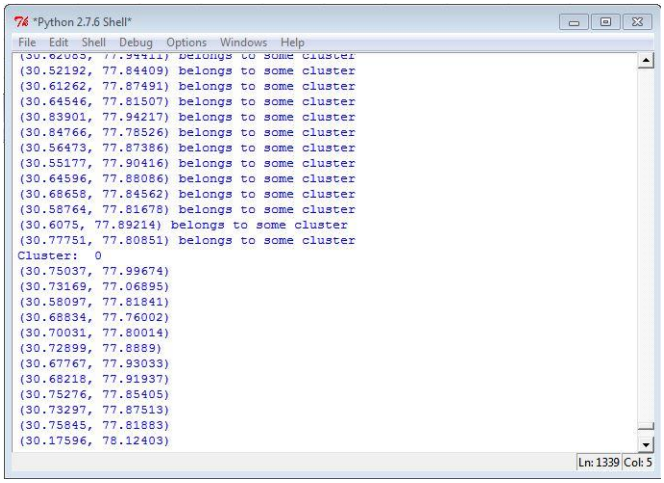

Fig. 1. DBSCAN Clustering Algorithm

Figure 2 shows the interface design developed for the disease outbreak detection system, consisting of input data panel having the district, disease vector and database input fields, output data format usng bhuvan, google and geoserver services and webview to display the cluster outputs.

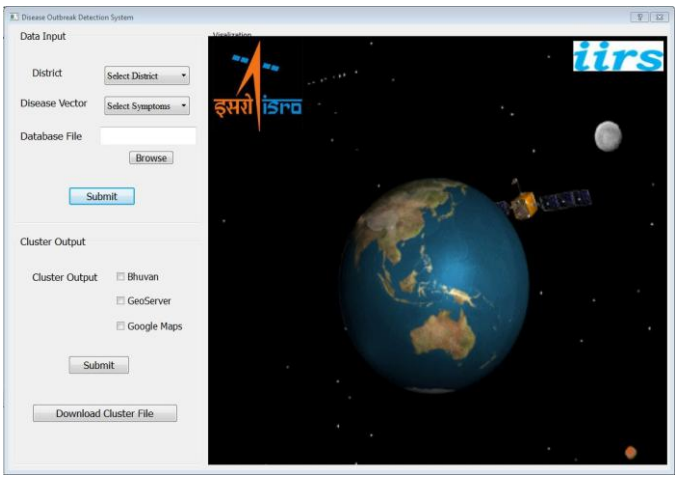

Fig. 2 Disease Outbreak Detection System Interface

After connecting to a database the lat, long and the case information will be enterd into the system which on processing will give clusters that be dispalyed on google map

service. Fig. 3 shows the individual points on the map on zoom level 1. On zoom out the points will merge and will be dispalyed on different colour clusters. Fig 4 and fig 5 are displaying the clusters on diffrenent zoom levels.

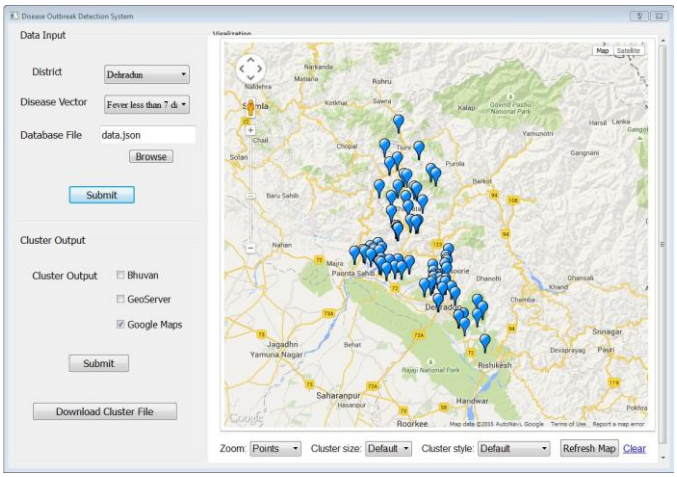

Fig 3 Individuals points on google maps

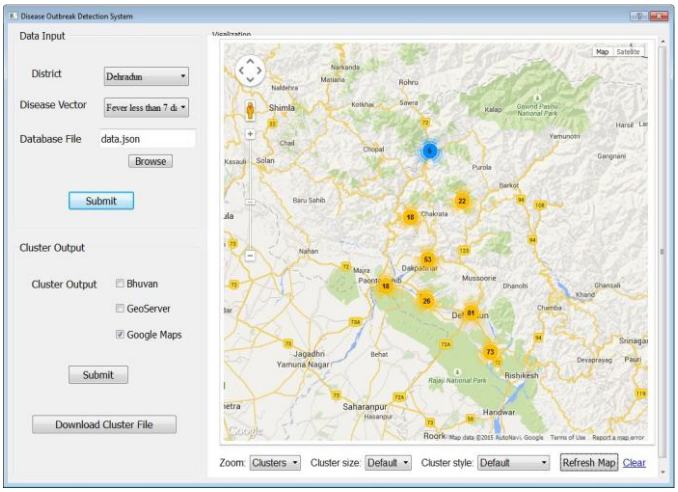

Fig 4. Clusters of different colours and size

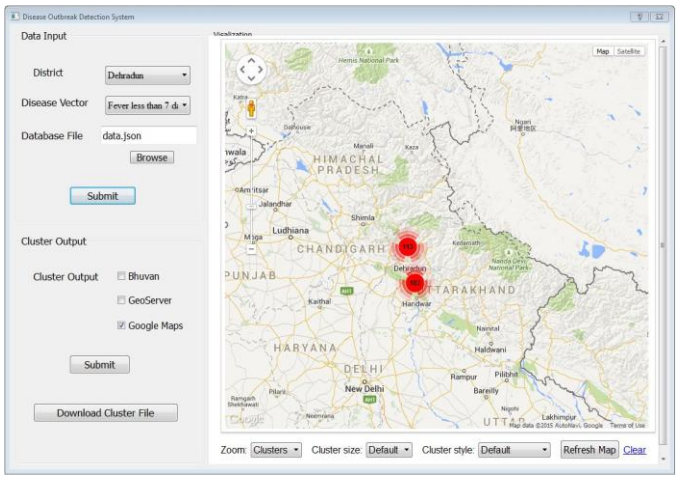

Fig 5. Clusters of complete data higher number of cases.

We can also display the outputs in form of clusters shape files on geoserver. Fig 6 shows the output using geoserver. On clicking the cluster we can get the information of number of cases, total area covered, total population and other clluster related information.

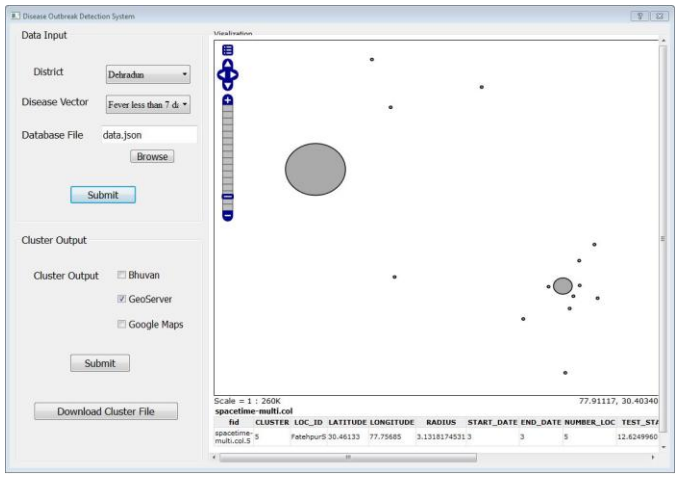

Fig 6. Displaying cluster using Geoserver

Bhuvan services also been used to display the th cluster shape file. The shape file is displayed on bhuvan wms service

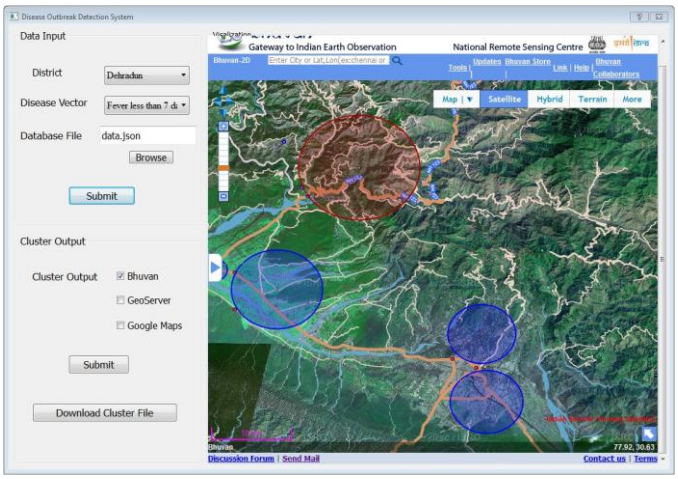

Fig 7. Clustering using Bhuvan WMS.

REFERENCES

[1] Aggarwal, C. et al.. A framework for clustering evolving data streams. conference on Very large data …. 2003.

[2] Barbara D., Requirement for Clustering Data Stream. SIGKDD Explorations International Conference on Knowledge and Data Mining, Boston, August 2000

[3] O'Callaghan L., Mishra N., Meyerson A., Guha S., and Motwani R. High-Performance Clustering of Streams and Large Data Sets. International Conference on Data Engineering (ICDE) 2002 (to appear).

[4] Kulldorff M., Bernoulli, Discrete Poisson and Continuous Poisson Models: A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997.

[5] Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F,Space-Time Permutation Model:. A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine,2005.

[6] Jung I, Kulldorff M, Richard OJ ,Multinomial Model:. A spatial scan statistic for multinomial data. Statistics in Medicine, 2010,

[7] Jung I, Kulldorff M, Klassen A ,Ordinal Model:A spatial scan statistic for ordinal data. Statistics in Medicine, 2007.

[8] Huang L, Kulldorff M, Gregorio D ,Exponential Model: A spatial scan statistic for survival data. Biometrics, 2007.

AUTHORS PROFILE

Vineet Kumar is an M.Tech student in Geoinformatics Department, IIRS, Dehradun, currently pursuing final year project over the development of real time disease outbreak detection system.

Koti Shiva Reddy, is a scientist 'SD' at Geoinformatics Department, IIRS. His research expertise is in the field of Service Oriented GIS and Public Health Data Modeling in GIS.