Spatial-Temporal Data Analysis in Epidemic Datasets

Siddharth Sujir Mohan

School of Computing

Informatics and Decision

Systems Engineering

[siddharthsujir.mohan@asu.edu](mailto:siddharthsujir.mohan@asu.edu)

**ABSTRACT:**

The first phase of the project aims at creating an epidemic word dictionary based on the data sets provided. The values each of the data file are normalized and quantified based on the Gaussian band that it falls under. The length of each Gaussian band varies depending upon the value of resolution provided. We then form a word by shifting h units at a time on the file, with the window length being w units.

The winavg and windiff for each window is computed by comparing the values with corresponding values of onehop neighbours at a time t. The norm of the win vector is found out, which gives the strength of the vector .A heat map is then plotted for the data file provided. The heat map also depicts the states having the maximum strength and minimum strength along with their neighbouring states.

**Key Words-** Epidemic, Data sets, one hop neighbours, window, shift length, norm of vector, heat map

**INTRODUCTION:**

**Terminology:**

* Epidemic: Number of new cases of a certain disease at a point in time that substantially exceeds what is expected based on recent experience. [1]
* Epidemic Datasets:Data about the number of epidemic cases in a particular state at a point in time**.**
* One hop neighbour:Neighbouring state of the state that is under consideration.
* Win:window of length w

**Goal Description:**

Given the epidemic dataset simulation files, the data has to be normalized and quantified to a Gaussian range. The quantised values are then grouped to form a window based on the given window length and by shifting h time units at a time. For each window, the filename, state, time and the win are written to epidemic\_word\_file.

For each word in the epidemic\_word\_file, the winavg and windiff has to be computed by finding the average win value of the neighbouring states. The winavg and windiff are written to epidemic\_word\_file\_avg and epidemic\_word\_file\_diff files respectively along with the corresponding time, state and filename.

The data set values of file selected are then to be plotted on a heat map. The strength of the states has to be found out by computing the norm of win vector. The states with the maximum strength and minimum strength are to be found out and depicted on the heat map, along with their one hop neighbour states.

**Assumptions:**

The following assumptions have been made.

* The datasets given are the ones that belong to the state in the United State of America.
* The row and column range of the datasets are same for all file given in the directory.
* For states with no neighbouring state, the value of win vector of neighbouring states is assumed to be zero.
* All Input files are under the same folder.

**Implementation:**

The project has been implemented using MATLAB .A separate module has been created for each task given in the project description.

**Task1:**

The task1.m program, reads the values from all the files in the given directory. For each data file, the program normalizes the data to values between 0 and 1. In order to quantize the values to Gaussian band, the length of each band is computed based on the given formula and the resolution ‘r’ value that has been provided.

The bands are then formed by adding the successive lengths. Once the bands are formed, the normalized values are then allotted to each band depending on the range which they fall under. The centre value of each band is used as the representative figure. The values for each state are grouped to a window based on the given window length and by shifting ‘h’ time units at a time. Each window is then written to epidemic\_word\_file.csv.

**Task2**

The task2.m program reads the values from epidemic\_word\_file.csv. For each state at time ‘t’, the program computes the average of win vector of its one hop neighbours at the same time t and adds it up with the win value of the state that is under consideration, as per the formula given in the description. The windiff is also computed in a similar way depending on the formula given.

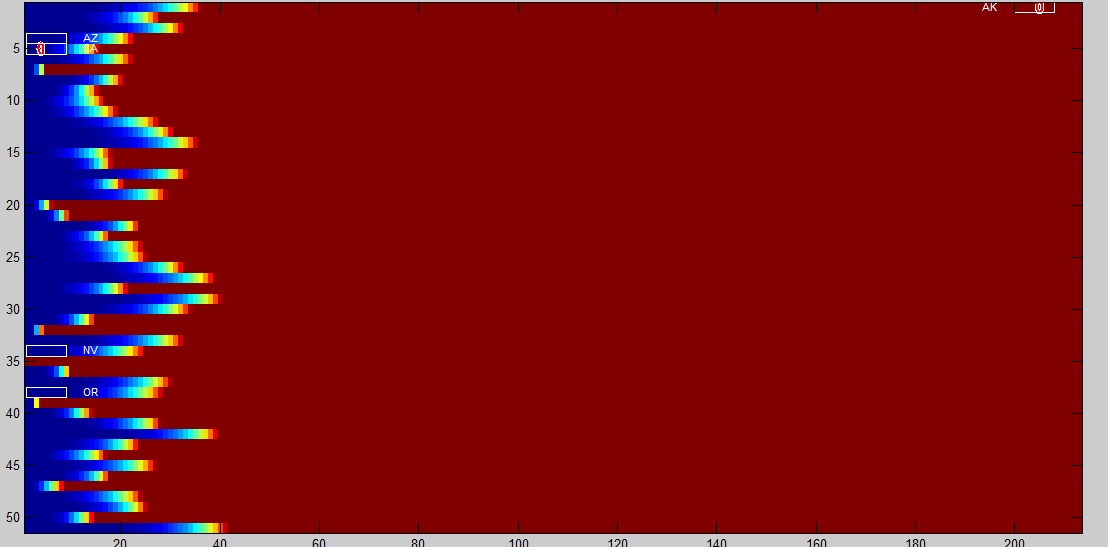
The winavg and windiff values are written to epidemic\_word\_file\_avg.csv and epidemic\_word\_file\_avg.csv respectively.

**Task3**

The task3.m, reads the dataset file f provided by the user and the lets the user to choose between epidemic\_word\_file.csv, epidemic\_word\_file\_avg.csv, and epidemic\_word\_file\_diff.csvreads the file selected by the user. The norm of the win vector is computed. The state of the vector with the maximum norm value is chosen as the maximum strength state and the state of the vector with the minimum norm value is chosen as the minimum strength state. The values of the data set file f are plotted on a heat map.

The heat map also shows the state with the maximum strength and the state with the minimum strength, along with their on hop neighbours.

The output of the task 3 will be as shown below



**Interface Specification**

The task 1 requires the user to enter, the directory of the datasets, window length w, shift length w and the resolution r, through the command prompt. The task 2 allows the user to enter the alpha value. In task 3, the user has to choose the epidemic data set file name, and select one file among the 3 files displayed on the prompt.

**System Requirement:**

The minimum requirements are as follows

* Intel 64-bit i7 Processor
* 8GB RAM

Software Used

* MATLAB R2013B

Execution Instruction

The program can be executed using MATLAB. On the MATLAB command prompt, enter the program name to execute the program.

**CONCLUSION:**

The phase one of the project, that involved creating an epidemic word dictionary using the given simulation datasets has been implemented. The values have been depicted using a heat map. The states that have the maximum and minimum strength have also been found out and represented on the heat map. The future scope of this project would be to extend this implementation to create an interactive heat map the represents the data simulation.

**REFERENCES**

1. Principles of Epidemiology, Second Edition. Atlanta, Georgia: Centers for Disease Control and Prevention :

<https://www.virtualbox.org/manual/UserManual.html> (Accessed: 27th August 2014)

1. MathWorks,(no date), Image function in MATLAB, available at

http://www.mathworks.com/help/matlab/ref/image.html (Accessed: 19th September 2014)

1. MathWorks,(no date), Insert shapes in image, MATLAB, available at

http://www.mathworks.com/help/vision/ref/insertshape.html (Accessed: 19th September 2014)

1. MathWorks,(no date), Vectors and Arrays, MATLAB, available at

http://www.mathworks.com/help/matlab/ref/colon.html (Accessed: 14th September 2014)