A Novel Disease Outbreak Prediction Model for Compact Spatial-Temporal Environments

Kam Kin Lao, Suash Deb, Sabu M. Thampi, and Simon Fong

Abstract. One of the popular research areas in clinical decision supporting system (CDSS) is Spatial and temporal (ST) data mining. The basic concept of ST concerns about two combined dimensions of analyzing: time and space. For prediction of disease outbreak, we attempt to locate any potential uninfected by the predicted virus prevalence. A popular ST-clustering software called “SaTScan” works by predicting the next likely infested areas by considering the history records of infested zones and the radius of the zone. However, it is argued that using radius as a spatial measure suits large and perhaps evenly populated area. In urban city, the population density is relatively high and uneven. In this paper, we present a novel algorithm, by following the concept of SaTScan, but in consideration of spatial information in relation to local populations and full demographic information in proximity (e.g. that of a street or a cluster of buildings). This higher resolution of ST data mining has an advantage of precision and applicability in some very compact urban cities. For proving the concept a computer simulation model is presented that is based on empirical but anonymized and processed data.

1 Introduction

Clinical information system is in imperative need for the human society, especially when people experienced some epidemic diseases like severe acute respiratory
syndrome (SARS), swine flu and enterovirus, etc., which has a high prevalence rate. They outbreak at a very rapid speed, and spread wide and far. There are research papers which advocate developing the clinical decision support system which predicts the time series and space area. However, the efficacy of clinical decision support system is based on the underlying analysis model. Some data related challenges are like: what kind of data attributes the system need to use? How about the scope of data? Is the data useful or not? Which analyzing method is efficient and effective? Any other parameter need to be concerned? What is the trend of disease outbreak? Etc.

Many researchers suggested embedding the clinical decision support system into the GIS (Geographic Information System) as it seems to be more accurate to detect the area whether is in a high prevalence rate and their adjacency areas [1], or even using the ST analyzing method to focus on the analyzing risk of the disease outbreak [2]. Actually these research papers assume the field of analysis is of large terrain or vast piece of land. It is useful for large countries. However, it may not be so applicable for compact urban cities like Macao, Hong Kong, Taipei etc. where the human population is very dense, but they are not necessarily evenly distributed. In the words, the radius approach might not work well in estimating the next infested areas. Different from the other researchers, we introduce a new and simple approach by dividing the city into respective regular polygons, each polygon which is square cell as assumed in this case, is of equal size; and they form grid over the coverage of the city regardless what shape the city is. The number of the square cells to be defined can be selected arbitrarily by the user that depends on the land area and the resolution required.

As a demonstrative case in this paper we use the data simulation of the enterovirus as the experiment part of our research, Macao land and the Tapai land will be separately divided by various numbers of cells and combined for analysis. At the start, the risk of the virus would be evaluated in order to find co-relationship among the areas, the analyzing model will predict how risky of each zone of the city, depends on the risk analyze (some factors may need to reference the previous disease record of the zone and the risk analyzing in the surrounding zones). After locating the high risk areas, the analyzer can group these zones and focus on the relations and/or correlations of them as try to know more about virus and its spread. The associated relationships among the areas are those that have the disease outbreak simultaneously. Technically it will involve using various classifiers of the decision tree and association rules analyzing model.

After applying these two models, some high risk areas and the relationship which the areas almost have the disease outbreak in the same time will be found. The analyzers can concentrate on analyzing the specific characteristic of the areas and deciding which attributes will have the significant relationship between the inflected areas. As the demographic information changes and the risk evaluation suggest, the experiment will vary for different time-series. Finally the analyzers can trace back the source of virus and identify the “flow” of various attributes; and investigate whether the virus has been mutated. This analyzing method will be novel as the part of risk evaluation for detecting among various areas, especially when the forecast of the disease outbreak is changing obviously, the risk index and the associated