Computational Statistics and Data Analysis 89 (2015) 134–146

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An adaptive minimum spanning tree test for detecting irregularly-shaped spatial clusters

Ruoyu Zhou[a], Lianjie Shu[a,*], Yan Su[b]

[a] Faculty of Business, University of Macau, Macau

[b] Department of Electromechanical Engineering, University of Macau, Macau

ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form 2 January 2015
Accepted 10 March 2015
Available online 26 March 2015

Keywords:
Minimum spanning tree
Spatial cluster detection
Arbitrary shape
Validity index

ABSTRACT

The clustering methodologies based on minimum spanning tree (MST) have been widely discussed due to their simplicity and efficiency in signaling irregular clusters. However, most of the MST-based clustering methods estimate the most likely cluster based on the maximum likelihood ratio from the resulting subtrees after the removal of edges of the MST. They can only estimate one cluster even if there are multiple clusters actually present over the study region. To overcome this limitation, we propose an adaptive MST (AMST) method to detect irregularly-shaped clusters. The basic idea is to first determine the best number of partition over the study region using a validity index and then to determine the significance of the candidate clusters. The comparison results with both the static and dynamic MST methods favor the proposed method.

1. Introduction

The detection of disease clusters in space is an important public health problem which has to deal with multiple statistical testing problems. A great number of tests have been proposed to handle this problem. According to Besag and Newell (1991), tests for spatial clusters could be generally classified into two categories: focused tests and general tests, depending on whether the assumption about cluster locations is made or not. Song and Kulldorff (2006) further divided the general tests into two groups: tests for global clustering and tests for localized clusters, depending on the ability to identify the locations of potential clusters. The global clustering tests look for evidence to determine whether there are clusters present over the whole study region without concern about cluster locations (Whittemore et al., 1987; Besag and Newell, 1991; Tango, 1995). On the other hand, tests for localized clusters are concerned with not only testing their statistical significance, but also detecting the location of clusters (Turnbull et al., 1990; Kulldorff, 1997).

Among these tests, the spatial scan test proposed by Kulldorff (1997) is conceptually intuitive and has been widely discussed as a tool for detecting localized clusters. The conventional scan test is to first compute the likelihood ratio statistics based on a circular scanning window with varying radius, and then to determine the area with the maximum likelihood ratio score. The area giving the maximum likelihood ratio is estimated as the most likely cluster. The spatial scan statistic test has been proved to be very powerful when the real cluster has a circular shape. However, the shape of the potential cluster is generally unknown in practice and the shape is generally irregular.

The use of a scanning window of a rigid geometry has been viewed as a limitation of the spatial scan test. To deal with this limitation, some modifications have been made to the spatial scan test, including the search for elongated, elliptical,

[*] Corresponding author.
E-mail address: ljshu@umac.mo (L. Shu).

http://dx.doi.org/10.1016/j.csda.2015.03.008
0167-9473/© 2015 Elsevier B.V. All rights reserved.

and irregular shapes. For example, Neill and Moore (2004) and Kulldorff et al. (2006) proposed the spatial scan statistics for elongated and elliptical-shaped clusters. However, these methods seem to be only slightly less constraining than the spatial scan tests based on circular-shaped clusters. The clusters with less-compact shapes will be overlooked by these methods.

In order to detect arbitrarily shaped clusters, various procedures have been developed. A sample of research in this respect includes the upper level set (ULS) method proposed by Patil and Taillie (2004), the simulated annealing (SA) approach proposed by Duczmal and Assunção (2004), and the *k*-nearest-neighbor method proposed by Tango and Takahashi (2005). The ULS is simple and fast to implement. However, it is shown not to have good performance as compared with other clustering methods. The SA method has the drawback that it requires the setting of awkward tuning parameters. These parameters are not easily interrelated, and there is lack of guidance to do this. The *k*-nearest-neighbor method extends the regular scanning window in the conventional spatial scan test to the flexible shape by using *k*-nearest neighbors. Each cell with its *k*-nearest neighbors at most is considered as a candidate cluster when determining the set of candidate clusters. Clearly, this method suffers from the computation complexity issue as the number of candidate clusters increases exponentially with the value of *k*. This makes it inefficient for large value of *k*. In general, there is no versatile method that can handle all types of clustering problem due to the arbitrary shape, unbalanced background population size, and variable densities.

Another effective clustering method consists of using the graph-based tests, especially the minimum spanning tree (MST), owing to its intuitive and effective data representation. Cluster design using MST was initiated by Zahn (1971) and later discussed by Maravalle and Simeone (1995) and Maravalle et al. (1997). It has now been successfully employed in many different settings, such as in image processing (Wang et al., 2009), pattern recognition (Paivinen, 2005), and biological data analysis (Xu et al., 2002), and public health surveillance (Assunção et al., 2006).

In the context of public health surveillance, Assunção et al. (2006) recently proposed two types of MST-based tests for fast detection of arbitrarily shaped disease clusters: the static MST (SMST) and the dynamic MST (DMST). They showed that the SMST includes the ULS method as a special case. Compared to the SMST, the DMST has larger detection power and is thus suggested for practical use. Note that the MST-based clustering method enjoys two good properties. First, it gets away from the specification of tuning parameters. Second, MST is an extremely economical way to represent the spatial structure of a graph and thus has the capacity of quickly identifying clusters without constraining the possible shapes. Due to these good properties, the MST-based method is more computationally efficient as compared to the SA and *k*-nearest-neighbor methods.

The MST-based methods determine the most likely cluster using the maximization of the likelihood ratio over a set of candidate clusters. Thus, both the SMST and DMST methods can only return/estimate a single cluster. In addition to a single cluster present, multiple clusters can also appear, which are more general in practice. This limitation would restrict the applications of SMST and DMST for clustering analysis. To meet the gap in part, we propose an adaptive MST (AMST) approach, aimed at automatically and simultaneously identifying the locations of clusters with arbitrary shapes. The key is to define a validity index taking both compactness and isolation of data into consideration to help determine best partition of a MST. Based on the clustering results from validity index, one can then evaluate the statistical significance of these candidate clusters.

The rest of this paper will be organized as follows. In Section 2, we briefly review the SMST and DMST methods. In Section 3, we present the AMST method. In Section 4, we compare the detection power and identification capability based on simulations. Both the case with a single cluster and the case with multiple clusters are considered. In Section 5, two simulation examples are provided to illustrate the use of the proposed method. Some concluding remarks are given in Section 5.

2. Static and dynamic MST tests

2.1. Notation

We begin by defining some notation. A whole study region contains I connected locations. For each location i, denote x_i as the number of cases observed, and n_i as its population size. Also, define $X = \sum_i x_i$ and $N = \sum_i n_i$ as the total number of cases and total population size over all the study region, respectively. A Poisson model is often assumed for the data, namely, $x_i \sim$ Poisson $(n_i\lambda_i)$, where λ_i is the disease rate in cell i, which is interpreted as the number of cases per unit population. Assuming that the potential cluster C is a subset of connected areas, define x_C and n_C as the disease count and associate population size inside cluster C respectively, and denote $x_{\overline{C}}$ and $n_{\overline{C}}$ as the respective disease count and population size outside C. Then, $x_{\overline{C}} = X - x_C$ and $n_{\overline{C}} = N - n_C$.

Under the null hypothesis of no cluster, it is explicitly assumed that the incidence rates of disease are all the same across all cells. That is, $H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_I = \lambda_0$. Under the alternative hypothesis H_1 of a cluster C with elevated incidence rate, it is assumed that $\lambda_i = \lambda_C$ $(\lambda_C > \lambda_0)$ for $i \in C$ and $\lambda_i = \lambda_0$ for $i \notin C$. Based on the Poisson model, the likelihood ratio statistic to null hypothesis H_0 and the alternative hypothesis H_1 in a fixed cluster C is

$$lr_C = \left(\frac{x_C}{\mu_C}\right)^{x_C} \left(\frac{X - x_C}{X - \mu_C}\right)^{X - x_C}, \tag{1}$$

where μ_C is the expected number of counts under the null hypothesis, namely, $\mu_C = n_C \cdot X/N$.