Neurocomputing 117 (2013) 161–172

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Spatial outlier detection based on iterative self-organizing learning model

Qiao Cai [a], Haibo He [b,*], Hong Man [a]

[a] Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[b] Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form
23 January 2013
Accepted 1 February 2013
Communicated by K. Li
Available online 16 March 2013

Keywords:
Self-organizing map
Spatial data mining
Spatial outlier detection
Iterative learning
Robust distance

ABSTRACT

In this paper, we propose an iterative self-organizing map (SOM) approach with robust distance estimation (ISOMRD) for spatial outlier detection. Generally speaking, spatial outliers are irregular data instances which have significantly distinct non-spatial attribute values compared to their spatial neighbors. In our proposed approach, we adopt SOM to preserve the intrinsic topological and metric relationships of the data distribution to seek reasonable spatial clusters for outlier detection. The proposed iterative learning process with robust distance estimation can address the high dimensional problems of spatial attributes and accurately detect spatial outliers with irregular features. To verify the efficiency and robustness of our proposed algorithm, comparative study of ISOMRD and several existing approaches are presented in detail. Specifically, we test the performance of our method based on four real-world spatial datasets. Various simulation results demonstrate the effectiveness of the proposed approach.

1. Introduction

With the continuous explosive increase of data availability in many real-world applications, computational intelligence techniques have demonstrated great potential and capability to analyze such data and support decision-making process. In general, there are five primary categories of data engineering research, including classification, clustering, regression, association, and deviation or outlier detection [1]. In this paper, our objective is to investigate the spatial outlier detection based on computational intelligence approaches.

The procedure of outlier detection can be considered as similar to the discovering of “nuggets of information” [2] in large databases. The motivation for this type of research is that in many practical situations, such outliers normally carry the most critical information to support the decision making process. Due to the wide range of application scenarios for outlier detection across different domains [19–21], such as financial industry, biomedical engineering, security and defense, to name a few, outlier detection has been an important research topic in the community for many years. For instance, an anomalous traffic pattern in a computer network might indicate the presence of malicious intrusion from unauthorized users or computers. In public health information systems, outlier detection techniques are widely employed to detect abnormal patterns in physical records that might indicate uncommon symptoms. In all of such situations, once the outliers are identified, they prompt a more focused human analysis to understand those data sets from a vast amount of original raw data.

We would like to point out that, due to the extensive research efforts in the community, there are different terminologies referring to the same or similar idea, such as outlier or anomaly detection [3,4], exception mining [5], mining rare classes [6], novelty detection [7], and chance discovery [8]. Data mining techniques concerning this issue involve both supervised and unsupervised learning paradigms. Generally speaking, supervised learning methods first establish a prediction model for regular and irregular events based on labeled data in the training set, and then make classifications for future test data. One of the shortcomings of such approaches is that they require a representative set of training data with the target function to train the model. Such labeled training data might be difficult or expensive to obtain in many real applications. Unsupervised learning, on the other hand, does not require labeled data. The performance of such approaches depends on the choice of feature selection, similarity measures, and clustering methods.

In this paper, we propose to use self-organizing map (SOM) with robust distance estimation for spatial outlier detection research in [27,28]. Although SOM was proposed a long time ago and there is a rich literature on SOM and related techniques in the community, the use of SOM specifically targeting for spatial outlier detection is a relatively new topic. With the continuous expansion of data availability in many of today's data intensive applications (the Big Data Challenge [45]), we consider the analyze of spatial

* Corresponding author.
E-mail address: he@ele.uri.edu (H. He).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.02.007

data has become more and more critical in many real world applications. Therefore, we hope the proposed SOM-based spatial outlier detection method in this work could provide important techniques and solutions to tackle the spatial data challenge. The major motivation for this approach is to take advantage of the data clustering capability of SOM to effectively detect outliers with both spatial and non-spatial features. Furthermore, to improve the learning and detection performance, we propose an iterative SOM approach with robust distance estimation for improved performance. The rest of this paper is organized as follows. Section 2 briefly introduces the development of related research on this topic. In Section 3, we present our proposed approach for spatial outlier detection. In Section 4, the detailed simulation results and analysis of our method are presented based on the U. S. Census Bureau databases for spatial outlier detection. Finally, we give a conclusion in Section 5.

2. Related work

In general, there are five major categories of approaches for outlier detection in the literature: distribution-based, clustering-based, distance-based, density-based, and depth-based methods. Distribution-based approaches are primarily concentrated on the standard statistical distribution models. Some representative distribution models like Gaussian or Poisson are frequently used to identify outliers that perform irregularly in such models [4]. In clustering-based approaches, the identification of outliers is normally considered as a side product while the primary goal of clustering is to find data cluster distributions [10]. However, these approaches have been successful in many applications, such as the CLARANS [11], DBSCAN [12], and CURE [13] approaches. Distance-based approaches rely on different distance metrics to measure the relationships between data items and to find outliers [14]. Some interesting methods have the capability of calculating full dimensional mutual distances with existing attributes [15,16] or feature space projections [3]. Density-based approaches are based on the analysis of data distribution density, such as the approach in [17], to determine a local outlier factor (LOF) for each data sample based on its corresponding local neighborhood density. In this way, those data samples with higher LOF can be considered as outliers. Finally, depth-based approaches can identify outliers based on geometric computation, which computes distinct layers of k-dimensional convex hulls [18].

An essential characteristic of spatial data analysis is that it involves both spatial attributes such as longitude, latitude and altitude, and associated non-spatial attributes, such as the population density and age distribution of each spatial point. Meanwhile, spatial data appear to be highly correlated. For example, spatial objects with the similar properties seem to cluster together in the neighboring regions. In fact, as discussed in [24], spatial autocorrelation problems involved in spatial dependency occur for all spatial objects when spatial properties are involved. The spatial relationships among the items in spatial datasets are established through a contiguity matrix, which may indicate neighboring relationships, such as vicinity or distance. Given such characteristic of spatial data mining, detection of spatial outliers aims at discovering specific data instances whose non-spatial attribute values are significantly distinct from the corresponding spatial neighbors. Informally speaking, a spatial outlier might be considered as a local instability whose non-spatial attributes are intrinsically relevant to the surrounding items, although they may be obviously distinct from the entire population. There are two major categories of outliers in spatial datasets: multi-dimensional space-based outliers and graph-based outliers [25]. The major difference between them is their spatial neighborhood definitions. Multi-dimensional space-based outliers are based on Euclidean distances, while graph-based outliers follow graph connectivity.

Most of the existing spatial outlier detection algorithms focus on identifying single attribute outliers, and could potentially misclassify normal items as outliers when genuine spatial outliers exist in their neighborhoods with extremely large or small attribute values. In addition, many practical applications involve multiple non-spatial attributes which should be incorporated into outlier detection. There are several reasons that spatial outlier detection still remains a great challenge. First of all, the definition of neighborhood is crucial to the determination of spatial outliers. Additionally, statistical approaches are required to characterize the distributions of the attribute values at various locations compared with the aggregate distributions of attribute values over all the neighboring data.

Several different approaches have been used to improve the classical definition of outlier by Hawkins [30]. Knorr and Ng [31] presented the concept of distance-based outliers for multi-dimensional datasets. Another approach for identifying distance-based outlier is to calculate the distance between certain point and its corresponding k nearest neighbors [33]. The ranked points are identified as outlier candidates based on the distance to its k nearest neighbors. In some specific models, local outliers appear to be more important than global outliers [34]. However, they are normally difficult to be identified using general distance-based techniques. The method based on a local outlier factor (LOF) was proposed to capture the phenomena that a sample is isolated from its surrounding neighborhood rather than the whole dataset. The local correlation integral (LOCI) method was also presented to discover local outliers [35]. This approach seems moderately similar to the LOF except for the definition of the local neighborhood. However, spatial attributes are not considered in these algorithms or approaches for outlier detection.

In spatial datasets, the non-spatial dimensions provide intrinsic properties of each data example, while spatial attributes describe location indices to define neighborhood boundary for spatial outlier detection. Thus, the physical neighborhood plays a crucial role in spatial data analysis. Common techniques for neighborhood characterization include KNN, grid technique, and grid-based KNN. Additionally, a variogram-cloud [22] displays spatial objects based on the neighboring relationships. For each pair of location coordinates, the square-root of the absolute difference between attribute values at the locations compared with the mutual Euclidean distance is depicted. Vicinal locations with significant attribute difference might be considered as a spatial outlier, even though attribute values at such locations may seem to be regular or normal if spatial attributes are ignored.

In our research, we consider a spatial outlier as a “spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood” [21]. The spatial outlier detection in the graph dataset was discussed with detailed algorithms in [21,37]. Two representative methods, i.e. Scatterplot [38] and Moran scatterplot [39], can be employed to quantitatively analyze spatial dataset and discover spatial outliers. A scatterplot illustrates attribute values on the X-axis and the average attribute values in the neighborhood on the Y-axis. A least-square regression line is used to identify spatial outliers. A positive spatial autocorrelation is suggested by the right upward sign for a scatter slope; otherwise, it turns to be negative. The purpose of a Moran scatterplot is to compare the normalized attribute values with the average normalized attribute values in the neighborhoods. In [40], several statistical outlier detection techniques were proposed, and they were compared with four algorithms. The z, iterative z, iterative r, and Median algorithms [36] were successfully used to identify spatial outliers. A measure for spatial local outliers was proposed