Discovery of Anomalous Windows through a Robust Nonparametric Multivariate Scan Statistic (RMSS)

Lei Shi, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA

Vandana P. Janeja, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA

ABSTRACT

This paper studies unusual phenomena by discovering anomalous windows in multivariate spatial data. Such an anomalous window is a group of contiguous spatial objects indicating the occurrence of unusual phenomenon in terms of multiple variables. The paper presents a novel Robust non-parametric Multivariate Scan Statistic (RMSS). In contrast to the existing work, the authors’ approach is designed to deal with anomalous window discovery in multivariate data. They propose their multivariate scan statistic that employs the robust Mahalanobis distance which enables taking into account multiple behavioral attributes at the same time and their correlations for the discovery of significant anomalous windows. The proposed multivariate scan statistic is non-parametric such that it does not rely on any prior assumption about the data distribution. It is robust such that it can handle data with large amount of outliers, up to 50% of the overall data size. It is also affine equivariant such that affine transformation such as stretch or rotation of the data would not affect the results. The authors evaluate their approach with (a) real-world multivariate climate data for discovering natural disasters and climate changes, (b) real-world multivariate traffic accident data for identifying accident hubs, which are route segments with underlying accident-prone issues, and (c) synthetic data of both continuous and discrete multivariate distribution for identifying clusters of known outliers under different outlier percentage in data. They compare their results to state of the art multivariate scan statistic method (Kulldorff et al., 2007). The evaluation shows the detection power of the authors’ method, and the significant improvement over the existing methods.

Keywords: Anomalous Window Discovery, Multivariate Scan Statistic, Multivariate Spatial Data, Robust Non-Parametric Multivariate Scan Statistics (RMSS), Spatial Data

DOI: 10.4018/jdwm.2013010102
INTRODUCTION

Study of unusual phenomena (Shi & Janeja, 2009; Kulldorff, 1997; Schoier & Borruso, 2012) finds use in various applications related to spatial data (Silva, Moura-Pires, & Santos, 2012) such as discovery of (i) disease outbreak in a region (Kulldorff, 1997), (ii) accident hubs along highways (Shi & Janeja, 2009), and (iii) leak of toxins in water or air, to name a few. Timely identification of such unusual phenomena is of utmost important, so that corresponding solutions can be prepared and applied in time to avoid and reduce the relevant personnel and property loss. Such unusual phenomena can be identified as anomalous windows in spatial data. An anomalous window is a group of contiguous spatial objects where the phenomenon takes place. These spatial objects are discovered by being quantified as unusual, in terms of their behavior, with respect to those of the other spatial objects in the data. Traditional approach of outlier detection such as nearest neighbor based methods (Knorr & Ng, 1997) or clustering based methods (Ester, Kriegel, Sander, J., & Xu, 1996) are not very suitable for such a discovery. It is primarily due to the reason that they are designed to identify individual outliers, and do not consider the quantification of the unusualness of such anomalous windows and the necessary spatial relationships between spatial objects in forming anomalous windows, which is the distinct characteristic of the spatial data. In contrast, spatial scan statistic has proven to be a promising technique in quantifying anomalous windows.

The traditional univariate spatial scan statistic (Kulldorff, 1997) discovers anomalous windows in practical applications by studying unusual behaviors of spatial objects in terms of one single attribute of interest. For instance in studying disease outbreaks the attribute of interest is the number of people afflicted by a disease. In studying accident hubs along highways the attribute of interest could be the number of fatalities or alternatively the number of crashes at a particular mile marker. However, such univariate spatial scan statistic methods cannot study multiple aspects of a phenomenon that have multi-dimensional or multi-domain influences.

Studying multiple aspects of a phenomenon and their interaction is very critical as the phenomenon may not necessarily be measured in the one dimension or even the one domain, but could have multi-dimensional and multi-domain influences. For instance the phenomenon of climate change can have influences in temperature, humidity, rainfall, wind, snow etc. at a location. The underlying spatial processes governing such a phenomenon, in this example climate change, as we stated, could influence multiple spatial properties of a region where it takes place. So suppose if an unusual temperature being observed is a result of the climate change, then it will not only affect this aspect of the climate but also influence the other weather patterns in the region, or even other environmental aspects such as crop growth, animal ecosystems and so on. While solely studying the phenomenon of climate change in a single aspect may not be sufficient or even feasible as it may well be hidden or even if visible may appear insignificant on its own, but the comprehensive influences on entire weather patterns, and environmental patterns are much more clear and convincing. Thus to fathom the underlying processes on such a phenomenon that is difficult to study in a single dimension or single domain, one has to look at the multiple relevant influences it has on spatial properties of other attributes within one domain or even other domains, in a unified manner with the analysis centered on space. This can be achieved by extending univariate spatial scan statistic to the multivariate spatial scan statistic for processing multivariate spatial data. Such multivariate spatial data can have attributes describing variety of spatial properties in the region.

Recent work on multivariate scan statistic (Kulldorff et al., 2007; Neill & Cooper, 2010) directly extends the univariate scan statistics
Social Media Mining for Assessing Brand Popularity
[www.igi-global.com/article/social-media-mining-for-assessing-brand-popularity/198973?camid=4v1a](www.igi-global.com/article/social-media-mining-for-assessing-brand-popularity/198973?camid=4v1a)

A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns, Part II: Sequences of Discrete Symbols
[www.igi-global.com/chapter/review-kernel-methods-based-approaches/58672?camid=4v1a](www.igi-global.com/chapter/review-kernel-methods-based-approaches/58672?camid=4v1a)