Computers & Geosciences 35 (2009) 1508–1516

Contents lists available at ScienceDirect

Computers & Geosciences

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

Discovery of feature-based hot spots using supervised clustering

Wei Ding [a,*], Tomasz F. Stepinski [b], Rachana Parmar [c], Dan Jiang [c], Christoph F. Eick [c]

[a] Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125-3393, USA
[b] Lunar and Planetary Institute, 3600 Bay Area Blvd., Houston, TX 77058, USA
[c] Department of Computer Science, University of Houston, Houston, TX 77204-3010, USA

ARTICLE INFO

Article history:
Received 7 September 2007
Received in revised form
6 October 2008
Accepted 11 October 2008

Keywords:
Hot spots
Clustering
Spatial data mining
Mars

ABSTRACT

Feature-based hot spots are localized regions where the attributes of objects attain high values. There is considerable interest in automatic identification of feature-based hot spots. This paper approaches the problem of finding feature-based hot spots from a data mining perspective, and describes a method that relies on supervised clustering to produce a list of hot spot regions. Supervised clustering uses a fitness function rewarding isolation of the hot spots to optimally subdivide the dataset. The clusters in the optimal division are ranked using the interestingness of clusters that encapsulate their utility for being hot spots. Hot spots are associated with the top ranked clusters. The effectiveness of supervised clustering as a hot spot identification method is evaluated for four conceptually different clustering algorithms using a dataset describing the spatial distribution of ground ice on Mars. Clustering solutions are visualized by specially developed raster approximations. Further assessment of the ability of different algorithms to yield hot spots is performed using raster approximations. Density-based clustering algorithm is found to be the most effective for hot spot identification. The results of the hot spot discovery by supervised clustering are comparable to those obtained using the G* statistic, but the new method offers a high degree of automation, making it an ideal tool for mining large datasets for the existence of potential hot spots.

1. Introduction

Spatial datasets abound in geosciences, making it difficult for the research community to turn all this data into knowledge. One solution is to apply spatial data mining techniques to geospatial datasets in order to automatically discover interesting relations or places that may exist in the dataset. Existing works on spatial data mining (Koperski and Han, 1995; Munro et al., 2003; Huang et al., 2004, 2006; Zhang et al., 2004) tend to focus on discovering systematic relations between spatial variables. For example, in a spatial co-location problem (Huang et al., 2004, 2006; Zhang et al., 2004), the goal is to find subsets of features that are located together in spatial proximity, throughout the spatial extent of the dataset. In other words, the goal is to discover globally valid proximity relationships between certain features. On the other hand, less attention has been given to the discovery of feature-based hot spots in spatial datasets. The term "hot spots" is most often used to describe clustered point-event patterns. Such hot spots are determined only by arrangement of the objects' spatial coordinates without taking into account the attribute values of the data. However, in this paper, we are concerned with feature-based hot spots—localized regions of high or low attribute values.

Multi-feature hot spots—places where multiple features attain values from the tails of their respective distributions—are of special interest. Multi-feature-based hot spots are interesting because they may indicate a rare local process or an unlikely set of circumstances that forces several features to have non-average values sync with one another. For example, in a dataset describing spatial distribution of temperature, humidity, and vegetation cover across a given area, a region may be found that is characterized by high temperature, low humidity, and dense vegetation cover. Such a place is worthy of closer examination, as this particular combination of variables is unexpected (vegetation is usually poor in hot and dry places). Closer examination may reveal an *a priori* unknown factor that accounts for such a combination (possibly irrigation).

The presently popular method of finding feature-based hot spots in spatial datasets relies on the G* statistic (Getis and Ord, 1992; Ord and Getis, 1995). The G* statistic detects local pockets of spatial association. The value of G* depends on an *a priori* given scale of the packets and is calculated for each object individually. Graphical visualization of the results of G* calculations reveals hot spots (aggregates of objects with values of G* higher than expected) and cold spots (aggregates of objects with values of G* lower than expected). Note that such aggregates are not formally-defined clusters, as the G*-based method has no built-in

* Corresponding author.
E-mail addresses: ding@cs.umb.edu (W. Ding), tstepinski@lpi.usra.edu (T.F. Stepinski), rparmar@uh.edu (R. Parmar), djiang@uh.edu (D. Jiang), ceick@uh.edu (C.F. Eick).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.10.007

clustering capabilities. Instead, hot spots are inferred from visualization, utilizing the ability of the human brain to isolate "clusters" in an image.

Our proposed method offers an alternative approach for identification of hot spots. It does not rely on local statistics; instead, it is rooted in data mining methodology, and, in particular, takes advantage of the notion of supervised clustering. Supervised clustering (Eick et al., 2004) uses a fitness function in order to maximize the purity of the clusters. The fitness function is constructed to reward aggregated objects having non-average values of their features. When subjected to such a fitness function, the clustering procedure is guided toward a solution that emphasizes hot spots. Thus, in our method hot spots are identified as formal clusters of objects—visualization is not necessary for their recognition. This makes our method especially useful in the context of automated mining of large datasets for identifying potentially interesting hot spots. For example, in the presence of multiple features, our method can be set up to compile a database of all possible hot spots, including hot spots of individual features, all combinations of double-feature hot spots, etc.

Methods of finding geometrically defined hot spots have been investigated in the past both explicitly and implicitly. Because the geometrically defined hot spots are clusters with respect to spatial coordinates, their detection lies at the heart of spatial data mining and has been investigated in Murray and Estivill-Castro (1998), Openshaw (1998) and Miller and Han (2001). More explicitly, detection of hot spots using a variable resolution approach (Brimicombe, 2005) was investigated in order to minimize the effects of spatial superposition. In Tay et al. (2003), a region-growing method for the discovery of hot spots was described, which selects seed points and then grows clusters from these seed points by adding neighboring points as long as a density threshold condition is satisfied. Definition of hot spots was extended in Williams (1999) and Kulldorff (2001) to cover a set of entities that are of some particular, but crucial, importance to the experts. This is a feature-based definition, somewhat similar to, but less specific than, what we are using in the present paper. This definition was applied to relational databases of spatio-temporal domain to find important nuggets of information. As mentioned earlier, an approach to identifying feature-based hot spots based on local statistics was developed in Getis and Ord (1992) and Ord and Getis (1995). Finally, in Eick et al. (2006), feature-based hot spots are defined in a similar sense, as in this paper, but their discovery is limited to single-feature datasets.

The overall framework of using the concept of supervised clustering for the identification of hot spots is presented in Section 2.1. Section 2.2 presents a description of four conceptually different clustering algorithms considered for use within the supervised framework. We report on the effectiveness of our method in Section 3 by evaluating a case study pertaining to the spatial distribution of ground ice on Mars. The optimal clustering solutions are subjected to detailed statistical analysis, with the aim of identifying the clustering algorithm best suited to the task of finding hot spots. In Section 4, we present an ancillary method aimed at transforming a clustering solution into a segmentation solution. The difference between a cluster and a segment is that whereas a cluster is a set of objects, a segment is defined as a polygon that has an area and transparent neighborhood relations with other segments. Thus, a segmentation solution can be subjected to additional statistical analysis that is not practical for a clustering solution; the result of such an analysis allows for additional discrimination between different clustering algorithms. For the end user, the segmentation solution provides more effective visualization, and facilitates a query of identified hot spots by additional attributes related to the properties of the area they occupy. Discussion and future work directions are given in Section 5.

2. Supervised clustering methodology

2.1. Framework

The relevant dataset consists of point objects, each characterized by a list of real-valued features. The basic tenet of our approach is to use a clustering algorithm to divide a dataset O into a set of clusters $X = \{c_1, \ldots, c_k\}$, $c_i \subseteq O$, in such a way as to maximize a fitness function $q(X)$. The clusters are disjointed and contiguous but not exhaustive; some objects in O may not be assigned to any cluster. The number of clusters, k, is either set *a priori* or the best value of k is determined by clustering algorithms, depending on the capabilities of clustering techniques.

For the task of hot spot identification, the fitness function must be constructed to reward isolation of hot spots. We propose the following fitness function q:

$$q(X) = \sum_{c \in X} (i(c) \times \|c\|^{\beta}) \tag{1}$$

where $i(c)$ is the interestingness measure of a cluster c—a quantity designed to reflect the degree to which clusters can be considered hot spots. The region "size" (number of objects in the cluster) is denoted by $\|c\|$, and the quantity $(i(c) \times \|c\|^{\beta})$ is a "reward" given to a cluster c. A cluster reward is proportional to its interestingness, but a bigger cluster receives a higher reward than a smaller cluster having the same value of interestingness to reflect a preference given to larger clusters. The premium put on the size of the cluster is controlled by the user-determined value of the parameter $\beta > 0$. We seek a clustering solution X such that the sum of rewards over all of its constituent clusters is maximized.

2.2. Interestingness of clusters

An entry in a geospatial dataset has the form $(\langle spatialcoordinates \rangle, \langle feature_1 \rangle, \ldots, \langle feature_m \rangle)$, where m is the number of features. The numerical values of the features come from their respective distributions, which could have quite different functional forms. Therefore, it is necessary to normalize the values of different features to a common meaning. The two most important properties of any distribution are its center (S), which indicates the location of the bulk of the data, and its scale (σ), which indicates dispersion around the center. For features having bell-shaped distributions, S and σ are easily estimated using the mean and standard deviation, respectively. However, mean and standard deviation are biased estimates of S and σ for features with skewed distributions of their values. Thus, in general, S and σ should be calculated using more robust statistical estimators. For S, a robust estimator is the trimmed mean calculated by discarding a certain percentage of the lowest and the highest values. Note that the median is a particular example of the trimmed mean. For σ, a number of robust estimators are used, including the median absolute deviation (MAD) as well as S_n and Q_n estimators introduced by Rousseeuw and Croux (1993).

Regardless of the method used to estimate S and σ, the feature values are transformed into their z-scores, $z_j = (x_j - S_j)/\sigma_j$, $j = 1, \ldots, m$. Strictly speaking, the term "z-score" is used only in the context of S being the mean and σ being the standard deviation, but an extension of that term to data normalization using any estimate of the center and the scale is quite natural. The z-score is the number of scales of a given feature value above or below its center. In this case, the centers of all features are transformed to 0; the positive values of z indicate upward