IterativeSOMSO: An Iterative Self-organizing Map for Spatial Outlier Detection

Qiao Cai[1], Haibo He[2], Hong Man[1], and Jianlong Qiu[2,3]

[1] Department of Electrical and Computer Engineering,
Stevens Institute of Technology, Hoboken, NJ 07307, USA
[2] Department of Electrical, Computer, and Biomedical Engineering,
University of Rhode Island, Kingston, RI 02881, USA
[3] School of Science, Linyi Normal University, Linyi 276005, China
{qcai,Hong.Man}@stevens.edu, he@ele.uri.edu, qjllinyi@yahoo.com.cn

Abstract. In this paper, we propose an iterative self-organizing map approach for spatial outlier detection (IterativeSOMSO). IterativeSOMSO method can address high dimensional problems for spatial attributes and accurately detect spatial outliers with irregular features. Detection of spatial outliers facilitates further discovery of spatial distribution and attribute information for data mining problems. The experimental results indicate our proposed approach can be effectively implemented for the large spatial dataset based on U.S. Census Bureau with approving performance.

Keywords: Neural network; Self-organizing map; Mahalanobis distance; Spatial data mining; Spatial outlier.

1 Introduction

Data mining, as a crucial technique in many of today's data intensive applications, aims to extract implicit and useful knowledge from large-scale arbitrary datasets. Among most data mining techniques, the procedure of outlier detection is similarly compared with discovering "nuggets of information" [1] in the large databases. In many situations, the outlier normally carries the important information. However, spatial outlier detection [2] still remains challenging and controversial for several reasons. Firstly, the definition of neighborhood is crucial to determine spatial outliers. Secondly, the statistical approaches for spatial outliers are required to illuminate the distribution of the attribute values for variety of locations compared with the aggregate distribution of attribute values over the all neighboring clusters [3].

In our previous research work, SOMSO [5][7] was proposed by integrating self-organizing map (SOM) [4] with Mahalanobis distance [6] to detect the spatial outliers. The advantage of SOMSO approach is that it can not only reduce data dimensions, but more importantly, the topological information of spatial location can be preserved to accurately seek similar spatial relationship in large databases. In this paper, we extend the SOMSO approach to be an iterative approach, the

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part I, LNCS 6063, pp. 325–330, 2010.

IterativeSOMSO method, to improve its efficiency and robustness. The key idea of IterativeSOMSO method is to use the iterative SOM mechanism to effectively determine the neighbor sets to reduce the influence of potential local outliers. Experiment results based on the U.S. Census Bureau database [8] demonstrate the effectiveness of this approach.

The rest of this paper is organized as follows. Section 2 presents the detailed IterativeSOMSO algorithm. In section 3, the detailed simulation analysis of this method is illustrated based on the U.S. Census Bureau databases for spatial outlier detection. Finally, we give a conclusion in section 4.

2 The Proposed Method: IterativeSOMSO

The proposed IterativeSOMSO algorithm can effectively detect spatial outlier with multiple spatial and non-spatial attributes. In this approach, we adopt the Mahalanobis distance concept to determine the threshold for identifying spatial outliers with multiple non-spatial attributes. With iterative utilization of SOM, the neighbor set can be effectively updated to eliminate the influence of potential local outliers for more robust detection.

[IterativeSOMSO Algorithm]

1. Given the spatial dataset $x = \{x_1, x_2, \ldots, x_n\}$ in a space with dimension $p \geq 1$, attribute function f with dimension $q \geq 1$.
2. Normally standardize the non-spatial attribute $f(x)$, i.e., $f(x) \leftarrow \frac{f(x)-\mu_f}{\sigma_f}$.
3. For each spatial point x_i, calculate the neighbor set $N(x_i)$ via SOM as following steps:
   (a) Initialize weight vectors with random small values, for $j = 1, 2, \ldots, N_{neuron}$, where N_{neuron} denotes the number of neurons in the lattice.
   (b) Randomly select a sample from the input data space.
   (c) Search best-matching (winning) neuron at each time iteration through minimum Euclidean distance.
   (d) Update the synaptic weight vector of all neurons.
   (e) Recursively implement step (b) until convergence of the feature map.
4. Compute the neighborhood function $g(x_i)$= average or median of the dataset, and comparison function $h(x_i) = f(x_i) - g(x_i)$.
5. Calculate Mahalanobis distance MD_i as (1).

$$MD_i = \sqrt{(x_i - \mu)^T S^{-1} (x_i - \mu)} \tag{1}$$

   where $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$, $S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$
6. Search largest Mahalanobis distance and its corresponding spatial index, then remove non-spatial feature of this data item.
7. Iteratively implement Step (4)-(6) until top M outlier candidates emerge.

8. Sort by Mahalanobis distance of the top M outlier candidates in descending order.
9. Let $\chi_q^2(\beta)$ denote chi-square distribution with certain confidence level β, If $MD_i^2 > \chi_q^2(\beta)$, x_i can be identified as a spatial outlier candidate.

The key idea of the proposed IterativeSOMSO algorithm is that distinctive properties of SOM is provided to determine how to organize synaptic weight vectors to represent the original spatial attributes to obtain spatial clusters. SOM can be considered as a special class of artificial neural networks based on competitive learning. The output neuron is essentially the winning neuron placed on the nodes of the lattice (usually one or two dimensions). For various input patterns, the neurons can be selectively adjusted or updated to adapt the competitive learning process. Briefly speaking, the learning SOM involves these stages: competitive phase, cooperative phase and adaptive phase. The principle goal of SOM is to project the input vector with higher dimensions into one or two dimensional discrete map in topologically ordered pattern, which can be effectively used to identify the neighbor set to facilitate spatial outlier detection. In this paper, we also compare the proposed method with the existing technique such as Grid based KNN method [10]. Based on the combination of Grid and KNN method, Grid based KNN approach might improve efficiency to find neighborhood in lower dimension, but it fails to apply this method in spatial outlier detection with high accuracy when spatial attribute dimension increases. A brief description is summarized as follow.

[Grid based KNN method]

1. Construct the specific grid for spatial data.
2. Find the grid index of each spatial point x_i.
3. Those who have the same grid index share the neighborhood relationship.
4. If the number of data points within the same grid is greater than k, search k-nearest neighbors for x_i and then update the neighbor set. Otherwise, keep the results of the neighbor set in step 3.

---

3 Simulation Result Analysis

The "house" dataset, primarily focused on the housing units and building permits in the United States, collects the detailed information about the housing or building ownerships and distribution density. The non-spatial attributes with 5 dimensions include house units in 2000, house units net change percentage from 2000 to 2005, house units per square mile of land area in 2005, housing units in owner-occupied percentage in 2000 and housing units in multi-unit structures percentage in 2000. The experiment shows that SOM approach is provided as an effective tool to detect spatial outliers. It differs from the traditional machine

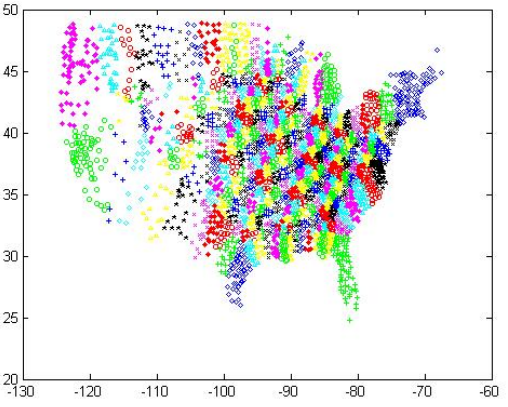

Fig. 1. Spatial clusters: The data points in the clusters with continuously identical marks and colors share common spatial properties in the neighborhood

learning and data mining techniques used in searching spatial neighboring clusters, which are merely concentrated on Euclidean distance for spatial attributes. However, competitive learning promotes synaptic neurons to collaborate with each other and adaptively form feature map with the property of topological ordering. By visual geometric computation in Fig. 1, the proposed method can ultimately acquire important information of the inherent connection for spatial data mining.

Table 1. The top 20 spatial outlier candidates for house dataset detected by Iterative-SOMSO algorithm

Rank	County	Mahalanobis Distance	Units 2000	2000-2005 units net change %	Units per square mile 2005	Units in owner-occupied % in 2000	Units in multi-unit structures % in 2000
1	New York, NY	47.0005	6.8489	0.5159	44.77	7.1244	9.2892
2	Los Angeles, CA	30.1396	29.1008	0.6165	0.8977	3.4559	3.2217
3	Cook, IL	20.1359	18.5294	0.5892	2.7144	2.1363	4.5226
4	Kings, NY	16.6126	8.0432	0.6668	16.7425	6.2006	7.9136
5	Bronx, NY	15.0883	4.0819	0.5752	14.8914	7.1903	8.1802
6	Maricopa, AZ	12.6488	10.9183	2.0930	0.0658	0.8695	1.5582
7	Harris, TX	12.5962	11.3482	1.0502	0.9227	2.4794	2.6778
8	Flagler, FL	11.3892	0.1134	9.0886	0.0347	1.3078	0.2440
9	Hudson, NJ	10.4354	1.8318	0.4953	6.5289	5.7256	7.7003
10	San Francisco, CA	10.3692	2.7849	0.5620	9.4251	5.1582	5.9408
11	Chattahoochee, GA	10.3348	0.3036	0.8650	0.1198	6.2138	1.0996
12	Queens, NY	9.8641	7.0208	0.6698	9.4402	4.1289	6.0901
13	Paulding, GA	9.4858	0.0704	7.2627	0.0390	1.6773	0.7771
14	Suffolk, MA	9.1950	2.2989	0.8840	6.1808	5.3033	7.4231
15	Loudoun, VA	9.0758	0.2259	7.3439	0.0893	0.7008	0.2786
16	King, TX	8.8654	0.3319	0.4893	0.1364	5.2505	1.1610
17	Kenedy, TX	8.6339	0.3309	0.4937	0.1364	5.1846	0.6198
18	Rockwall, TX	8.4423	0.1953	6.3160	0.0792	1.1363	0.2546
19	Dallas, TX	8.2839	7.3522	0.2013	1.1706	2.8357	2.9764
20	Eureka, NV	8.0410	0.3242	0.3352	0.1363	0.0513	0.8944

Table 2. The top 20 spatial outlier candidates for house dataset detected by Grid based KNN algorithm

Rank	County	Mahalanobis Distance	Units 2000	2000-2005 units net change %	Units per square mile 2005	Units in owner-occupied % in 2000	Units in multi-unit structures % in 2000
1	New York, NY	47.1676	6.8489	0.5159	44.77	7.1244	9.2892
2	Los Angeles, CA	33.1943	29.1008	0.6165	0.8977	3.4559	3.2217
3	Cook, IL	19.9999	18.5294	0.5892	2.7144	2.1363	4.5226
4	Kings, NY	17.2194	8.0432	0.6668	16.7425	6.2006	7.9136
5	Bronx, NY	15.6216	4.0819	0.5752	14.8914	7.1903	8.1802
6	Harris, TX	12.4259	11.3482	1.0502	0.9227	2.4794	2.6778
7	Maricopa, AZ	12.3113	10.9183	2.0930	0.0658	0.8695	1.5582
8	Flagler, FL	11.2978	0.1134	9.0886	0.0347	1.3078	0.2440
9	Hudson, NJ	10.8455	1.8318	0.4953	6.5289	5.7256	7.7003
10	Queens, NY	10.6102	7.0208	0.6698	9.4402	4.1289	6.0901
11	San Francisco, CA	10.1129	2.7849	0.5620	9.4251	5.1582	5.9408
12	Chattahoochee, GA	10.0224	0.3036	0.8650	0.1198	6.2138	1.0996
13	San Diego, CA	9.9802	9.0266	0.1974	0.1967	2.4662	2.4646
14	Suffolk, MA	9.9380	2.2989	0.8840	6.1808	5.3033	7.4231
15	Loudoun, VA	9.0355	0.2259	7.3439	0.0893	0.7008	0.2786
16	Henry, GA	9.0182	0.0551	7.2163	0.1149	1.4662	0.3719
17	Paulding, GA	8.9215	0.0704	7.2627	0.0390	1.6773	0.7771
18	Orange, CA	8.8396	8.3907	0.1501	1.4840	1.6744	2.2620
19	Kenedy, TX	8.5330	0.3309	0.4937	0.1364	5.1846	0.6198
20	Alexandria, VA	8.4137	0.2447	0.1052	5.5310	4.4984	5.5250

To better visualize spatial clusters, the topological information will be shown by the cluster density, which can illustrate the number of spatial data items on each neuron. The analysis of cluster density can help us to understand the quantity of spatial data with similar spatial patterns. Besides, the histogram of spatial clusters is employed to display the neighborhood based on the feature map as Fig. 2.

Table 1 and Table 2 illustrate the top 20 ($M = 20$) spatial outlier candidates for house dataset detected by the proposed IterativeSOMSO and the Grid-based KNN algorithm, respectively. From these two tables one can see that both methods can provide comparable results. Since the iterative procedure in IterativeSOMSO can eliminate unknown influence arising from local outliers, we believe

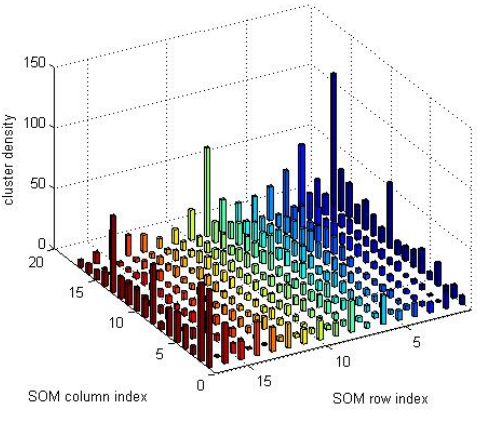

Fig. 2. Histogram of the cluster distribution

the outliers detected by the IterativeSOMSO approach might be more reliable. In terms of computational cost, Grid based KNN will have large computational cost when the dimension of spatial attribute is high. Based on the simulation results, we hope that IterativeSOMSO algorithm may provide an effective approach for such challenging spatial outlier detection applications.

4 Conclusion

In this work, we propose an IterativeSOMSO approach for spatial outlier detection. Experimental results and comparative analysis illustrate the effectiveness of this method. There are a few interesting future directions along this topic. For instance, theoretical analysis of the propose method in terms of convergence is critical to understand the fundamental mechanism of this approach. Also, large-scale experiments and comparative study are necessary to fully justify the effectiveness of this approach. Furthermore, the computational cost of this approach should also be investigated from both a theoretical and empirical point of view. We are currently investigating all these issues and will report their results in the future. Motivated by our results in this paper, we believe the IterativeSOMSO method might be a powerful technique for spatial outlier detection with multiple attributes.

References

1. Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons Ltd, Chichester (2004)
2. Shekhar, S.: Spatial Databases: A Tour. Prentice-Hall, Englewood Cliffs (2003)
3. Shekhar, S., Zhang, P., Huang, Y., Vatsavai, R.: Trends in spatial data mining. In: Data Mining: Next Generation Challenges and Future Directions, pp. 357–380. AAAI/MIT Press (2003)
4. Kohonen, T.: Self-organizing Maps. Springer, Heidelberg (2001)
5. Cai, Q., He, H., Man, H.: SOMSO: A Self-Organizing Map Approach for Spatial Outlier Detection with Multiple Attributes. In: Proc. Int. Joint Conf. on Neural Networks, pp. 425–431 (2009)
6. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining, pp. 276–277. The MIT Press, Cambridge (2001)
7. Cai, Q., He, H., Cao, Y.: Learning from Spatial Data: A Self-Organizing Map Approach for Spatial Outlier Detection. In: Proc. Int. Conf. on Cognitive and Neural Systems (2009)
8. U.S. Census Bureau, United States Department of Commerce, `http://www.census.gov`
9. Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J.: Engineering Applications of the Self-Organizing Map. Proc. of the IEEE 84, 1358–1384 (1996)
10. Lu, C., Chen, D., Kou, D.: Detecting Spatial Outliers with Multiple Attributes. In: Proc. of 15th IEEE Int. Conf. on Tools with Artificial Intelligence, pp. 122–128 (2003)