Spatial Categorical Outlier Detection: Pair Correlation Function Based Approach

Xutong Liu, Feng Chen, Chang-Tien Lu
Department of Computer Science, Virginia Tech
{xutongl,cfeng,ctlu}@vt.edu

ABSTRACT

Spatial Categorical Outlier Detection (SCOD) has attracted considerable attentions from the areas of spatial data mining and geological analysis. When encountering an SCOD problem, some researchers introduce to utilize Spatial Numerical Outlier Detection measures by mapping categorical attributes to continuous ones. However, such approaches fail to capture the special properties of spatial categorical data, which is prone to incur the masking and swamping issues. In this paper, we model spatial dependencies between spatial categorical observations and propose a Pair Correlation Function(PCF) based method to detect SCOs. First, a new metric, named Pair Correlation Ratio(PCR), is estimated for each pair of categorical combinations based on their co-occurrence frequency at different spatial distances. Then discrete PCRs are fitted in a continuous function of distances. The outlier score is computed using the average PCRs between referenced object and its spatial neighbors. Observations with the lowest PCRs are labeled as potential SCOs. Extensive experiments demonstrated that PCF based method outperformed existing approaches.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Application-Data Mining, Spatial Databases

General Terms

Algorithms, Design

Keywords

Spatial categorical data, Pair correlation, Outlier detection

1. INTRODUCTION

With the ever-increasing volume of spatial categorical data, identifying hidden but potentially interesting patterns of anomalies has attracted considerable attentions.Spatial Categorical Outlier(SCO) analysis, which aims at detecting abnormal objects in spatial context, becomes one of the important data mining branches. The identification of SCOs can help extract important knowledge in many applications, including geological data analysis, meteorological knowledge mining, satellite image analysis, and hotspot identification.

During past decades, numerous Traditional Categorical Outlier Detection (TCOD) algorithms (e.g., [5, 6, 10])have been proposed, which may not be appropriate in spatial context.TCOD is determined by global differences and does not consider spatial relationship when identifying anomaly patterns. As indicated by the geographic rule of thumb, "*Nearby things are more related than distant things* [14]," requires more consideration on spatial autocorrelation in spatial analysis. In this sense, spatial outliers can be recognized as local outliers since they are determined by local "abnormal behavior." Most existing spatial outlier detection approaches(e.g., [1, 2, 13]) concentrate on numerical data. Actually, the non-spatial attributes of spatial data are usually category-typed, where sometimes attributes have no intrinsic order information. A typical example is Rock type whose values may include *Igneous*, *Sedimentary*, and *Metamorphic*. Such special property makes anomaly detection in categorical domain more complicated than that in numerical one.

When encountering categorical data, some introduce Spatial Numerical Outlier Detection(SNOD) methods by directly mapping categorical attributes to continuous ones. There exist several critical issues: 1)**Mis-utilization**: statistically, the concept of SCO is different with that of Spatial Numerical Outlier(SNO). Although both of them focus on the identification of abnormal behaviors, SCO is determined by co-occurrence infrequencies, while SNO is determined by numerical distances; 2)**Complicated function**: the mapping process is absolutely not straightforward, especially for nominal attributes; 3)**Swamping and masking issues**: without estimating relevance correctly, some true outliers may be missed and normal ones misclassified as outliers.

To capture the co-occurrence frequency, Pair Correlation Function(PCF) has been proven to be very effective [9]. In this paper, we investigate its benefits on SCOD and design corresponding algorithm. As the first paper that focuses on SCOD, the key contributions include:

Definition of SCO: A SCO is defined as a spatial observation which occurs infrequently with regard to its spatial neighbors.

Design of PCF-SCOD algorithm: Using the concepts of PCF, relevance values are computed for each pair of categories at different specified distances, which help estimate the outlying values for spatial objects.

Extensive experiments: The experiments on *3* real datasets demonstrated proposed method outperformed ex-

ACM SIGSPATIAL GIS '11, November 1-4,2011.Chicago,IL,USA
Copyright 2011 ACM ISBN 978-1-4503-1031-4/11/11 ...$10.00.

isting competing algorithms on accuracy.

2. PRELIMINARY CONCEPTS

This section first introduces PCF techniques, and then summarizes some key notations used in this paper.

2.1 Pair Correlation Function

In mathematical mechanics, PCF $g(r)$ is defined as the observed probability of finding an object at a given distance, r, from a fixed reference particle [12]. The mathematical definition of $g(r)$ is

$$g(r) = \frac{d_{n(r)}/N}{d_{v(r)}/V} = \frac{d_{n(r)}}{d_{v(r)}} \cdot \frac{V}{N} = \frac{d_{n(r)}}{4\pi r^2 d_r} \cdot \frac{V}{N} \tag{1}$$

Where N and V denote the number of units and volume of the entire system, respectively; $d_{n(r)}$ and d_v represent those of the shell region; r is the distance from reference unit to the shell of interest. The relevances among spatial objects are determined by the frequency of co-occurrence of a pair of categories at specific distances. PCF is capable of estimating how observations are packed together, which could be utilized to capture the relationship among objects.

2.2 Preliminary Definition

To formalize the task of SCOD, we need to understand some basic definitions.

Definition 2.1 (Spatial Categorical Dataset) *Let X denote a spatial location on a domain S of the d dimensional Euclidean space R^d. Let A be the categorical attribute and D non-empty set over this attribute.A set $\mathcal{D} \subseteq S \times D$ is called a spatial categorical dataset over S and D. Each record $r_i \in \mathcal{D}(i \in 1, ...n)$ can be denoted as a vector $(r.X, r.A)\prime$.*

Categorical attributes can be classified into two types: ordinal and nominal. The key characteristic of nominal attributes is that different values an object can take are absolutely not inherently ordered. This paper is focused on the dataset that consists solely of nominal attributes.

The anomalous behavior in spatial domain can be truly captured by the local difference, which is determined by the relevances between a specific object and its spatial neighbors. In the paper, k-Nearest Neighbor (kNN) is utilized to construct spatial neighborhood.

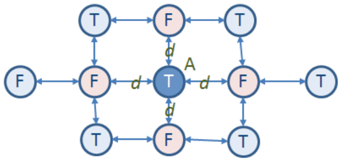


Figure 1: An example for differentiating SNO and SCO

Definition 2.2 (Spatial Neighborhood) *Given a dataset $\mathcal{D}$ with n points, parameter k and for $r_i \in \mathcal{D}$, its spatial neighborhood is estimated by the top k points according to its spatial Euclidean distance vector with the rest of observations in the dataset such that $\forall j \in 1, ..., n$, $j \neq i$, $r_j \in kNN(r_i) : d^E(r_i, r_j) \leq d^E_k(r_i)$.*

In numerical domain, an outlier is defined as the one whose non-spatial attribute is significantly different with those of its neighbors. Such definition is not applicable in categorical domain. As shown in Fig. 1, based on SNOD approach, object A will be recognized as an outlier since it has the attribute of T which is very different with its neighbors', F. However, the contrary is the case in categorical domain. This is because the pair of attributes, $< T, F >$ or $< F, T >$ occurs frequently with the spatial distance of d. Object A should be treated as a normal observation.

Definition 2.3 (SCO) *Let r_i be an observation in $\mathcal{D}$ and r_{i-1},...,r_{i-k} be its spatial neighbors. Its outlierness, for $k \geq 1$,is defined as*

$$OutScore(r_i) = -\sum_{j=1}^{k} PCR(r_i, r_{i-j})/k \tag{2}$$

r_i will be considered as an outlier if it belongs to the top l objects which have the highest outliernesses. $PCR(r_i, r_{i-j})$ denotes the frequency of each pair of category pair at a specified distance. In one word, an SCO is an observation which has lower co-occurrence frequency compared with its spatial neighbors.

Section 3 will discuss PCR function in detail.

3. PCF-SCOD APPROACH

An SCO is defined as an observation which rarely co-occurs with its neighbors. Here, PCF is utilized to extract the local "abnormal behaviors." The main components in PCF-SCOD are described as follows.

PCR estimation: PCR is computed for each pair of categories at different distances. With the set of discrete points in a 2-D space, determined by distance against PCR, a continuous PCR function can be statistically learned, which help estimate the PCRs among spatial objects.

Outlierness computation: The spatial neighbors of an object can be formed using kNN. And the outlying degree of each object is determined by the mean of PCRs between itself and its spatial neighborhood.

Outlier identification: Finally, the outlier scores are ranked in an descending order and top l objects are identified as SCOs.

The first component is crucial since it determines the estimation quality of the relevances among observations. Subsection 3.1 introduces PCR function particularly.

3.1 PCR Function

For each random variable r, $r.A$ is a multilevel categorical variable taking values in $\mathcal{L} = A^1, ..., A^L$. We denote Equation(3) as the frequency of observing category A^l in the dataset,

$$Freq(A^l) = P[A(r_i.A) = A^l] = \frac{n^{A^l}}{n} \tag{3}$$

where n^{A^l},n represent the numbers of objects whose non-spatial attribute is A^l and the objects in the whole dataset, respectively.

Definition 3.1 (Pair Correlation Ratio-PCR) *Considering a spatial pair correlation process in which there are two observations, r_i,r_j in $\mathcal{D}$, each of them is tagged with one category, A^l and $A^{l'}$, respectively. The PCR of r_i,r_j is defined as the normalized pair frequency of the pair of categories, $< A^l, A^{l'} >$, happen to occur at r_i and r_j.*

The mathematical definition of PCR is

$$PCR(r_i, r_j) = \frac{PF(< A^l, A^{l'} >, d^E(r_i.X, r_j.X))}{Freq(A^l) \cdot Freq(A^{l'})} \tag{4}$$

As shown in Equation(4), $PCR(r_i, r_j)$ is only determined by the co-occurrence frequency of categories objects takes and their spatial Euclidean Distance, not their specific spatial locations.

We calculate the pair frequency by the following procedures:

- **Distance bin formulation**: Compute the spatial distances among objects, identify b small bins based on the maximum and minimal distances, whose sizes are computed as follows.

$$d = \left| d^E(Max) - d^E(Min) \right| / b \tag{5}$$

- **Identification of pair set $\mathcal{D}^c$**: Based on the spatial distance, map each pair of objects into the corresponding bin.

$$\mathcal{D}^c = \{< r_i, r_j >, (c-1)\cdot d \leq d^E(r_i.X, r_j.X) < c\cdot d, c \in [1, b]\} \tag{6}$$

- **Identification of pair set $\mathcal{D}_{A^l A^{l'}}$**: Map each pair of objects into a subset in which each pair takes a specific category pairs.

$$\mathcal{D}_{A^l A^{l'}} = \{\langle r_i, r_j\rangle, [(r_i.A = A^l(A^{l'}))\ \&\ (r_j.A = A^{l'}(A^l))]\} \tag{7}$$

- **Identification of pair set $\mathcal{D}^c_{A^l A^{l'}}$**: Store each pair of objects in $D_{A^l A^{l'}}$ into the corresponding distance bins.

$$\mathcal{D}^c_{A^l A^{l'}} = \mathcal{D}^c \cap \mathcal{D}_{A^l A^{l'}} \tag{8}$$

- **Pair frequency computation**: Compute the pair frequency of the pair of categories in the c^th bin as follows

$$PF(\langle A^l, A^{l'}\rangle, c\cdot d) = |\mathcal{D}^c_{A^l A^{l'}}|/|\mathcal{D}^c| \tag{9}$$

  Here,$|\mathcal{D}^c_{A^l A^{l'}}|$ and $|\mathcal{D}^c|$ represent the number of pair objects in $\mathcal{D}^c_{A^l A^{l'}}$ and $\mathcal{D}^c$, respectively.

Overall, for each pair of $\langle r_i, r_j\rangle$, we can estimation d pair frequency values corresponding with d bins. Based on the d discrete points in a 2-D space, we can statistically learn a pair frequency function $PF(\langle A^l, A^{l'}\rangle, d^E)$ by polynomial and curve fitting, subjecting to the following constraints:

1. $PF(\langle A^l, A^{l'}\rangle, d^E) = PF(\langle A^{l', A^l}\rangle, d^E)$
2. $PF(\langle A^l, A^{l'}\rangle, 0) = \begin{cases} Freq(A^l) & A^l = A^{l'} \\ 0 & A^l \neq A^{l'} \end{cases}$
3. $\sum_{l'=1}^{L} PF(\langle A^l, A^{l'}\rangle, d^E) = Freq(A^l)$

Algorithm 1 PCF-SCOD Approach

```
for $i = 1$ to $n$ do {Identify neighborhood, distance matrix}
    $[Neighbor(i,:), DistMa(i,:) = kNN(X, r_i.X, k)]$
end for
$d^E_{Max} = max(DistMat)$;{Identify the maximum spatial distance}
$d^E_{Min} = 0$;{Set the minimum spatial distance}
$d = \frac{|d^E_{Max} - d^E_{Min}|}{b}$;{Calculate the size of unit bin}
$[n_A, Freq_A] = CateFreq(A)$; {Construct the category matrix}
for $c = 1$ to $b$ do {Identify pair dataset, $\mathcal{D}^c$}
    $\mathcal{D}^c = PairSetIdentify(DistMa, Cond^{\mathcal{D}^c})$
end for
for $l = 1$ to $n_A$ do {Identify pair dataset, $\mathcal{D}_{A^l A^{l'}}$}
    for $l' = 1$ to $n_A$ do
        $\mathcal{D}_{A^l A^{l'}} = PairSetIdentify(A, Cond^{\mathcal{D}_{A^l A^{l'}}})$
    end for
end for
for $l = 1$ to $n_A$ do
    for $l' = 1$ to $n_A$ do
        for $c = 1$ to $b$ do
            $\mathcal{D}^c_{A^l A^{l'}} = D^c \bigcap D_{A^l A^{l'}}$; {Identify pair dataset $\mathcal{D}^c_{A^l A^{l'}}$}
            $PF(A^l, A^{l'}, c\cdot d) = \frac{|\mathcal{D}^c_{A^l A^{l'}}|}{|D^c|}$ {Calculate pair frequency.}
        end for
        $PF(A^l, A^{l'}, d^E) = FitModel(PF(A^l, A^{l'}, [0 : d^E_{Max}])$;
    end for
end for
for $i = 1$ to $n$ do {Calculate PCR matrix}
    for $j = 1$ to $k$ do
        $f = Neighbor(i, j)$;
        $PCRMat(i, j) = \frac{PF(r_i.A, r_f.A, DisMat(i,f))}{|Freq_{r_i.A}|\cdot|Freq_{r_f.A}|}$;
    end for
end for
$RelevanceMat = mean(PCRMat)$; {Compute relevances}
$RankList = Rank(RelevanceMat, ascend)$; {Rank objects}
$O_l = Outlier(RankList, 1 : l)$ {Mark the outliers}
```

3.2 Algorithm of PCF-SCOD

Algorithm 1 describes PCF-SCOD approach as the following 7 steps.

Neighborhood Construction: First, construct distance matrix, $Dismat$, in which each entry records the spatial distance between each pair of objects, r_i and r_j. With it, the spatial neighborhood matrix, $Neighbor$, can be identified for each spatial object.

Bin Formulation: Then, with the stored values in $Dismat$, identify its maximum and minimum values. And the size of unit bin, d can be computed.

Frequency Computation: For each category,calculate its corresponding occurrence frequency, $Freq(A^l)$, at which the spatial objects take it in the dataset $\mathcal{D}$.

Pair-Sets Identification: Three critical pair-sets are identified in which function $PairSetIdentify$ is used to extract the pair objects which satisfy certain conditions, $Cond^{\mathcal{D}^c}$ and $Cond^{\mathcal{D}_{A^l A^{l'}}}$. After that, compute pair frequencies for each pair categories with different specified spatial distances. Finally, a continuous PF function is learned statistically.

Outlierness Computation.With PF function, PCR can be identified for any pair of spatial objects. Furthermore, PCR matrix is computed based on the mean of PCR values between the reference observation and its neighbors.

Outlier Detection: The top l objects with lower PCR values are recognized as outliers.

Time Complexity. To form the distance and neighborhood matrices will take $O(n^2)$. It takes $O(n)$ to construct the category frequency matrix. Identifying $\mathcal{D}^c$, $\mathcal{D}_{A^l A^{l'}}$ and $\mathcal{D}^c_{A^l A^{l'}}$ takes around $O(b\cdot n^2)$, $O(n_A^2 \cdot n^2)$ and $O(b\cdot\ n_A^2 \cdot (|\mathcal{D}^c|+|\mathcal{D}_{A^l A^{l'}}|))$. Finally, computing the PCR matrix costs $O(k\cdot n)$. In summary, assuming $n \gg k$, $n \gg b$, $n \gg n_A$ and $n \gg (|\mathcal{D}^c| + |\mathcal{D}_{A^l A^{l'}}|)$. The total time complexity of PCF-SCOD approach is $O(n^2)$=$(O(n^2)$+$O(n)$+$O(b\cdot n^2)$+$O(n_A^2 \cdot n^2)$+$O(b\cdot\ n_A^2 \cdot (|\mathcal{D}^c| + |\mathcal{D}_{A^l A^{l'}}|))$+$O(k\cdot n))$.

Table 1: Experiment Datasets

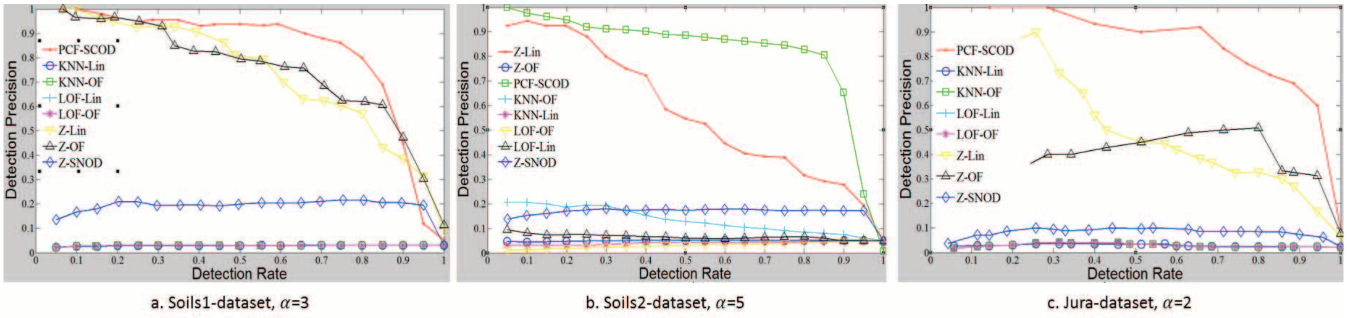
Dataset	Size	Categories
Jura	359	1:Argovian 2:Kimmeridgian 3:Sequanian 4:Quaternary
$Soil_1$	1000	1:Leptosol 2:alcisol 3: utcrops 4:Sand Dunes
$Soil_2$	3000	1:Luvisol 2:Leptosol 3: Plinthosol 4:Vertisol, 5:Nitisol 6:Lixisol 7:Fluvisol

4. EXPERIMENTS AND RESULTS

We conducted extensive experiments on 3 real datasets to compare the performances among the proposed PCF-SCOD, with other popular outlier detection approaches.

Experiment Settings

we chose Z-test [13], kNN [11] and LOF [4] methodologies. To compute the similarities among nominal categorical data, we used Lin and OF measurements [3]. Therefore, there were overall 6 different comparable approaches: Z-OF, Z-Lin, LOF-Lin, LOF-OF, kNN-Lin and kNN-OF. Also, we directly applied Z-test to categorical datasets by assuming that the nominal categories can be ordered. We executed all the approaches on 3 real datasets, including *Jura* [7], $Soil_1$ and $Soil_2$ [8]. Table 1 describes the detailed information of them. To demonstrate the effectiveness of the proposed method, we generated some synthetic outliers on the real datasets. We assumed the raw dataset as a ground truth, and randomly selected 2%, 3% and 5% of the data to be anomalies by modifying them from its original category to anyone of others. For each contamination rate, we



a. Soils1-dataset, α=3

b. Soils2-dataset, α=5

c. Jura-dataset, α=2

Figure 2: Comparison of algorithm performances on three spatial categorical datasets

generated the synthetic outliers 10 times, and then calculated the mean and standard deviation of accuracy for each method. To compare the accuracies among all methods, we used the common evaluation measures: the precision is plotted against recall. The parameter k was set as *8* and b(bin number) value in PCF-SCOD was *10*.

Experiments Results

Fig. 2 depicts the comparison our method against other existing approaches on the single attribute datasets. Each point in curves corresponds to the average performance over *10* randomly generated datasets. We observed that PCF-SCOD achieved 20% − 40% improvement over Z-OF, Z-Lin and Z-Gooall3, and 60% − 70% over LOF-Lin, LOF-OF, kNN-Lin, kNN-OF and Z-test.

The results confirmed two observations: firstly, the concept of SCOs is different with that of SNOs. Two objects taking different attributes is not necessarily irrelevant with each other. The co-occurrence exactly illustrates their higher relevance. This can be demonstrated by comparison results of Z series against PCF-SCOD; Secondly, when identifying SCOs, the existing TCOD and SCOD approaches can't avoid the well-known swamping and masking problems at all. TCOD approaches treat the spatial and non-spatial attribute equally and don't consider the spatial dependency and spatial correlation which are the specific properties of spatial data. And for the Z-OF and Z-Lin, they outperformed TCOD methods since they differentiate spatial and non-spatial attributes. However, they performed worse than PCF-SCOD. First, they treat SCO in an SNOD way. Second, the dissimilarity between objects is computed based on the global frequencies, not local frequencies.

5. CONCLUSION

In this paper, we utilize PCF concepts on SCOD and design an algorithm that can identify SCOs with single attribute. This is the first spatial outlier detection approach that processes spatial categorical data by capturing the special properties of nominal categorical attributes. The proposed approach can process not only nominal, but ordinal categorical datasets. The experiments conducted on real datasets demonstrated the effectiveness of the PCF-SCOD method.

6. REFERENCES

[1] N. R. Adam, V. P. Janeja, and V. Atluri. Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. In *ACM SAC '04*, pages 576–583, 2004.

[2] L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27(2):93–115, 1995.

[3] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM*, pages 243–254, 2008.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. ACM SIGMOD '00, pages 93–104, 2000.

[5] V. Chandola, S. Boriah, and V. Kuman. Understanding categorical similarity measures for outlier detection. Technical report, University of Minnesota, 2008.

[6] K. Das and J. Schneider. Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 220–229, New York, NY, USA, 2007. ACM.

[7] P. Goovaerts. *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press, 1997.

[8] http://www.iiasa.ac.at/Research/LUC/External-World-soil
database/HTML/.

[9] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. Statistical analysis and modelling of spatial point patterns. 2008.

[10] A. Koufakou, E. G. Ortiz, M. Georgiopoulos, G. C. Anagnostopoulos, and K. M. Reynolds. A scalable and efficient outlier detection strategy for categorical data. IEEE ICTAI '07, pages 210–217, 2007.

[11] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. ACM SIGMOD '00, pages 427–438, 2000.

[12] T. Reed and K. Gubbins. *Applied statistical mechanics: thermodynamic and transport properties of fluids*. Butterworth-Heinemann reprint series in chemical engineering. Butterworth-Heinemann, 1973.

[13] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *KDD*, pages 371–376, 2001.

[14] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.