Mining at Most Top-K% Spatio-temporal Outlier Based Context: A Summary of Results

Zhanquan Wang, Chunhua Gu, Tong Ruan, and Chao Duan

Department of Computer Science and Engineering,
East China University of Science and Technology, Shanghai, China
{zhqwang,chgu,ruan.tong}@ecust.edu.cn,duanchaoaa@126.com

Abstract. Discovering STCOD is an important problem with many applications such as geological disaster monitoring, geophysical exploration, public safety and health etc. However, determining suitable interest measure thresholds is a difficult task. In the paper, we define the problem of mining at most top-K% STCOD patterns without using user-defined thresholds and propose a novel at most top-K% STCOD mining algorithm by using a graph based random walk model. Analytical and experimental results show that the proposed algorithm is correct and complete. Results show the proposed method is computationally more efficient than naive algorithms. The effectiveness of our methods is justified by empirical results on real data sets. It shows that the algorithms are effective and validate.

Keywords: Spatio-Temporal Outliers, Top-K%.

1 Introduction

Spatio-temporal outlier detection, called anomaly detection in space and time, is an important branch of the data mining research [1][2][3]. At most top-K% spatial-temporal context outlier detection (TopSTCOD) represents some special objects that have some anomalous behavior without composite interesting measures in the space and time. Formally, given a collection of objects over a common Spatio-temporal framework, and a neighborhood relation over neighbors, a TopSTCOD mining algorithm aims to discover correct and complete sets of interesting and non-trivial TopSTCODs. A TopSTCOD represents a pattern set whose interest measures are in the top-K% of the complete set of objects and have higher values than patterns which are not found in STCOD method. Discovering TopSTCODs is important for many spatio-temporal application domains. For example. In the Masai Mara national reserve (MMNR) in Kenya[10], there are many species, such as wildebeest and zebra, they are gregarious species, but two groups often lives nearby. The existing methods can't find the patterns without composite interesting measures, and these thresholds values are mostly domain-specific and without domain knowledge, it is very difficult to set up suitable interest measure thresholds to mine the STCODs. If the user-defined threshold is too small to mine the patterns, it is highly likely that too many patterns will be generated. If the threshold values are too large, it is also possible to discover

H. Deng et al. (Eds.): AICI 2011, Part II, LNAI 7003, pp. 688–695, 2011.

too few patterns and miss possible significant ones. So it is very important and interesting for biologists to research their behavior and habit without interesting measure thresholds. There are other applications are such as military, ecology, and homeland defense [2][4][5].

To the best of our knowledge, this is the first work to discover top-K% spatio-temporal contextual outliers detection at spatio-temporal dataset; It includes the statement which are applicable to the real applications; A new and computationally efficient TopSTCOD mining method is presented; It includes comparisons of approaches and experimental designs. This paper focuses on TopSTCODs by statement of top K%. The rest of the paper is organized as follows. Section 2 reviews some background and related works in outlier detection data mining. Section 3 proposes basic concepts to provide a formal model of TopSTCOD. TopSTCOD and fast TopSTCOD mining algorithms are presented in section 4. The experimental results are proposed in section 5 and section 6 presents conclusions and future work.

2 Related Work

The quality of identified contextual outliers heavily relies on the meaningfulness of the specified context [2][6][8][9][12][19]. However a STCOD mining algorithm proposed in author's previous work which requires user-defined thresholds: a spatial contextual outlier measure and the time prevalence measure. The spatial contextual outlier is used to determine if the pattern is spatially anomalous. The time prevalence measure is used to determine if the pattern is frequent. These thresholds values are mostly domain-specific and without domain knowledge, it is very difficult to set up suitable interest measure thresholds to mine STCODs. If the user-defined threshold is too small to mine the STCODs, it is highly likely that too many patterns will be generated. If the threshold values are too large, it is also possible to discover too few patterns and miss possible significant ones. This study aims to discover TopSTCODs with no need for user-defined spatial threshold and time prevalence threshold by using a random walk graph and spectral analysis[2][11] which is a powerful tool to study the structure of a graph at each timeslot, where we use transition matrix to study how to get unknown contextual information for spatio-temporal data.

3 Basic Concepts and Statement of Modeling

The random walk graph and contextual outliers is omitted due to space limit [6]. The focus of this study is to discover at most top K% spatio-temporal context detection objects (TopSTCODs) over a spatio-temporal framework and a neighborhood relation R (or relation of social network). First we introduce basic concepts, and then explain the modeling of at most top-K% STCODs. Fig.1 shows an example of spatio-temporal contextual outlier. In spatio-temporal data, we proposed a spatiao-temporal framework. The context-based spatio-temporal outlier detection is different from the general outlier detection, which does not only consider spatial attributes and time attributes, but also consider the contextual information. The framework is as follow: a framework of spatio-temporal framework STF, an object (node) set:

$O = \{o_1,...o_i,...,o_n\}(1 \leqslant i \leqslant n)$, n is number of nodes at each timeslot. $T = \{t_1,...,t_j,...,t_m\}(1 \leqslant j \leqslant m)$, m is the number of timeslot. So we can define $STF = \{O_1,...,O_m\}$, $STF = OXT$, O_i is the object set at the timeslot t_i.

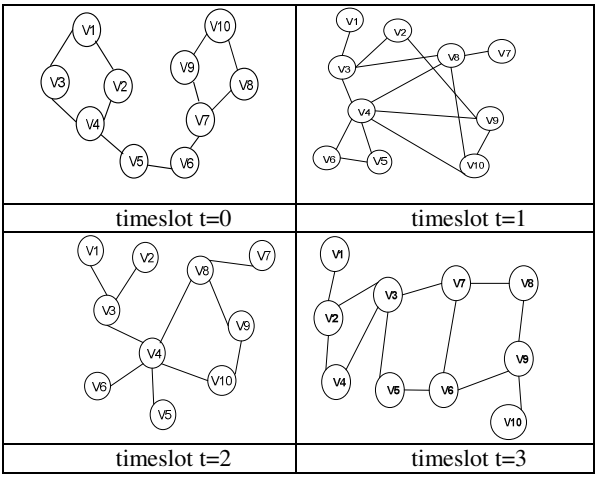

Fig. 1. An example for TopSTCOD

3.1 Modeling of at Most Top K% Spatio-Temporal Context Detection

A method that can detect the contextual outlier for spatio-temporal dataset with the non-main eigenvector is proposed. In the model, every non-main eigenvector of the transitional matrix defines a unique graph of 2-labeling/2-coloring. Intuitively, given a 2-coloring graph, each sub-graph could be regard as a context. Assume S^+ is a sub-graph and S^- is another, we can get the probability of a node be visited from the beginning of S^+ or S^-. There are some nodes called contextual outliers if the probability of they are visited by S^+ or S^- is equal.

Definition 1: Assume （S^+, S^-） is a 2-coloring of G, S^+ is a index set of the node marked +, and S^- is a index set of the node marked -. They are satisfied the follow condition: $S^+ \neq \varphi, S^- \neq \varphi, S^+ \cup S^- = \{1,...,n\}$. Then we can call （S^+, S^-） a pair of contexts of G. And the random walk in G can be called a contextual random.

Definition 2: (fixed expectation)**:** Assuming G is a random walk graph , W a transitional matrix, μ_i is the expectation of random variable[8], if μ_i satisfies the follow condition: $\mu_i = c\sum_{j=1}^{n}\mu_j w_{ij}, \forall i, 1 \leqslant i \leqslant n$, where c is a constant of

time-independent, then we call $\mu = (\mu_1,...,\mu_n)^T$ is the fixed expectation corresponding the contextual random walk of S^+ and S^-.

If W is a transitional matrix of positive, then every non-main eigenvector of W can uniquely determine a pair of contexts and the corresponding fixed expectation. Particularly, assume v is a eigenvector which corresponding to the eigenvalue $\lambda < 1$ of W. So we can regard v as a non-main eigenvector of W and get the follow lemma:

Lemma 1: Given a non-main eigenvector v of a positive transitional matrix, then $\sum_{i=1}^{n} v(i) = 0$, where $v(i)$ is an item of v.

According to lemma 1, we can define a 2-coloring of G with v, it can provide a pair of contexts: $S^+ = \{i : \mathbf{v}(i) > 0\}, S = \{i : \mathbf{v}(i) < 0\}$.

Now considering the contextual random walk with (S^+, S^-) in G, we can get follow theorems:

Theorem 1 (the fixed expectation of a contextual random walk)**:** If assuming $\mu = (\mu_1,...,\mu_n)^T$, $\mu_i = \frac{\mathbf{v}(i)}{\sum_{j=1}^{n} |\mathbf{v}(j)|}, \forall i, 1 \leqslant i \leqslant n$,where v is non-main eigenvector corresponding to the eigenvalue λ of W, so definition 4 is satisfied. Therefore, μ is a fixed expectation of contextual random walk graph. Theorem 1 indicates that every non-main eigenvector uniquely determines a 2-coloring graph (S^+, S^-) and its fixed expectation μ.According to the theorem 1, we can define the contextual outlier with fixed expectation.

Definition 3 (contextual outlier value)**:** Assume G is a random walk graph, W a positive transitional matrix, then the contextual outlier value (COV) of node i is $|\mu_i|$, and μ_i is the fixed expectation which defined according to Theorem 1.

According to the definition above, we can know that the contextual outlier value of any node is between 0 to 1. A small value indicates that the node is a contextual outlier.

We compute its contextual outlier value for all nodes in time slot t=0,1,2,3.

$\mu =$	$\mu =$	$\mu =$	$\mu =$
$\begin{bmatrix} 0.1168 \\ 0.1096 \\ 0.1096 \\ 0.1332 \\ 0.0309 \\ -0.0309 \\ -0.1332 \\ -0.1096 \\ -0.1096 \\ -0.1168 \end{bmatrix}$	$\begin{bmatrix} 0.0652 \\ 0.0845 \\ 0.1664 \\ -0.1467 \\ -0.1766 \\ -0.1766 \\ 0.0364 \\ 0.0931 \\ 0.0370 \\ 0.0175 \end{bmatrix}$	$\begin{bmatrix} -0.1307 \\ -0.0452 \\ -0.2766 \\ -0.0956 \\ -0.0271 \\ -0.0271 \\ 0.0749 \\ 0.1585 \\ 0.1120 \\ 0.0523 \end{bmatrix}$	$\begin{bmatrix} 0.1228 \\ 0.0318 \\ 0.1903 \\ 0.1232 \\ 0.0318 \\ -0.0858 \\ -0.0448 \\ -0.1389 \\ -0.1176 \\ -0.1129 \end{bmatrix}$
timeslot t=0	timeslot t=1	timeslot t=2	timeslot t=3

Given a set of spatio-temporal objects with a neighborhood relation R (or relation of social network), an at most top-K% STCOD is a subset of spatio-temporal object which are neighbors in space and time.

Definition 6: Given a spatio-temporal dataset, and a set T of time slots, a pattern is in the top-K% STCOD list if it is in the first K% of the number of all objects based on the lowest value of in the spatial contextual measures at each timeslot ($ascend(fixexpecation_{ti})/n \leqslant K\%$) and the highest values in the time measures for all timeslots($descend_{timeslots}((c\_ascend(fixedexpectation_{ti})/n \leqslant)K\%) \leqslant K\%$), where ascend is count of sorting ascend for the contextual outlier value at each timeslot. descend is value of sorting descend time prevalence index values.

Table 1. Valuse of the contextual outlier

nodes	t=0	t=1	t=2	t=3	Time prevalence index values
v1	0.1168	0.0652	0.1307	0.1228	0
v2	0.1096	0.0845	0.0452	0.0318	1/4
v3	0.1096	0.1664	0.2766	0.1903	0
v4	0.1332	0.1467	0.0956	0.1232	0
v5	0.0309	0.1766	0.0271	0.0318	3/4
v6	0.0309	0.1766	0.0271	0.0858	2/4
v7	0.1332	0.0364	0.0749	0.0448	1/4
v8	0.1096	0.0931	0.1585	0.1389	0
v9	0.1096	0.037	0.112	0.1176	0
v10	0.1168	0.0175	0.0523	0.1129	1/4

Fox example, in fig.1. there are ten objects. A top-20% STCOD will include two objects (v5,v6) which are the top-20% of ten objects which are at most 20% spatio-temporal outlier based context. Table.1 shows the detailed part.

3.2 Analysis for Model

At most top-K% spatio-temporal contextual outliers which are produced from our methods are correct because the patterns satisfy threshold pairs. The patterns are complete because our algorithms can find any STCODs as long as it satisfies our definitions and rules. The model average time complexity is $O(n^2 m)$. We omit the detail analysis due to space limit.

4 Mining TopSTCODs

In the section, we discuss the implementation of our spatio-temporal contextual outlier values in practice. We propose a hierarchical algorithm which iteratively

partitions the data set for each time slots until the size of the sub graph is smaller than a user specified threshold pairs. In every time slots, we acquire the contextual outlier value of spatial object with the method of contextual outlier detection mentioned above. Set a K%. The naïve algorithm1 of context-based spatio-temporal outlier detection is omitted due to space limit, We only describe the fast algorithm for spatio-temporal contextual outlier which is more efficient than naïve method.

Algorithm2: top K% spatial-temporal contextual outlier detection

Inputs: spatial-temporal data set STD, number of timeslot m, K;
Output: at most top K% spatial-temporal contextual outliers

```
1: TopSTO ← φ,
2: foreach t_i ∈ T do
3:    foreach i∈ TopSTO do
4:       Tr(t_i ,i)=0;
5:    end;
6:  creat random walk graph G and transition matrix W;
7:  TopSTO (G, W,K) ;
8:    foreach i∈ L do
9:          if  i(score)PERCENT > K then
10:              Tr(t_i ,i)=1;
11:         end;
12:      end;
13: end
14: foreach i∈ TopSTO do
15:    TOS(i)= ∑_{j=1}^{m} Tr(t_j ,i) / m ;
16:       if  TOS(i) PERCENT > K  then
17:            TopSTO = TopSTO ∪ {i} ;
18:       end
19: end
```

The time complexity of our algorithm is a time polynomial. The algorithm involves calculating the first and second largest eigenvalue and eigenvector of an n × n matrix, where n is the number of nodes. So its complexity is mainly determined by the complexity of eigen-decomposition. Second time cost is sorting data by outlier values, and the average time complexity is $O(n^2 m)$, the time cost for judging outlier is $O(nm)$. Therefore, we would adopt a efficient method to calculate eigenvalue and eigenvector. As a characteristic of outliers is that they are composed by minority objects, so we don't necessarily to consider all the objects in the process of calculation. We can set a K%, only put the object whose outlier value is smaller than threshold K% into the sort list L. It can keep the most normal data without any unnecessary operation in order to improve algorithm. When calculating outlier in the time series, we don't need to deal with all objects, but only a minority objects, so the size of the sort list will be greatly reduced. When judging spatial contextual outlier, we can set number set a K%, the object whose outlier value is smaller than K% will regard as outlier, this is the first modified part.

The second modified part is: when judging outlier in the time slots, some nodes have satisfied the judging condition before deal with all the times, so it is unnecessary to deal with the later times. The improved algorithm is described as the fast algorithm.

5 Experimental Evaluation

In this section, we present our experimental evaluations of several design decisions and workload parameters on our TopSTCOD mining algorithms. We used two real-world training dataset,(mail log for users and vehicle data set). We evaluated the behavior of two algorithms. Fig.2 shows the experimental setup to evaluate the impact of design decisions on the performance on these methods. Table.2 shows real dataset in vehicles. Experiments were conducted on a Windows XP, 2.0GHz Inter Pentium 4 with 1.5GB of RAM.

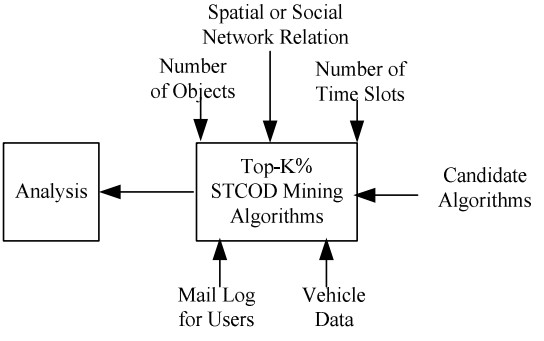

Fig. 2. Experimental setup

5.1 Effect of Number of Time Slots

According to the location information of the vehicle we establish transition matrix with spatial relation R(R=100). Experiment analysis would be done according to number of time slots. We evaluated the effect of number of timeslots on the execution time of both algorithms by mining at most top-10% STCOPs. The parameter K parameter was set at 10. The execution time of both algorithms increase, as the number of timeslots is increased (Fig.3). The TopSTCOD-Miner is computationally more efficient than the naïve approach because of its early pruning strategy (Fig.3). As the number of time slots increases, the ratio of the increase in execution time is smaller for TopSTCOD -Miner than with the naïve approach.

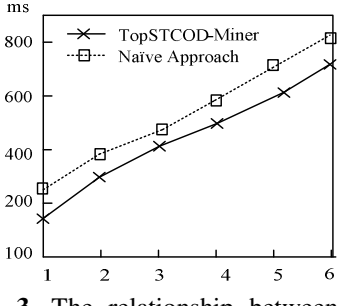

Fig. 3. The relationship between the number of timeslots and time

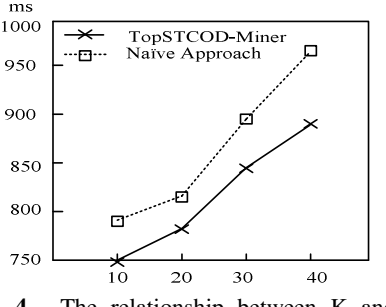

Fig. 4. The relationship between K and running time

5.2 Effect of Parameter K

In the third experiment we evaluated the effect of number of K on the execution times of algorithms. The neighborhood distance and number of timeslots are 100m and 10. Fig. 4 shows that the execution time of both algorithms increases and the TopSTCOD-Miner outperforms the naïve approach as the parameter K increases.

From the experiment analysis above, we can draw a conclusion that the improved algorithm is more efficient than the original algorithm.

6 Conclusion

We defined at most top-K% spatio-temporal contextual outlier detection and its mining problem. We also presented a novel and computationally efficient algorithm for mining these patterns and its improved method, and proved that the model is correct and complete in finding at most -K% spatio-temporal contextual outliers. Our experimental results using the vehicle dataset from the real world provide further evidence of the viability of our approach. For future work, we would like to explore the relationship between the proposed composite interest measures and spatio-temporal statistical measures of interaction[2]. We plan to develop other new computationally efficient algorithms for mining TopSTCODs; further study variation of transition matrix due to multi-scale of space or time, and space and time.

Acknowledgement. The authors are very grateful to the financial support from the National Natural Science Foundation of China under Grant No. 60703026

References

1. Han, J.W., Kamber, M.: Data mining concepts and techniques. Morgan Kaufmann Publishers, San Francisco (2001)
2. Shekhar, S., Chawla, S.: Spatial databases: a tour. Prentice Hall, Englewood Cliffs (2003)
3. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. In: SIGMOD Conference (2000)
4. Moonesinghe, H.D.K., Tan, P.N.: Outlier detection using random walks. In: ICTAI (2006)
5. Kou, Y., Lu, C.T., Chen, D.: Spatial weighted outlier detection. In: SDM (2006)
6. Wang, X., Davidson, I.: Discovering contexts and contextual outliers using random walks in graphs. In: ICDM 2009 (2009)
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. 41(3) (2009)
8. Skillicorn, D.B.: Detecting anomalies in graphs. In: ISI (2007)
9. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: ICDM, pp. 413–422 (2008)
10. Barnet, V., Lewis, T.: Outlier in statistical data. John Wiley&Sons, New York (1994)
11. Chung, F.: Spectral graph theory. American Mathematical Society, Providence (1997)
12. Song, X., Wu, M., Jermaine, C.M., Ranka, S.: Conditional anomaly detection. IEEE Trans. Knowl. Data Eng. 19(5) (2007)