Spatio-temporal Outlier Detection Based on Context: A Summary of Results

Zhanquan Wang, Chao Duan, and Defeng Chen

Department of Computer Science and Engineering,
East China University of Science and Technology, Shanghai, China
zhqwang@ecust.edu.cn, duanchaoaaa@126.com, chendefeng918@163.com

Abstract. Spatio-temporal outlier detection plays an important role in some applications fields such as geological disaster monitoring, geophysical exploration, public safety and health etc. For the current lack of contextual outlier detection for spatio-temporal dataset, spatio-temporal outlier detection based on context is proposed. The pattern is to discover anomalous behavior without contextual information in space and time, and produced by using a graph based random walk model and composite interest measures. Our approach has many advantages including producing contextual spatio-temporal outlier, and fast algorithms. The algorithms of context-based spatio-temporal outlier detection and improved method are proposed. 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, Context, Composite Interest Measures.

1 Introduction

Spatio-temporal outlier detection, also called anomaly detection in space and time, is an important branch of the data mining research [1][2][3]. Most of the existing approaches identify outliers from a global point view for spatio-temporal datasets. However, sometimes an instance may not be an outlier when compared to the rest of the spatio-temporal dataset but may be a spatio-temporal outlier in the context of a subset of dataset from point view of space and time. This type of outlier detection is called contextual spatio-temporal outlier detection. There are many applications in the real world, 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. It is important for us to find that some wildebeets often live in other group, it is a anomalous behavior, the reverse is also true. The existing methods can't find the pattern. We propose one new method to detect spatio-temporal outlier based context using a graph rand and define composite interest measures. Firstly a transition matrix can be generated from data set according to the relation, Then contexts and contextual outlier at each timeslot are defined using a graph based random walk model. The spatio-temporal contextual outlier can be produced by

H. Deng et al. (Eds.): AICI 2011, Part I, LNAI 7002, pp. 428–436, 2011.

defining composite interest measures which are suitable to the real application. Our contributions are: discover spatio-temporal contextual outliers detection (STCOD) without a prior contextual information is proposed; It includes the definition and interpretation for composite interest measures which are applicable to the real applications; A new and computationally efficient STCOD mining method is presented; It includes comparisons of approaches and experimental designs. 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 STCOD. STCOD 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

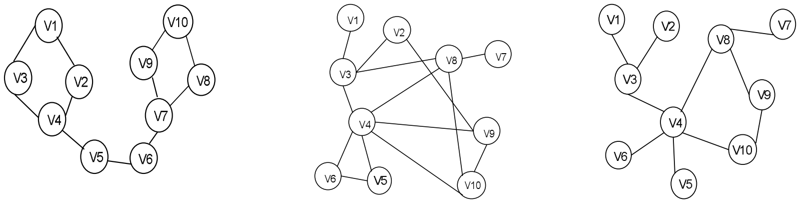
The quality of identified contextual outliers heavily relies on the meaningfulness of the specified context [9], many existing approaches require a priori contextual information, for example, S.Salvador[7] can get states and rules for time series anomaly detection which needs priori contextual information. S. Shekhar[2] proposes many outlier methods, but don't deal with time series anomaly and can't deal with it under unknown contextual information. X. Song[3] proposes conditional anomaly detection, it can deal with the specified context, but need contextual information. Tan[19] introduces random walk model for global outlier detection and used the principal eigenvector of the transition matrix as global outlier score, but it don't process contextual outliers and spatio-temporal dataset. Skillicorn[8] uses spectral graph analysis to explore anomalous structure in a graph. The methods are focused on outlier detection in a global point of view, and don't address contextual outliers and spatio-temporal dataset. X. Wang[6] can deal with the patterns without priori contextual information by proposing a probabilistic approach based on random walk, but it can't process spatio-temporal dataset. Our works explore meaningful contexts and contextual outlier for spatio-temporal applications 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 and how to define composite interest measures which are applicable to real applications.

3 Statement and Algorithm Modeling

For spatio-temporal data, we proposed a spatio-temporal framework to deal with context-based spatio-temporal outlier detection. The spatio-temporal framework is as follow: a framework of spatio-temporal framework STF, an object(node) set: $O=\{o_0,\ldots,o_1,\ldots,o_{n-1}\}(0\leqslant i\leqslant n-1)$, n is number of nodes at each timeslot. T=$\{ t_0,\ldots,t_1,\ldots,t_{n-1}\}(0\leqslant i\leqslant m-1)$, m is the number of timeslot. So we can define $STF=\{O_0,\ldots O_{m-1}\}$,$STF=OXT$, O_i is the object set at the timeslot t_i.

3.1 Spatio-temporal Contextual Outlier Detection

The random walk graph and contextual outliers is omitted due to space limit[7]. A method that can detect the contextual outlier for spatio-temporal dataset with the non-main eigenvector is proposed. In the model, a unique graph of 2-labeling/2-coloring is defined in every non-main eigenvector of the transitional matrix. 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 nodes visited by S^+ or S^- is equal. Fig.1 shows an example of spatio-temporal contextual outlier.



timeslot t=0 timeslot t=1 timeslot t=2

Fig. 1. An example for STCODs

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

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: $\mu_i = c\sum_{j=1}^{n}\mu_j w_{ij}, \forall i, 0 \leqslant i \leqslant n \quad 1$, where c is a constant of time-independent, then we call $\mu = (\mu_0, \mathrm{L}, \mu_{n \quad 1})^T$ is the fixed expectation for the contextual random walk of S^+ and S^-.

If W is a positive transitional matrix, then every non-main eigenvector of W can uniquely determine a pair of contexts and the corresponding fixed expectation. Particularly, assume v is an eigenvector which corresponds 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: $S^{+} = \{i : \mathbf{v}(i) > 0\}, S \quad = \{i : \mathbf{v}(i) < 0\}$.

Now consider 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 assume $\mu = (\mu_0, \ldots, \mu_{n\ 1})^{T}$, $\mu_i = \dfrac{\mathbf{v}(i)}{\sum_{j=1}^{n} |\mathbf{v}(j)|}, \forall i, 0 \leqslant i \leqslant n \quad 1$,where v is non-main eigenvector corresponding to the eigenvalue λ of W, so definition 2 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 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. The detailed part is shown in [9]

$\mu = \begin{bmatrix} 0.1168 \\ 0.1096 \\ 0.1096 \\ 0.1332 \\ 0.0309 \\ -0.0309 \\ -0.1332 \\ -0.1096 \\ -0.1096 \\ -0.1168 \end{bmatrix}$	$\mu = \begin{bmatrix} 0.0652 \\ 0.0845 \\ 0.1664 \\ -0.1467 \\ -0.1766 \\ -0.1766 \\ 0.0364 \\ 0.0931 \\ 0.0370 \\ 0.0175 \end{bmatrix}$	$\mu = \begin{bmatrix} -0.1307 \\ -0.0452 \\ -0.2766 \\ -0.0956 \\ -0.0271 \\ -0.0271 \\ 0.0749 \\ 0.1585 \\ 0.1120 \\ 0.0523 \end{bmatrix}$
timeslot t=0	timeslot t=1	timeslot t=2

Definition 4: Given a spatio-temporal dataset, a spatial contextual outlier measure is value to judge if the node is contextual outlier. We know v5,v6 in time slot t=0, v7,v9,v10 in time slot t=1, v5,v6 in time slots t=2 are the contextual outlier if spatial contextual outlier threshold is 0.04.

Definition 5: Given a spatio-temporal pattern and a set T of timeslots, such that $\mathrm{T} = \{t_1,...,t_j,...,t_m\}(0 \leqslant \mathrm{j} \leqslant \mathrm{m} - 1)$. The time contextual outlier measure of the pattern is the fraction of timeslots where the patterns occur the total number of timeslots.

Given a spatio-temporal dataset of STF, and spatial contextual outlier threshold θ_c, the composite measure of pattern P_i is a composition of the spatial contextual outlier and the time prevalence measures as shown below.

$$\Pr ob_{t_i \in T}(c\_outlier(P_i, timeslot \quad t_i) \leqslant \theta_c) \tag{3}$$

Where $\Pr ob$ stands for probability of overall prevalence time slots, and $c\_outlier$ stands for spatial contextual outliers.

Consider a spatio-temporal dataset of STF and a threshold pair (θ_c, θ_t), where θ_t is time prevalence threshold. $STCOD$ is a prevalent pattern if it's composite measures satisfy the following.

$$\Pr ob_{t_i \in T}(c\_outlier(P_i, timeslot \quad t_i) \leqslant \theta_c) \geqslant \theta_t \tag{4}$$

Where $\Pr ob$ stands for probability of overall prevalence time slots, $c\_outlier$ stands for spatial contextual outliers. θ_c is the spatial contextual threshold and θ_t is the time prevalence threshold. We can know v5, v6 are spatio-temporal contextual outlier if $\theta_t = 0.6$.

3.2 Analysis for Model

Spatio-temporal contextual outliers 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 STCODs

In the section, we discuss the implementation of our spatio-temporal contextual outlier score in practice. We propose a novel 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 threshold θ_c, our method can judge if the node is a spatial contextual outlier, then we get spatio-temporal outlier based on context according to θ_t. 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: fast spatial-temporal contextual outlier detection

```
Inputs: spatio-temporal data set STF, G,W, $\theta_c$ ,$\theta_t$ ;
Output: spatial-temporal contextual outliers(STCOD)
STCOD$\leftarrow\phi$ , TOS(i) $\leftarrow$0,Lt$\leftarrow\phi$ , Sum(i) $\leftarrow$0;i$\leftarrow$0
for each $t_i \in$ T do
  for each i$\in$ STF do
      Tr($t_i$ ,i)=0;
  end;
  creat random walk graph G and transition matrix W;
  STCOD (G, W,α,$\theta_c$ ) ;
     Add L to Lt;
     for each i$\in$ L do
           if TOS(i) > $\theta_t$ then
                 STCOD=STCOD$\cup${i};
                 Delete i item from L and Lt;
           else Tr($t_i$ ,i)=1;
           end;
       end;
     for each i$\in$ L do
       Sum(i)= Sum(i)+ Tr($t_i$ ,i);
       TOS(i)= Sum(i) $/m$ ;
     end;
     i++;
end;
for each i$\in$ Lt do
     if TOS(i) > $\theta_t$ then
           STCOD=STCOD$\cup${i};
     end
end
return STCOD;
```

The time complexity of 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 value. Therefore, we would adopt an efficient method to calculate eigenvalue and eigenvector. As a trick of outliers is that they are composed by minority objects, so the first modified part is: we don't necessarily consider all the objects in the process of calculation. When making spatio-temporal contextual outlier detection, we can set a threshold β, only put the objects whose outlier value is smaller than threshold β into the sort list L. It can keep the most normal data without any unnecessary operation in order to improve algorithm. When calculate 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 a threshold θ_c, the objects whose outlier value is smaller than θ_c will regard as outlier. In our algorithm, we adopt the latter approach. We can put the judging work into algorithm1, just need set $\beta = \theta_c$; When judging outlier in the time slots, some nodes have already satisfied the judging condition before dealing with all the time slots and not be judged as spatio-temporal contextual outliers, so the second modified part is: it is unnecessary to continue to deal with these nodes at the rest timeslots. 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 STCOD 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. Experiments were conducted on a Windows XP, 2.0GHz Inter Pentium 4 with 1.5GB of RAM.

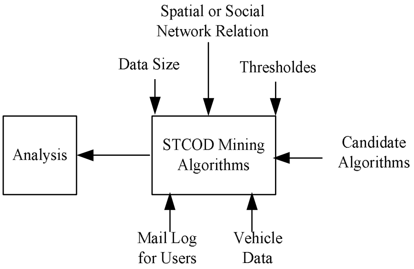


Fig. 2. Experimental setup

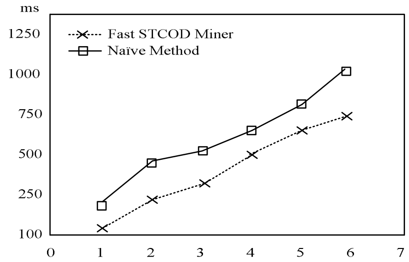


Fig. 3. The relationship between the number of timeslots and time

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 and threshold θ_c , θ_t . The solid line represent the naïve algorithm, and dotted line represent the improved algorithm (fast algorithm). From the Fig.3, we can find that the running time of the improved algorithm is shorter than the original algorithm, and the gap is increasing along as the number of time slots increasing. Because the improved algorithm has exclude the most normal data with the thresholds, which reduces a lot of operate time.

On the Fig.4, it shows the relationship between threshold θ_c and running time. The number of time slots and the other threshold θ_t are 6 and 0.7, we can see that with the increase of the threshold θ_c, running time of the fast algorithm is increasing, but the overall running time is less than the naïve algorithm.

On the Fig.5, it shows the relationship between threshold θ_t and running time. Assume the times m and the other threshold θ_c are both fixed value, we can see that with the increase of the threshold θ_t, the overall running time of the fast method is less than the naive algorithm.

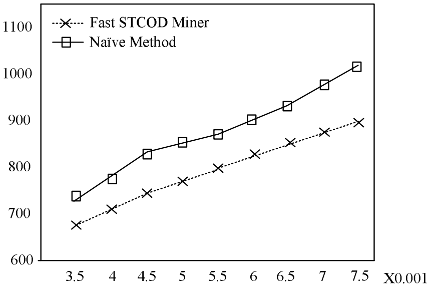


Fig. 4. The relationship between threshold θ_c and running time

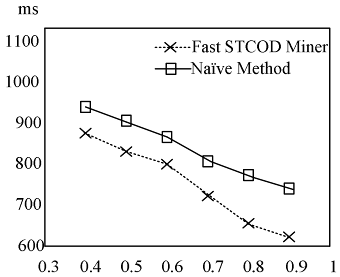


Fig. 5. The relationship between threshold θ_t and running time

From the experiment analysis above, we can draw a conclusion that the improved algorithm is more efficient than the original algorithm. The experimental evaluations were presented for mail log for users from our department. It also shows our methods can deal with STCOD patterns and effective and validate. We don't show the detailed results for mail log data for users due space limits.

6 Conclusion

We defined spatio-temporal contextual outlier detection and its mining problem, and proposed a new monotonic composite interest measure. We presented a novel and computationally efficient algorithm for mining patterns and its improved method, and proved that the model is correct and complete in finding 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 STCODs. So 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. Song, X., Wu, M., Jermaine, C.M., Ranka, S.: Conditional anomaly detection. IEEE Trans. Knowl. Data Eng. 19(5) (2007)
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)
7. Salvador, S., Chan, P., Brodie, J.: Learning states and rules for time series anomaly detection. In: Proc.17th Intl. FLAIRS Conf. (2004)
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)