A Rough Set Approach to Spatio-temporal Outlier Detection

[1] University of Naples Parthenope - 80143 Naples, Italy
{alessia.albanese,alfredo.petrosino}@uniparthenope.it
http://cvprlab.uniparthenope.it
[2] Indian Statistical Institute, Kolkata 700 108, India
sankar@isical.ac.in
http://www.isical.ac.in/

Abstract. Detecting outliers which are grossly different from or inconsistent with the remaining spatio-temporal dataset is a major challenge in real-world knowledge discovery and data mining applications. In this paper, we deal with the outlier detection problem in spatio-temporal data and we describe a rough set approach that finds the top outliers in an unlabeled spatio-temporal dataset. The proposed method, called Rough Outlier Set Extraction (ROSE), relies on a rough set theoretic representation of the outlier set using the rough set approximations, i.e. lower and upper approximations. It is also introduced a new set, called Kernel set, a representative subset of the original dataset, significative to outlier detection. Experimental results on real world datasets demonstrate its superiority over results obtained by various clustering algorithms. It is also shown that the kernel set is able to detect the same outliers set but with such less computational time.

1 Introduction

Spatio-temporal data mining is a growing research area dedicated to the disclosure of hidden knowledge in large spatio-temporal databases, mainly through detecting periodic patterns and outliers detection. This paper addresses the problem of outlier detection in spatio-temporal data using rough set theory, proposed by Pawlak [6]. Only a few methods for outlier detection, in general and in spatio-temporal context, exploit rough set theory in order to define degrees of outlierness based on rough set concepts. Nguyen in [3] discusses a method for the detection/evaluation of outliers, as well as how to elicit background domain knowledge from outliers using multi-level approximate reasoning schemes. Y. Chen, D. Miao, and R. Wang in [4] demonstrate the application of granular computing model using information tables for outlier detection. F. Jiang, Y. Sui and C. Cao in [5] propose a new definition of outliers that exploits the rough membership function. In contrast to those approaches that interpret the rough set theory from the operator-oriented point of view [2], our method exploits the set-oriented view of rough set theory in order to define the concept of outlier in

A. Petrosino et al. (Eds.): WILF 2011, LNAI 6857, pp. 67–74, 2011.

terms of its lower and upper approximations, keeping into account those objects that can neither be ruled in nor ruled out as members of the target concept.

We also introduce a new set, named Kernel Set. This is a selected subset of elements able to describe the original dataset both in terms of data structure and obtained results. We have shown the advantages of considering the Kernel Set in term of computation time by comparing the Rough Outlier Sets extracted by the original dataset with those extracted by the Kernel Set.

At this aim, the paper is organized as follows. Section 2 defines the problem. Section 3 reports the new rough set approach ROSE (Rough Outlier Set Extraction) to detect Spatio-Temporal (ST) Rough Outlier Set. Section 4 defines the Kernel Set and explains its significance to outlier detection. Sections 5 and 6 present executed tests on a real world dataset and the performance evaluation of the algorithm. Finally, conclusion remarks are given in Section 7.

2 Spatio-temporal Outlier Detection Problem

Let $S =< U, A >$ be an information system with U a normalized dataset and A its set of attributes. U can be written as follows:

$$U = \{p_i \equiv (z_{i1}, z_{i2}, ..., z_{im}) \in [0,1]^m, \quad i = 1, ..., N\}$$

where $p_i,\ i = 1, ..., N$ is a m-dimensional feature vector and A=$\{a_1, a_2, a_3, ..., a_m\}$ is its attribute set.

The proposed definition of the **Outlier Detection Problem** is as following:

Definition 1. *Given U, an integer $n > 0$ and a measure $d_{p_i}(U)$, defined over every $p_i \in U$, the Outlier Detection Problem consists of finding $\overline{n} \geq n$ objects $p_1, p_2, ..., p_n, p_{n+1}, ..., p_{\overline{n}} \in U$ such that*

$$d_{p_1}(U) \geq d_{p_2}(U) \geq ... \geq d_{p_n}(U) = d_{p_{n+1}}(U) ... = d_{p_{\overline{n}}}(U) > d_{p_j}(U), \\ \forall j = \overline{n} + 1, ..., N$$

The concept of measure is used to determine the degree of dissimilarity of each object with respect to others. Then, the *n-Outlier Set* can be formally defined:

Definition 2. *A n-Outlier Set $O \subseteq U$ is the set of $\overline{n} \geq n$ objects:*

$$O = \{p_1, ..., p_n, p_{n+1}, ..., p_{\overline{n}} \in U \ / \ \ d_{p_1}(U) \geq ... \geq d_{p_{n(U)}} = d_{p_{n+1}}(U) ... = \\ d_{p_{\overline{n}}}(U) > d_{p_j}(U) \ \ \forall j = \overline{n} + 1, ..., N\}$$

where $d_{p_i}(U),\ \forall i = 1, ..., N$ is a measure defined and computed on U.

From definition 2 it follows that $\tau = d_{p_n}(U)$ is the **outlierness threshold**, i.e.

$$\tau = inf\{max_1(d_p(U), d_q(U)), ..., max_n(d_p(U), d_q(U))\}, \forall p, q \in U \qquad (1)$$

Starting from the definition of spatial and temporal outlier due to Birant and Alp [8], we propose the following definitions applied to ST data. In this case U is a ST normalized dataset in which, at least, three attributes must be present, i.e. : the two spatial attributes and the temporal one.

Definition 3. *Given U, an integer $n > 0$ and a measure on spatial and temporal components $d^{s,t}_{p_i}(U)$, defined over every $p_i \in U$, an object $p \in U$ is a ST-***Outlier** *iff $d^{s,t}_p(U) \geq \tau$ where τ is defined as in (1).*

In a ST–context, a feasible measure to be associated to each object is based on the distances from its spatial and its temporal k-nearest neighbors [10]. Precisely:

$$d^{s,t}_p(U) = \alpha \cdot d^s_p(U) + \beta \cdot d^t_p(U) \tag{2}$$

where:

$$d^s_p(U) = \sum_{j=1}^{k} d^s(p, N^s(p, p_j)) \;\; and \;\; d^t_p(U) = \sum_{j=1}^{k} d^t(p, N^t(p, p_j)), \quad \forall p \in U \tag{3}$$

$k > 0$ is nearest neighbors number, $N^s(p, p_j)$ and $N^t(p, p_j)$ are the j-th spatial and temporal nearest neighbor of p, respectively and α, β are such that $\alpha+\beta = 1$. Definition 1 defines **ST-Outlier Detection Problem**, selecting a measure as in (2).

3 Rough Outlier Set Extraction (ROSE)

The goal of the proposed approach is to exploit the rough set theory to define the *Outlier Set* such as a *Rough Outlier Set (ROS)*. Let $S =< U, A >$ be an information system with U a ST normalized dataset and A its attribute set.

Given $n > 0$ (outlier number), we want to describe $O \subseteq U$ *(n–Outlier Set)* as

$$< \underline{B}(O), \overline{B}(O) > (Rough\ n - Outlier\ Set) \tag{4}$$

where $\underline{B}(O)$ is the *B–Lower approximation* and $\overline{B}(O)$ is the*B–Upper approximation* of *n–Outlier Set* with respect to an attribute subset $B \subseteq A$.

The *B–Lower approximation* $\underline{B}(O)$ is defined as the set of objects that can be certainly classified as members of the set O on the basis of the knowledge in B, while the objects in the *B–Upper approximation* $\overline{B}(O)$ as possible members of O on the basis of the knowledge in B.

At this aim, let I_B be the *B–indiscernibility relation* on the universe U:

$$I_B = \{(p_i, p_j) \in U \times U : a(p_i) = a(p_j),\ \forall a \in B\}$$

The equivalence classes $[p_j]_B$ or granules G_j of the partition induced by I_B on U are such that:

$$U = \bigcup_{j=1}^{N} G_j \;\; and \;\; G_j \cap G_j = \emptyset, \;\; i \neq j.$$

The measure in (2) is used as a spatio-temporal weight $\overline{\omega}_{G_j}(s, t, i)$, to be assigned to every granule G_j, depending on space, indicated by s, and/or on time, by t, and on iteration, by i and then the considered attribute subsets B are spatio-temporal attributes, only spatial and only temporal attribute. In our framework, the *B–Lower* and *B–Upper approximation*, at iteration i, can be defined as follows:

Definition 4. *The B–Lower approximation* $\underline{B}_i(O)$ *of n-Outlier Set O, at iteration i, is:* $\underline{B}_i(O) = \{G_j \subseteq U : \overline{\omega}_{G_j} > \tau_i\}$

$$where\ \tau_i = inf\ \{max_1^i\ (\overline{\omega}_{G_j}, \overline{\omega}_{G_k}), ..., max_n^i\ (\overline{\omega}_{G_j}, \overline{\omega}_{G_k})\},\ \forall\ G_j, G_k \subseteq U \quad (5)$$

Definition 5. *The B–Upper approximation* $\overline{B}_i(O)$ *of n–Outlier Set O, at iteration i, is:*

$$\overline{B}_i(O) = \{G_j \subseteq U : \overline{\omega}_{G_j} > \overline{\tau}_i\} \quad where: \quad \overline{\tau}_i = \tau_{i-1},\ \forall\ \ i >= 2 \qquad (6)$$

The threshold τ_1 is computed as the minimum value among the n higher values of weights assigned to the granules at first iteration.

The iterative procedure will stop when the thresholds does not vary anymore then the best Lower and Upper approximations in (4) have been reached.

ROSE Algorithm. The *Rough Outlier Set Extraction Algorithm* is designed to receive as input the universe U, k the nearest neighbors number and n the number of outliers to detect. The output of the procedure is the *ROS* (*Upper*, *Lower*, *Boundary* and *Negative Region*). At each iteration, the procedure randomly selects a subset of objects and computes their weights considering spatial and/or temporal components depending on the attribute subset B, with respect to, the *ROS* has been calculating. *UpdateUpperApprox* and *UpdateLowerApprox* functions compute the *lower* and *upper approximation* of *ROS*, using the current τ and previous τ_prev thresholds as defined in (5) and (6) respectively. A pruning strategy identifies objects from U having their weight under the threshold in order to build the *Negative Region*.

4 Kernel Set and Relevance to Outlier Detection

Let us now define *Kernel Set* $K \subseteq U$ that is a representative subset of the universe U that characterizes the overall dataset. Intuitively, this subset of U is able to maintain the structure of the universe U.

Definition 6. *Given* U *and two integers* $n > 0$, $k > 0$ *(number of nearest neighbors),* $d(U)$ *a measure defined on* U*, the Kernel Set* K *is built by adding each object* $p \in U$ *such that one of the following properties holds:*

1. $d_p(U) \geq \tau$
2. *if* $d_p(U) < \tau$*, then* $\exists q \in U$ *such that* $p \in NN^k(q)$ *and* $d_q(U) < \tau$ *and* $d_q(K - \{p\}) \geq \tau$

where $NN^k(q)$ *is the set of k-nearest neighbors of* q *and* $d(K)$ *is the restriction of* $d(U)$ *on* $K \subseteq U$*.*

The Definition 6 states that the objects that belong to the **Kernel Set** are:

1. object p for which $d_p(U) \geq \tau$ and hence belongs to *n–Outlier Set*.
2. object p that, even if $d_p(U) < \tau$, is one of the nearest neighbors of an object q for which hold $d_q(U) < \tau$ and $d_q(K - \{p\}) \geq \tau$.

The second property states that, once these objects p have been added to K, the measure of the object q become less than τ also in K as it is in U. Otherwise, the global structure of the dataset should be altered. The procedure allows to build also the *Kernel Set*, following the definition (6).
Some properties of Kernel Set have been proved:

1. a Kernel Set K contains the *n–Outlier Set*: $K \supseteq O$.
2. The Outlier Set O_K, computed from Kernel Set K is a superset of O computed from U: $O_K \supseteq O$.

The motivation of *Kernel Set* is that it is significative to outlier detection.

Indeed, outlier detection is a time consuming task, the use of *Kernel Set*, instead of U, as input of the *ROSE* procedure, have two major advantages:

- same results in terms of *rough outlier set* is obtained
- computational time is reduced due to the lower cardinality of *Kernel Set* respect to U.

5 Experimental Results and Discussion

The validation results are reported on the real-world dataset, named School Buses [7], consisting of 145 trajectories of two school buses collecting and delivering students around Athens metropolitan area in Greece for 108 distinct days.

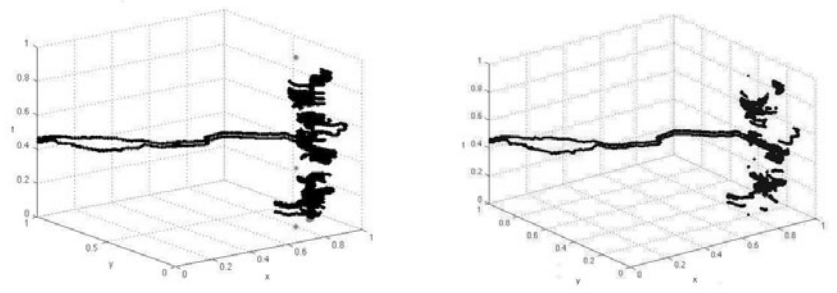
Let $< U, A >$ be the information system. U is the ST School Buses dataset, normalized and with some injected only temporal outliers (Figure 1(a)):

$$\mathrm{U} = \{p_i \equiv (z_{i,1}, z_{i,2}, z_{i,3}) \in [0,1]^3,\ i = 1, ..., N\}$$

where $(z_{i,1}, z_{i,2})$ are cartesian coordinates of the *i-th* object, $z_{i,3}$ its time-stamp. In this case, $A = \{x, y, t\}$ is the attribute set.

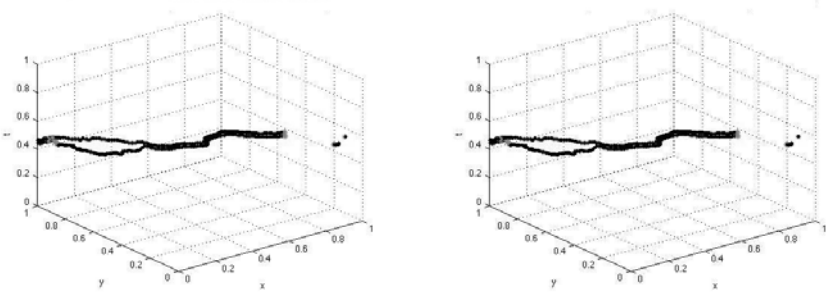
Spatial Rough Outlier Set Extraction from U
We want to describe $O \subseteq U$ as: $< \underline{B}(O), \overline{B}(O) >$ where, in this case, $B = \{x, y\} \subseteq A$, constituted by the spatial attributes. Specifically, the lower approximation and boundary at last step of *spatial-ROS* are represented and shown in Figure 2(a), where boundaries are reported in gray color. Many interesting objects should be missed without keeping into account the boundary.

Spatio-temporal Rough Outlier Set Extraction from U
We want to describe $O \subseteq U$ as: $< \underline{B}(O), \overline{B}(O) >$ where, in this case, $B = \{x, y, t\} = A$, so we are describing the *ST-ROS*. The ST-outliers will be the more relevant spatial and temporal outliers (see injected temporal outliers marked as gray stars in Figure 1(a)). The lower approximation includes the most part of the spatial and temporal outliers, while the upper includes the remaining part of the temporal outliers and some other spatial outliers. Figure 3(a) shows the lower approximation, while Figure 3(b) shows the lower approximation with boundaries in gray color.



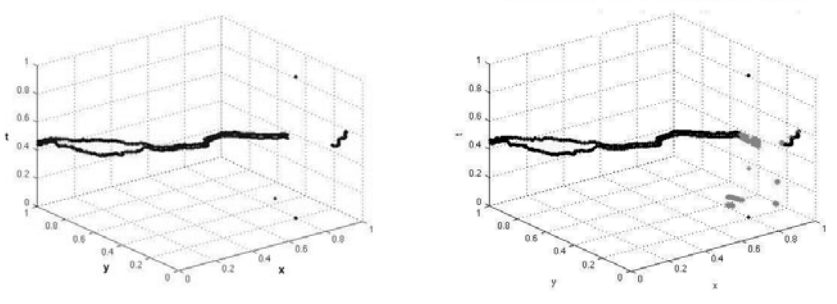
(a) (b)

Fig. 1. School Buses Testing Subset: (a) Injected Temporal Outliers (b) Kernel Set



(a) (b)

Fig. 2. Last Step of S-Rough Outlier Set: (a) Lower Approximation U Boundary from Dataset (b) Lower Approximation U Boundary from Kernel Set



(a) (b)

Fig. 3. Last Step of ST-Rough Outlier Set: (a) Lower Approximation (b) Lower Approximation U Boundary

Spatial Rough Outlier Set Extraction from Kernel Set
Figure 1(b) shows the *Kernel Set* of School Buses testing subset. Now, we want to show benefits of considering this set, comparing *S-ROS* extracted by the universe U with one extracted by the *Kernel Set*.

Starting from the *Kernel Set* and selecting only spatial components, the *ROS* is built by our approach *ROSE*. At this aim, let $< U, A >$ be the information system, with $U = K$. Then $B = \{x, y\}$. Figure 2(b) shows the lower approximation with boundaries in gray color at the last iteration. Comparing these results with the last iteration of *ROSE* for the extraction of the *Spatial Rough Outlier Set* from the entire universe U, shown in Figure 2(a), we can appreciate that the results are the same with a considerable computational time benefit.

Table 1. (a) Spatial and (b) Spatio-Temporal Outlier Detection - Quantitative Evaluation of Algorithms - Chosen Initial Centroids

Methods	α **Index**	ρ **Index**	γ **Index**	DB **Index**
$ROSE$	0.9836	0.0164	0.9987	N.A.
$RFCM$	0.5448	0.4551	0.9250	0.0736
$RPCM$	0.4725	0.5274	0.7919	1.1077
$RFPCM$	0.5645	0.4354	0.9007	0.8983

Methods	α **Index**	ρ **Index**	γ **Index**	DB **Index**
$ROSE$	0.8941	0.1059	0.9514	N.A.
$RFCM$	0.3549	0.6450	0.6444	1.8066
$RPCM$	0.3283	0.6716	0.5914	1.1077
$RFPCM$	0.3651	0.6348	0.6618	1.3299

6 Quantitative Measures and Indices

In this section, we use performance indices as introduced by Maji and Pal in [9] such as α, ρ and γ indices, as well as the *DaviesBouldin (DB)* measure as introduced in [1], to evaluate the performance of ROSE compared with some other *rough–fuzzy* clustering algorithms [9], i.e.: RFCM - Rough Fuzzy C-Means, RPCM - Rough Possibilistic C-Means, RFPCM - Rough Fuzzy Possibilistic C-Means. Parameter setting: $c = 2$ (cluster number, i.e. *inlier cluster* and *outlier cluster*), ω and $\tilde{\omega}$ (importance of lower and boundary) both equal to 0.5. We report only the final prototypes of the best solution, obtained for a particular choice of initial centroids. Table 1(a) and Table 1(b) report the best results obtained for RFCM, RPCM and RFPCM. Table 1(a) and Table 1(b) compare the performance of these algorithms with respect to α, ρ, γ and $DBindex$ in Spatial and ST-Outlier Detection respectively. Although the hybridization versions of c-means algorithm were not designed as outlier detectors, generate good prototypes for $c = 2$. In Spatial Outlier Detection, the RFPCM provides the best results and the results of other two are quite similar to that of the RFPCM; while in ST-Outlier Detection, the RPCM outperform them. The proposed ROSE algorithm performs better than RFCM, RPCM and RFPCM algorithms, both in terms of qualitative measures and of outliers detected, as shown in figures 3(b) and 2(a).

7 Conclusions

The paper reports a new rough set based outlier detection method, called *ROSE*, that has been theoretically grounded based on a definition of Outlier Set as Rough Set. The results of the proposed method have been shown and have been also compared with some other *rough–fuzzy* clustering algorithms, incorporating the concepts of rough sets, producing reasonable results both from quantitative and qualitative standpoints. A definition of a new set, called *Kernel Set*, has been also provided. The *Kernel Set* is a subset of U, significative to outlier detection. It has been shown that this set is able to detect the same outliers with less computational time.

References

1. Bezdek, J.C., Pal, N.R.: Some new indexes for cluster validity. IEEE Trans. Syst., Man, Cybern. B, Cybern. 28(3), 301–315 (1988)
2. Yao, Y.Y.: Two views of the theory of rough sets in finite universes. International Journal of Approximate Reasoning 15, 291–317 (1996)
3. Nguyen, T.T.: Outlier Detection: An Approximate Reasoning Approach. Springer, Heidelberg (2007)
4. Chen, Y., Miao, D., Wang, R.: Outlier Detection Based on Granular Computing. Springer, Heidelberg (2008)
5. Jiang, F., Sui, Y., Cunge: Outlier Detection Based on Rough Membership Function. Springer, Heidelberg (2006)
6. Pawlak, Z.: Rough Sets, Theoretical Aspects of Reasoning about data. Kluwer, Dordrecht (1991)
7. Frentzos, E., Gratsias, K., Pelekis, N., Theodoridis, Y.: Nearest Neighbor Search on Moving Object Trajectories. In: Anshelevich, E., Egenhofer, M.J., Hwang, J. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 328–345. Springer, Heidelberg (2005)
8. Birant, D., Kut, A.: Spatio-Temporal Outlier Detection in Large Databases. Journal of Computing and Information Technology 14(4), 291–297 (2006)
9. Maji, P., Pal, S.K.: Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices. IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics 37(6) (December 2007)
10. Albanese, A., Petrosino, A.: A Non Parametric Approach to the Outlier Detection in Spatio-Temporal Data Analysis. In: D'Atri, et al. (eds.) Springer book Information Technology and Innovation Trends in Organizations (2011)