Rough Sets, Kernel Set and Spatio-Temporal Outlier Detection

Alessia Albanese, *Member, IEEE,* Sankar K. Pal, *Fellow, IEEE,*
and Alfredo Petrosino, *Senior, IEEE*

Abstract

Nowadays, the high availability of data gathered from wireless sensor networks and telecommunication systems, has drawn the attention of researchers on the problem of extracting knowledge from spatio-temporal data. 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 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. We have also introduced a new set, named Kernel Set, that is a subset of the original dataset, which is able to describe the original dataset both in terms of data structure and of obtained results. Experimental results on real world datasets demonstrate the superiority of ROSE, both in terms of some quantitative indices and outliers detected, over those obtained by various rough fuzzy clustering algorithms and by the state-of-the-art outlier detection methods. It is also demonstrated that the kernel set is able to detect the same outliers set but with less computational time.

Index Terms

Spatiotemporal Data, Outlier Detection, Spatiotemporal Uncertainty Management, Rough Set and Granular Computing.

A. Albanese and A.Petrosino are with Department of Applied Science, University of Naples Parthenope, Naples, ITALY
S. K. Pal is with Indian Statistical Institute, Kolkata, INDIA.

The work was done by S.K. Pal as a J.C. Bose Fellow of the Govt. of India.

Digital Object Indentifier 10.1109/TKDE.2012.234

I. Introduction

Spatio-temporal data mining is a growing research area dedicated to the discovery of hidden knowledge in large spatio-temporal databases, mainly through detecting periodic and/or frequent patterns and outliers. Particularly, outlier detection finds its applications in a broad spectrum of fields, such as fraud detection, intrusion detection in computer networking, and detecting motion or abnormal regions in image processing. The presence of outliers makes the modeling difficult due to the discordance the outliers introduce into the data; in this sense, the outlier detection task is attractive for two main reasons: the isolation of outliers, as a preventive step, can improve the performance of the predictive modeling by offering better data quality; on the contrary, the identification of outliers can be the main goal of the analysis as, for example, in fraud detection. The most investigated approaches for outlier detection include: 1) distribution-based approaches that make use of standard statistical distribution to model the data declaring as outliers the objects that deviate from the model; 2) depth-based techniques which are based on computational geometry and compute different layers of convex hulls declaring as outliers the objects belonging to the outer layers; 3) distance-based approaches which compute the proportion of database objects that are a specified distance from a target object; 4) density-based approaches which assign a weight to each sample based on their local neighborhood density. A different classification is based on the outlier detection output and divides into: labeling and scoring techniques. Labeling methods partition the data into two non-overlapping sets (outliers and non-outliers) and scoring methods offer a ranking list by assigning to each datum a factor reflecting its degree of outlierness. These former methods exploit a hard decision about the sets, the latter ones deal with a sort of soft decision about the membership of each datum to the set. The proposed method is the first rough method that improves and upgrades the ”scoring methods”, proposing an effective soft granular computing based solution exploiting the uncertainty region (boundary) in order to obtain more reliable results. Indeed, rough-set theory [41] is a paradigm to deal with uncertainty, vagueness and incompleteness and it is proposed for indiscernibility in classification according to some similarity. Rough sets were extensively used for data mining but rarely for outlier detection in general-domain, the same for spatio-temporal specific-domain is hardly ever addressed and never for outlier detection in spatio-temporal data. In some sense, the few available outlier detection approaches interpret the rough set theory from the ”operator-oriented point of view” [53]. In contrast, our method, called ROSE (*Rough Outlier Set Extraction*), exploits the set-oriented point of view of rough set theory in order to define the concept of outlier in terms of its lower and upper approximations (*rough outlier set*), keeping into account those objects that can neither be ruled in nor ruled out as members of the target concept. Performance of ROSE in detecting outliers is found to be superior to best rough–fuzzy clustering algorithms in terms of various quantitative indices and to several state-of-the-art outlier detection methods.

Moreover, we introduce the concept of *kernel set*. Given a dataset, the kernel set is a selected subset of elements able to describe the original dataset in terms of dataset structure. The paper includes two different versions of the ROSE algorithm on a test dataset: one adopting, as input set, the entire set and the other adopting its kernel set. Experimental results show the advantages of considering the kernel set, in term of computational time, by comparing the *rough outlier set* extracted by the original dataset with one extracted by the kernel set.

The paper is organized as follows. In section II an overview on outlier detection approaches is given. Section III reports some preliminaries about rough set theory relevant to this work, indeed our approach is rough set based. Section IV introduces the problem and reports the new rough set approach ROSE to extract the spatio-temporal rough outlier set. Section V introduces the new set kernel set. Sections VI-A, VI-B, VI-C present executed tests on three real world (benchmark and test) datasets and the performance evaluation of the algorithm. Finally, conclusion remarks are given in Section VII about ongoing and future work.

II. Related Work

Most of the existing surveys on anomaly detection focus on a particular application domain or on a single research area, while the surveys, like [25], [14], [36] and two more recent brief surveys [44] and [49] are complete works that give the state-of-the-art of anomaly detection techniques. They group anomaly detection into multiple categories and discuss techniques under each category. The discussed research issues include many topics to be taken into account to choose the appropriate outlier detection approach: (i) the detection method (parametric, i.e., distribution based [7], depth-based [30], [29], [20]; graph-based methods [33], [48]; non-parametric, i.e., distance-based [31], [4], [43], [46]; density-based [12], [45], [54], [40], [55], [6]; clustering-based methods [24], [1], [21], [38]; and semi-parametric, i.e., neural network-based, support vector machine-based techniques); (ii) the nature of the detection algorithm, i.e., supervised, unsupervised, semi-supervised detection; (iii) the nature of data, i.e., numerical, categorical, [11], [18] or mixed data [32], [37]; (iv) the size and the dimensionality of the dataset, [2], [57], [47]; (v) the nature of the target application [13], [22], [5]. This concerns the outlier detection methods in general domain. Concerning with specific spatio-temporal (ST) domain, only a few outlier detection methods have been proposed. Wu, et al. [52] propose a spatio-temporal outlier detection algorithm called *Outstretch*, which discovers the sequences of spatial outliers over several time periods. Birant and Kut [9] present a ST-outlier detection approach based on clustering concepts called *ST-DBSCAN* which is an improved version of the clustering technique *DBSCAN* [45] that supports also temporal aspects. Cheng and Li [17] further propose a four-step approach to detect spatio-temporal outliers, i.e., classification, aggregation, comparison and verification. Wang et al. [50] also propose an approach to outlier detection in spatio-temporal domain. In a more recent work, Liu et al. [34] deal with the problem of detecting spatio-temporal outliers and causal relationships among them from traffic data streams.

Rough set theory has been recently introduced in the ST-domain literature for different aspects. In ST-domain, using the notion of rough sets, Bittner [10] defines approximations of ST-regions and relations between those approximations. Concerning outlier detection in general domain some works have been proposed: Nguyen [39] discusses a method for the detection and evaluation of outliers, as well as how to elicit the background domain knowledge from outliers using multi-level approximate reasoning schemes; Y. Chen et al. [15] demonstrates an application of granular computing model using information tables for the outlier detection; F. Jiang et al. [27] proposes a definition for outliers based on a rough outlier factor (ROF) as degree of outlierness for every object with respect to a given subset of universe. More recently, the same authors [28] propose a novel definition of outliers - sequence-based outliers - in information systems of rough set theory and an algorithm to find out such outliers. Concerning spatio–temporal outlier detection, no rough set theory based approach has been proposed up to now.

III. Rough Set Theory

Rough set theory, proposed by Pawlak [41], is a new and highly accepted paradigm used to deal with uncertainty, vagueness and incompleteness. The main idea is based on the indiscernibility relation that describes indistinguishability of objects. Rough Set Theory (RST) can be approached as an extension of the Classical Set Theory, for use when representing incomplete knowledge.

Concepts are represented by lower and upper approximations, according to which rough set methodology focuses on approximate representation of knowledge derivable from data [42].

A. *Indiscernibility and Set Approximation*

Let U be the universe of the discourse and A be the finite and non empty set of attributes, then $S = \langle U, A \rangle$ is an information system. Let B a subset of A. With every subset of attributes $B \subseteq A$, an equivalence relation I_B on U can be easily associated:

$$I_B = \{(p, q) \in U \times U \ / \ \forall a \in B, a(p) = a(q)\} \tag{1}$$

I_B is called ***B*–*indiscernibility relation***.
If $(p, q) \in I_B$, then objects p and q are indiscernible from each other by attributes B. The equivalence classes of the partition induced by the ***B*–*indiscernibility relation*** are denoted by $[p]_B$. These are also known as *granules*. We can approximate any subset X of U using only the information contained in B by constructing the lower and upper approximations of X. The sets $\{p \in U : [p]_B \subseteq X\}$ and $\{p \in U : [p]_B \cap X \neq \emptyset\}$, where $[p]_B$ denotes the equivalence class of the object $p \in U$ relative to I_B, are called the ***B*–*lower*** and ***B*–*upper approximation*** of X in S and respectively denoted by $\underline{B}(X), \overline{B}(X)$. The objects in $\underline{B}(X)$ can be certainly classified as members of X on the basis of knowledge in B, while objects in $\overline{B}(X)$ can only be classified as possible members of X on the basis of B.

IV. Spatio-Temporal Outlier Detection

In this section, the spatio-temporal outlier detection problem is introduced by providing the problem formalization from a theoretical standpoint, together with its computational solution. A strict distinction between the spatial and temporal components is proposed in our definition of the problem. This may result useful in many contexts, e.g., datasets which are characterized by only spatial information (we intend for spatial not only location information but also features detected at each location), where the temporal information is implicitly attached or is not present at all. In all such cases, the distinction allows us to consider just the spatial component, saving space and time. In this way, time can be differently weighted for finding more efficiently temporal outlierness and for handling different scenarios where spatial and temporal components get different importance in the dataset. The proposed approach finds also spatio-temporal outliers.

A. Problem Definitions

Let us consider an information system $S =< U, A >$ with U a spatio-temporal 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 the attribute set. In the following, we consider that at least three attributes must be present, i.e., the spatial attributes and the temporal one.

Given U, an integer $n > 0$ and a measure $d_{p_i}(U)$, defined over every $p_i \in U$, the general definition of the **Outlier Detection Problem** is as following:

Definition 1: 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

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

According to this definition, the concept of measure is used to determine the degree of dissimilarity of each object with respect to all others. Then, the *n–Outlier Set* can be formally defined as:

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 the definition 2 it follows that $\tau = d_{p_n}(U)$ is the **outlierness threshold**, i.e., the minimum value among the n maximum values of measures computed in U (associated to objects belonging to the *n–Outlier Set*), i.e.,

$$\tau = inf\{max_1(d_p(U), d_q(U)), .., max_n(d_p(U), d_q(U))\}, \forall p, q \in U \tag{2}$$

Starting from the definition of spatial outlier and temporal outlier due to Birant and Alp [9] asserting: ”a spatial outlier is a spatial referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood”, and ”a temporal outlier is an object whose non-spatial attribute value is significantly different from those of other objects in its temporal neighborhood”, we propose the following definitions applied only to spatio-temporal data:

Definition 3: A Spatial Outlier (**S–Outlier**) is an object whose spatial attribute value is significantly different from those of its closer objects (spatial neighborhood).

In this framework, the *Spatial Outlier* definition corresponds to:

Definition 4: Given U, an integer $n > 0$ and a measure on spatial component $d^s_{p_i}(U)$, defined over every $p_i \in U$, an object $p \in U$ is a **S–Outlier** $iff \;\; d^s_p(U) \geq \tau$ where τ is defined in (2).

Following definition 4, it holds that:

Proposition 1: A Spatial Outlier (**S–Outlier**) is an object that belongs to the spatial *n–Outlier Set* indicated by O_s.

Similarly, we propose the following definition of *Temporal Outlier*, applied to only spatio-temporal data:

Definition 5: A Temporal Outlier (**T–Outlier**) is an object whose temporal attribute value is significantly different from those of its closer objects (temporal neighborhood).

In this framework, the *Temporal Outlier* definition corresponds to:

Definition 6: Given U, an integer $n > 0$ and a measure on temporal component $d^t_{p_i}(U)$, defined over every $p_i \in U$, an object $p \in U$ is a **T–Outlier** $iff \;\; d^t_p(U) \geq \tau$ where τ is defined in (2).

Equally, following definition 6, it holds that:

Proposition 2: A Temporal Outlier (**T–Outlier**) is an object that belongs to the temporal *n–Outlier Set* indicated by O_t.

Definition 3 states that a spatial outlier has no objects or a small group of objects in its spatial neighborhood. The same is valid for a temporal outlier according to Definition 5. Following both definitions the following holds:

Definition 7: A Spatio–Temporal Outlier (**ST–Outlier**) is an object that respects both the definitions above.

To obtain a real degree of outlierness, an appropriate measure should be associated to each object; i.e., the Euclidean distance computed between each object and all the other objects belonging to U. In real applications, characterized by an huge amount of data, this idea is unfeasible due to its high computational complexity $(O(N^2))$ where $N = |U|$.

We preserve two aims: on one hand, we exploit the well-known outlier definition based on k-nearest neighbors [43], in order to associate to each object, a measure based on the distances among the object itself and its k-nearest neighbors rather than all N objects with $k \ll N$; on the other hand, we make use of a pruning strategy that discards objects that surely cannot belong to the *n–Outlier Set*, in order to address the problem of alleviating the computational cost.

In a Spatio–Temporal context, the measure associated to each object is based upon the distances from its spatial k-nearest neighbors and its temporal k-nearest neighbors [3]. Precisely:

$$d_p^{s,t}(U) = \alpha \cdot d_p^s(U) + \beta \cdot d_p^t(U) \tag{3}$$

where:

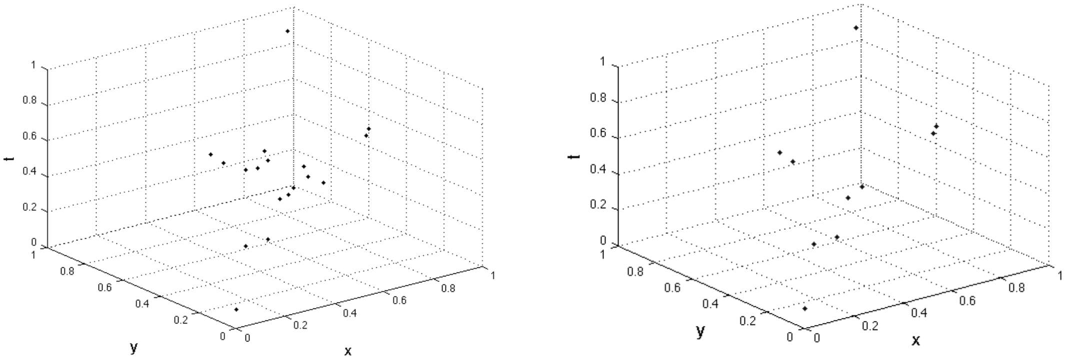
$$d_p^s(U) = \sum_{j=1}^{k} d^s(p, N^s(p, p_j)), \quad \forall p \in U \tag{4}$$

$$d_p^t(U) = \sum_{j=1}^{k} d^t(p, N^t(p, p_j)), \quad \forall p \in U \tag{5}$$

$k > 0$ is the number of nearest neighbors to keep into account, $N^s(p, p_j)$ and $N^t(p, p_j)$ are, respectively, the j-th spatial nearest neighbor and the j-th temporal nearest neighbor of p, and α, β weight such that $\alpha + \beta = 1$. Definition 1, that introduces the Outlier Detection Problem, defines the Spatio-Temporal Outlier Detection Problem, by selecting a measure as in (3).

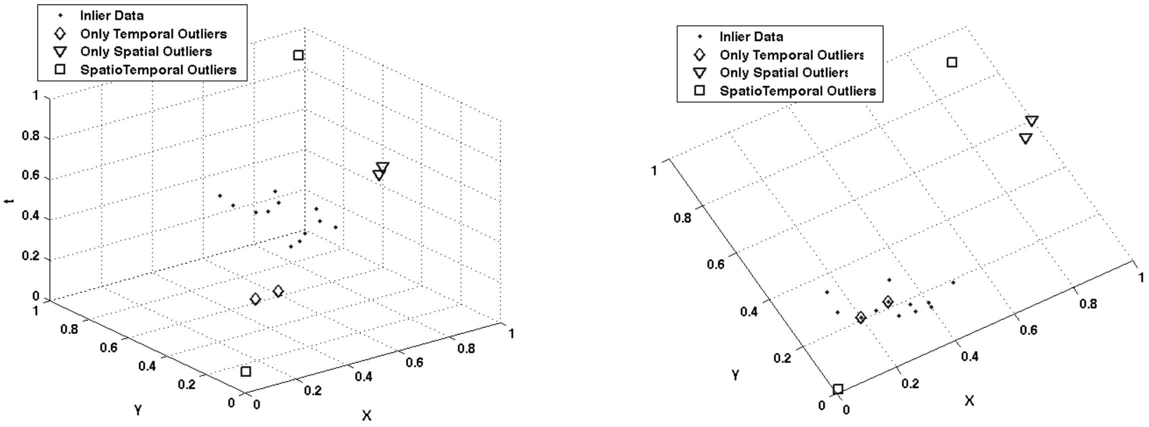
To better illustrate the meanings of the previous and the following definitions, let us consider the Example, a spatio-temporal dataset E = $\{p_i \equiv (z_{i1}, z_{i2}, z_{i3}) \in [0,1]^3, \quad i = 1, ..., 18\}$ where p_i is a 3-dimensional feature vector and $A = \{a_1, a_2, a_3\}$ is the essential attribute set, i.e., a_1, a_2 are the spatial attributes and a_3 is the temporal attribute.

E is a labeled dataset, containing 18 elements as reported in Table I of Appendix and plotted in the Figure 1. By fixing $k = 3$ and $n = 4$, the outlier sets (spatial, temporal outlier sets), on the basis of the previous definitions, are computed as follows. A *4–Spatial Outlier Set* $O_s \subseteq E$ is the set of objects $p \in E$ that significantly deviate from the rest of data with respect to the spatial component, i.e., $O_s = \{(0.95, 0.55, 0.50), (1, 0.60, 0.50), (0.01, 0.01, 0.1), (0.9, 0.9, 0.95)\}$. A *4–Temporal Outlier Set* $O_t \subseteq E$ is the set of objects $p \in E$ that significantly deviate from the rest of data with respect to the temporal component, i.e., $O_t = \{(0.01, 0.01, 0.1), (0.20, 0.21, 0.3), (0.30, 0.22, 0.3), (0.9, 0.9, 0.95)\}$. If $n = 2$, a *2–Spatio–Temporal Outlier Set* $O_{s,t} \subseteq E$ is the set of objects $p \in E$ that significantly deviate from the rest of data with respect to the spatial and the temporal component, i.e., $O_{s,t} = \{(0.01, 0.01, 0.1), (0.9, 0.9, 0.95)\}$. O_s, O_t and $O_{s,t}$ are shown in figure 2(a) as diamond and square, as triangle and square and only square respectively. In figure 2(b), a 2D projection has



(a) (b)

Fig. 1. (a) The example dataset E and (b) its kernel set.



(a) (b)

Fig. 2. Example dataset: (a) Detected outlier sets (b) their xy-projection.

been reported in order to better visualize that the spatial outliers and spatio-temporal outliers are spatially far from the rest of data.

B. Rough Outlier Set Extraction (ROSE)

1) Theory: The goal of our approach is to exploit the rough set theory to define the *Outlier Set* such as a *Rough Outlier Set*.

Let $S =< U, A >$ be an information system with U a spatio temporal normalized dataset and A

its attribute set. If $n > 0$ is the required outlier number, we want to describe $O \subseteq U$ *(n–Outlier Set)* as

$$< \underline{B}(O), \overline{B}(O) > (Rough\ n - Outlier\ Set) \tag{6}$$

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 *B–Upper approximation* $\overline{B}(O)$ is defined as the set of possible members of O on the basis of the knowledge in B.
With 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 (3) 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, s, and/or on time, t, and at iteration, i. The attribute subsets B include spatio-temporal attributes, or only spatial and only temporal attribute in order to define spatio-temporal outlier set, or only temporal set and only spatial outlier set respectively. In this framework, the *B–Lower* and *B–Upper approximations* at iteration i can be defined as follows:

Definition 8: 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

$$\tau_i = inf\ \{max_1^i\ (\overline{\omega}_{G_j}, \overline{\omega}_{G_k}), ..., max_n^i\ (\overline{\omega}_{G_j}, \overline{\omega}_{G_k})\},\ \forall\ G_j, G_k \subseteq U \tag{7}$$

Definition 9: The *B–upper approximation* $\overline{B}_i(O)$ of n-Outlier Set O, at iteration i, is:

$$\overline{B}_i(O) = \{G_j \subseteq U : \overline{\omega}_{G_j} > \overline{\tau}_i\}$$

where

$$\overline{\tau}_i = \tau_{i-1},\ \forall\ \ i >= 2 \tag{8}$$

The threshold τ_1 is computed as the minimum value among the n higher values of weights assigned to the granules at first iteration, then, at second iteration, τ_2 will be the new minimum value among the new n higher values of weights re–assigned to the granules at second iteration and $\overline{\tau}_2 = \tau_1$.

The iterative procedure will stop when the following convergence criterion will be satisfied:

Lemma 1: The construction of the lower approximation $\underline{B}(O)$ or the upper approximation $\overline{B}(O)$ of an *n–Outlier Set* O converges if it exists an index k such that the threshold does not vary anymore, i.e.,

$$if \quad \overline{\tau}_k = \tau_k \;\; then \;\; \underline{B}_k(O) = \overline{B}_k(O) \tag{9}$$

Proof: See Appendix.

Hence, the *Rough n–Outlier Set* is represented by:

$$< \underline{B}_{k-1}(O), \overline{B}_{k-1}(O) > \tag{10}$$

In case of $B = A$ (every attribute is considered), the granules are:

$$\forall p_j \in U : \{p_j\} \equiv G_j \qquad \forall j = 1, ..., N \tag{11}$$

so both spatial and temporal components are taken into account.

As instance, let us consider the labeled Example dataset. In this case, the attribute set is $A = \{x, y, t\}$, i.e., x and y are cartesian coordinates and t is the temporal component.

Spatial Outliers

In the case of spatial outliers, the reduction is made in terms of temporal component, i.e., $B = \{t\}$; so we have the following partition of the universe:

$$I_B = I_{\{t\}} = \{\{p_1, p_2\}, \{p_3, p_9\}, \{p_4\}, \{p_5\}, \{p_6\}, \{p_7, p_8\}, \{p_{10}\},\\ \{p_{11}\}, \{p_{12}\}, \{p_{13}\}, \{p_{14}\}, \{p_{15}\}, \{p_{16}\}, \{p_{17}\}, \{p_{18}\}\}$$

The concept of *Spatial Outlier* can be appropriately defined on the basis of knowledge in $B = \{t\}$. Specifically, the *B–lower approximation* of the *Spatial Outlier Set* O_s, is composed by the granules completely included into O_s, i.e., $\underline{B}(O_s) = \{\{p_7, p_8\}, \{p_{17}\}, \{p_{18}\}\}$ and the *B–upper approximation* is composed by the granules that have non trivial intersection with O_s, i.e., $\overline{B}(O_s) = \{\{p_7, p_8\}, \{p_{17}\}, \{p_{18}\}\}$. In this case, the upper approximation does not give any additional information.

Temporal Outliers

In the case of temporal outliers, the reduction is made by spatial components, i.e., $B = \{x, y\}$, getting:

$$I_B = I_{\{x,y\}} = \{\{p_1, p_{12}\}, \{p_2, p_{13}\}, \{p_3\}, \{p_4\}, \{p_5\}, \{p_6\}, \{p_7\},$$
$$\{p_8\}, \{p_9\}, \{p_{10}\}, \{p_{11}\}, \{p_{14}\}, \{p_{15}\}, \{p_{16}\}, \{p_{17}\}, \{p_{18}\}\}$$

The concept of *Temporal Outlier* can be equivalently get on the basis of knowledge in $B = \{x, y\}$. The *B–lower approximation* of the *Temporal Outlier Set* O_t, is composed by the granules completely included into O_t, i.e., $\underline{B}(O_t) = \{\{p_{17}\}, \{p_{18}\}\}$ and the *B–upper approximation* is composed by the granules that have a non trivial intersection with O_t, i.e., $\overline{B}(O_t) = \{\{p_1, p_{12}\}, \{p_2, p_{13}\}, \{p_{17}\}, \{p_{18}\}\}$. In this case, the notion of rough set arises; indeed the upper approximation give additional information.

2) ROSE Algorithm: The *Rough Outlier Set Extraction (ROSE) Algorithm* is designed to receive as input the universe U, the number k of nearest neighbors and the number n of outliers to detect. The output of the (iterative) procedure is the *Rough Outlier Set* (Upper, Lower Approximation and Negative Region). The algorithm selects, at each step, a small subset of objects, called WorkingSet, from the overall dataset U. To this aim, ExtractElements extracts a number of elements equal to a fixed percentage of the cardinality of U that has to be greater than k. The following main steps are computed. For all selected objects, the procedure computes the Euclidean distances among the objects in the WorkingSet and all the objects of U, considering the spatial components, the temporal components or both of them (general case $B = A$) depending upon the chosen attribute subset B with respect to the Rough Outlier Set has been computed. Algorithm ROSE related to the general case has been shown. UpdateUpperApprox and UpdateLowerApprox at first iteration, create the same set of n top outliers at that step, i.e., the n objects that have an associated measure higher than the others. Then, at next iterations, UpdateUpperApprox and UpdateLowerApprox compute the Lower and Upper approximation of *Rough Outlier Set*, using the τ (computed by LowerWeight) and τ_prev thresholds as respectively defined in (7) and (8). At each iteration i, the pruning strategy selects objects from U that have their measure under the computed threshold in order to build the Negative Region. The LowerWeight function computes the τ threshold (and consequently τ_prev is the saved value of τ before to be updated). At each iteration, the thresholds have been computed as the weight

Algorithm 1 ROSE - Rough Outlier Set Extraction

```
$begin \quad ROSExtraction(U, n, k)$
$LowerOutlierSet = null; UpperOutlierSet = null$
$w_{s,t,k}(q) = 0$
$\tau\_prev = 0; \tau = 0$
$WorkingSet = ExtractElements(U)$
while $(WorkingSet != null)$ do
  for $p \in U$ do
    for $q \in WorkingSet$ do
      if $(LowerOutlierSet == null \ and \ UpperOutlierSet == null)$
      $or \ (w_{s,t,k}(q) \geq \tau\_prev))$ then
        $d_s(p, q) = CalculateSpDistance(p, q)$
        $d_t(p, q) = CalculateTempDistance(p, q)$
        $BuildTreeKNN(p, q, d_s, d_t, k)$
      else
        $AddNegativeRegion(p)$
      end if
    end for
  end for
  for $q \in WorkingSet$ do
    $w_{s,t,k}(q) = CalculateWeight(q)$
    $UpperOutlierSet = UpdateUpperApprox(\tau\_prev, n, w_{s,t,k}(q))$
    $LowerOutlierSet = UpdateLowerApprox(\tau, n, w_{s,t,k}(q))$
  end for
  $\tau = LowerWeight(UpperOutlierSet)$
  if $(\tau != 0)$ then
    $\tau\_prev = \tau$
  end if
  $U = U - WorkingSet$
  $WorkingSet = ExtractElements(U)$
end while
$end \quad ROSExtraction()$
```

minimum value among the weight maximum n values, as defined in equation 7). The difference set between the Universe set and the Negative Region is the Kernel Set.

3) ROSE Algorithm - Time Complexity : The ROSE algorithm has worst-case time complexity $O(|U|^2)$, but practical complexity $O(|U|^{1+d})$, with $d < 1$ and U the universe.

V. THE KERNEL SET: RELEVANCE TO OUTLIER DETECTION

The present section introduces a new set, called kernel set, and states that it is a relevant set for outlier detection. Given a dataset U, the kernel set is a subset, of lower cardinality, that can be used instead of U, in order to detect the same outlier set. The time complexity reduction of the use of kernel set is quantified by measuring kernel set dimensionality over that of U.

A. Definition

Let us now define a new set, called *Kernel Set*, $K \subseteq U$, as a selected subset of the universe U that characterizes the overall dataset. Intuitively, this set is a subset of objects of U that maintains the general structure of the universe U. The Kernel Set is built by construction, in an iterative way, adding each object having specific properties.

Definition 10: 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 10 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 $d_q(U) < \tau$ 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 becomes less than τ also in K as in U. Otherwise, the global structure of the dataset should be altered.

Also, the *Kernel Set* is built for the Example dataset like:
$K = \{(0.01, 0.01, 0.1), (0.9, 0.9, 0.95), (0.95, 0.55, 0.5), (1.0, 0.6, 0.5), (0.2, 0.21, 0.3), (0.3, 0.22, 0.3), (0.3, 0.16, 0.55), (0.35, 0.15, 0.6), (0.15, 0.26, 0.76), (0.16, 0.34, 0.77)\}$.
This set is also reported in Figure 1(b). The Kernel Set contains all elements of the *Outlier Set*.

B. Properties

Let us start to prove the following propositions related to the new set.

Proposition 3: The measure computed in K is an upper bound of the measure computed in U such that:

$$d_p(U) \leq d_p(K), \quad \forall p \in U$$

$where\ d_p(U) = \sum_{j=1}^{k} d(p, N(p, p_j))\ and\ \ N(p, p_j)\ is\ the\ j\text{-}th\ nearest\ neighbor$ of p.
Proof: See Appendix.
The following proposition is valid:

Proposition 4: A **Kernel Set** contains the *n–Outlier Set*: $K \supseteq O$.

Proof: $\forall p \in O : d_p(U) > \tau \Rightarrow p \in K$
The proof clearly follows from definition of K. ■

Proposition 5: The Outlier Set O_K, computed starting from Kernel Set K is a superset of O computed from U:

$$O_K \supseteq O$$

Proof: See Appendix.

C. Significance to Outlier Detection

The *kernel Set* is a meaningful subset of the universe U with the following properties:

- *Kernel Set* is a subset with lower cardinality than U
- the ”same results” in terms of *rough outlier set* are obtained using *Kernel Set* instead of U
- *Kernel Set* can be considered as the model learned during a training phase.

In the following, we propose the comparison between the obtained results, in terms of rough outlier set, executing ROSE algorithm, once using, as input, the entire universe U and another time computed using, as input, the *kernel Set* K.

D. *Computational Benefits*

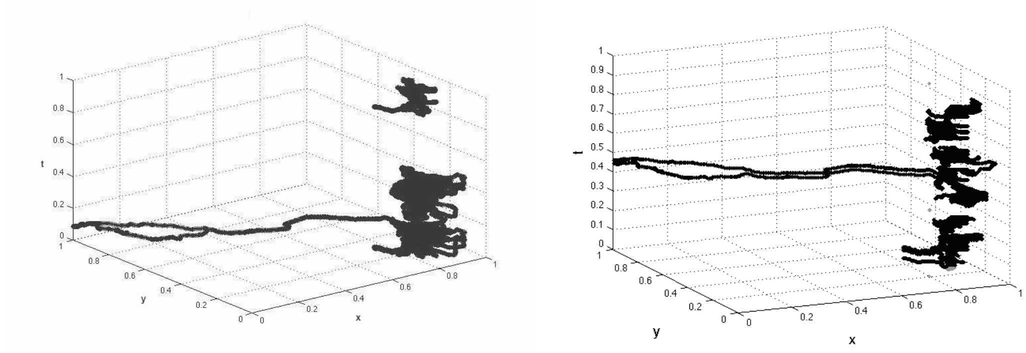
Let us consider the two versions (or runnings) of ROSE algorithm, in order to appreciate the computational benefits. At the first run, ROSE algorithm receives, as input, the entire dataset U, whilst at the second run, ROSE receives the kernel set K of U that is a subset of U. A computational benefit, coming from using kernel set instead of the entire universe, is derived. Indeed, $O(|U|^{1+d}) < O(|K|^{1+d})$, being $K \subset U$. To quantify the computational benefits coming from the use of the kernel set, we evaluate the dimensionality of kernel set K with respect to U. The experimental results have been provided in the following section VI-D.

VI. Experimental Results and Discussion

Our outlier detection method is based on rough set theory and is specific for spatio temporal data. At the best of our knowledge there is no rough approach to outlier detection for spatio temporal data to compare with. Hence, three different experimental tests have been executed. The first test is oriented to demonstrate the ability of the outlier detection algorithm and the role of the kernel set working on a real world spatio temporal dataset; the comparisons on this dataset are made using rough-fuzzy clustering methods. The second test is intended to compare our results with other outlier detection methods (also rough-oriented) for general domain on a UCI repository dataset. The third test is oriented to compare our performance with outlier detection methods (not rough approach) tailored for spatio temporal domain, on a spatio temporal dataset. Two subsections VI-D and VI-E end this section: one concerning an experimental evaluation of the dimension reduction percentage of the kernel set with respect to its starting dataset U and one concerning a sensitivity analysis about the parameters k and n of the algorithm.

A. *School Buses dataset*

For the first test, we make tests on a real-world dataset, named School Buses [19]. The dataset is publicly available and consists of 145 trajectories (about 69000 entries) of 2 school buses collecting and delivering students around Athens metropolitan area in Greece for 108 distinct days. The structure of each record is as follows: $\{obj\_id, traj\_id, date, time, lat, lon, x, y\}$ where obj_id is the school bus identification, $traj\_id$ is the unique trajectory identification, the date and time are the sampling timestamps every 30 seconds (date in $dd/mm/yyyy$ format and time in *hh:mm:ss* format), the (lat, lon) and (x, y) are the bus location, in WGS84 and in



(a) (b)

Fig. 3. School Buses dataset: (a) Normalized dataset (b) Testing subset with added temporal outliers highlighted in gray color.

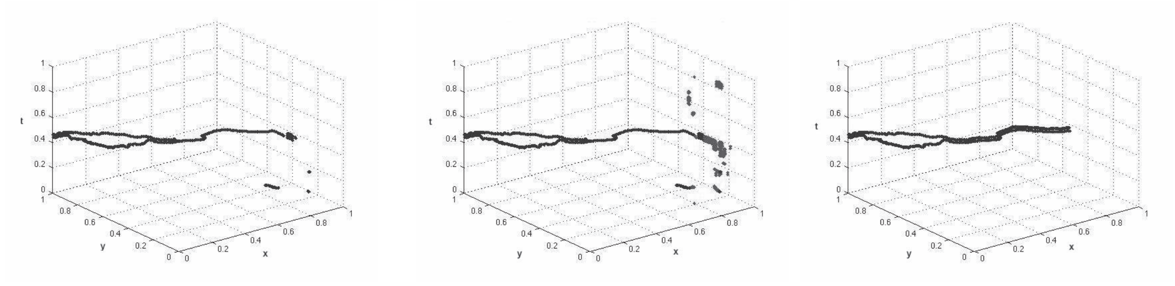
GGRS87 reference systems, respectively. In our case, the obj_id and $traj\_id$ are not considered, date and time fields are converted in just one field t consisting of a time string corresponding to the elements year, month, day, hour, minute and second. Moreover, the lat and lon are redundant and are not considered, since x and y give the same information. Hence, the normalized representation of the dataset is illustrated in figure 3(b): in a 3D cartesian reference system, x and y are the spatial coordinates and the third dimension is time t. In Figure 3(a) the trajectory map of School Buses is shown. In the following Figure 3(c), the testing dataset consisting of half of the original dataset (about 30000 entries) with some added temporal outliers is shown.

1) Rough Outlier Set Extraction - Spatial Rough Outlier Set Extraction from U : Let U denote the spatio-temporal normalized School Buses dataset

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

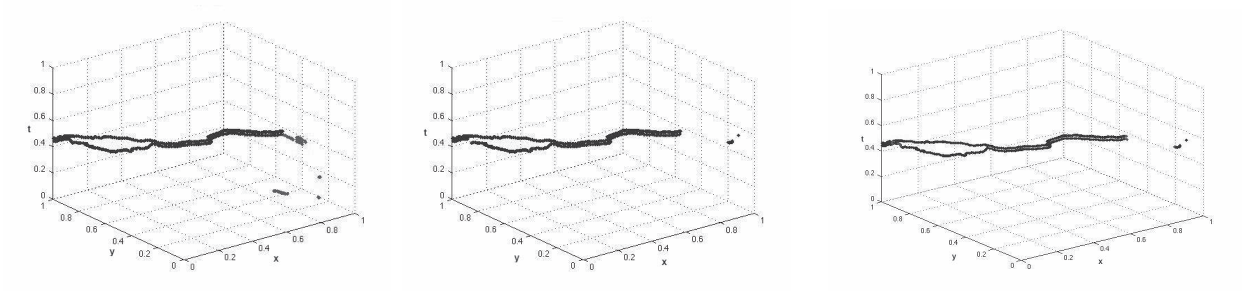
where $(z_{i,1}, z_{i,2})$ are cartesian coordinates of the *i–th* object, $z_{i,3}$ is the relative timestamp. Let $< U, A >$ be the information system, with the attribute set $A = \{x, y, t\}$, i.e., x and y are the spatial components and t is the temporal component.

We want to describe $O \subseteq U \quad (Outlier\ Subset)$ as the *rough outlier subset* $< \underline{B}(O), \overline{B}(O) >$ where $B \subseteq A$ is constituted by the spatial attributes, (x, y). Selecting only spatial components, the results of selected iterations, an $intermediate$ step, the *last–1* and the *last* one have been shown. Specifically, the lower, upper approximation (lower and boundary) at an intermediate step of *Spatial Rough Outlier Set* are represented and shown in Figure 4(a) and Figure 4(b), where boundaries are reported in gray color. Figure 4(c) and 5(a) show the lower, upper approximation (lower and boundary) at $last$-1 step, while Figure 5(b) and (c) shows the same approximations at $last$ step. In the last figure, we can see the advantages of keeping into account the boundary. Otherwise, many interesting objects (belonging to the boundary) should be missed.



(a) (b) (c)

Fig. 4. (a) Intermediate Step: Lower Approx (b) Intermediate Step: Lower Approx U Boundary (c) Last-1 Step: Lower Approx.

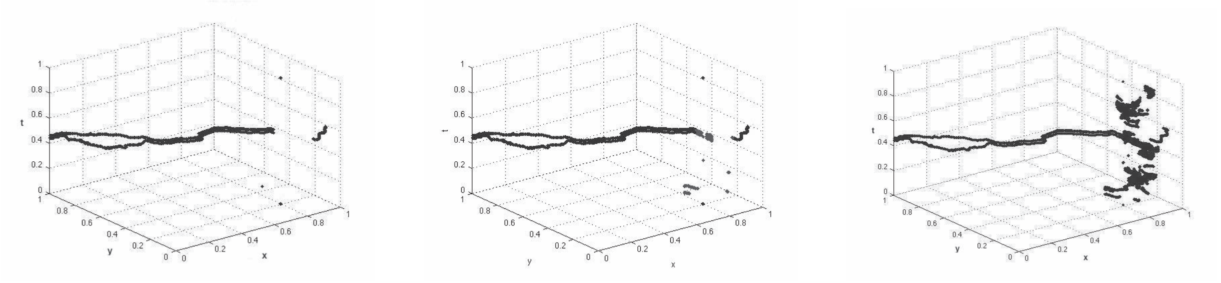


(a) (b) (c)

Fig. 5. (a) Last-1 Step: Lower Approx U Boundary (b) Last Step: Lower Approx (c) Last Step: Lower Approx U Boundary

2) Rough Outlier Set Extraction - Spatio-Temporal Rough Outlier Set Extraction from U: Let $< U, A >$ be the information system, with the attribute set $A = \{x, y, t\}$, i.e., x and y are the spatial components and t is the temporal component. Now we are considering $B = A$, so we are looking for *spatio–temporal Rough Outlier Set*.

The spatio-temporal outliers will be more relevant than spatial and temporal outliers (see temporal outliers injected in the Figure 3(b)). Hence, the lower approximation includes the most part of spatial and temporal outliers, while the upper approximation includes the remaining part of temporal outliers and some other spatial outliers have been detected. In this section, we show the lower, lower approximation with boundary at last step. Figure 6(a) shows the lower approximation, while Figure 6(b) shows the lower approximation with boundaries in gray color.

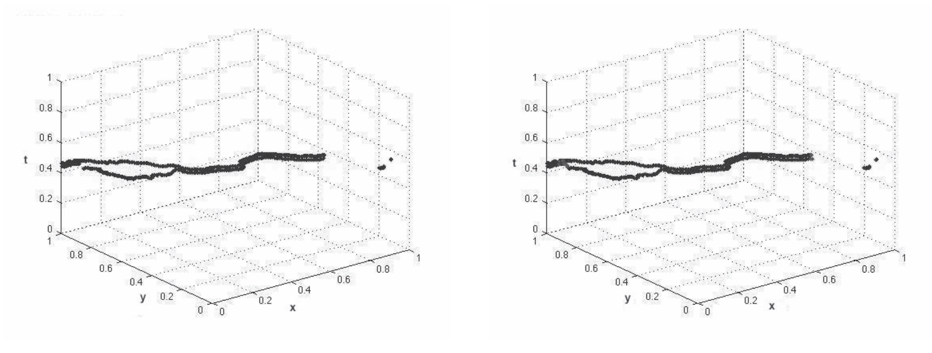


(a) (b) (c)

Fig. 6. (a) Last Step: Lower Approx (b) Last Step: Lower Approx U Boundary (c) School Buses Dataset: its Kernel Set.

3) Rough Outlier Set Extraction - Spatial Rough Outlier Set Extraction from the Kernel Set: The section reports the tests aimed to demonstrate the use of the Kernel Set. This set is a selected subset, able to describe the original dataset both in terms of data structure and in terms of obtained results. In particular, we want to show the advantages of using this set and the benefits of considering it. To this aim, we show the Rough Outlier Set extracted by the universe U and the Rough Outlier Set extracted by the Kernel Set. The results show the advantages of considering this set. Figure 6(c) shows the kernel set of School Buses dataset. Starting from the *Kernel Set*, the Rough Outlier Set is built by our approach ROSE. Let be $B \subseteq A$ constituted by the spatial attributes, i.e., (x, y). Selecting only spatial components, the results of last iteration of the test of Spatial Rough Outlier Set Extraction from the Kernel Set is reported. Figure 7(a) shows the lower approximation at the last iteration, while Figure 7(b) shows the lower approximation with boundaries in gray color. Thus, we compare these results with the last test of Rough Outlier Set Extraction from the entire Universe U, shown in Figure 5(c). Comparing Figure 5(c) and Figure 7(b) we can appreciate that the results are quite similar with an interesting computational benefit coming from considering the *Kernel Set* instead of the entire universe U.

4) Quantitative Measures and Indices: In this section, we use performance indices as introduced by Maji and Pal in [35] such as α index, ρ index and γ index, to evaluate the performance of our algorithm compared with *Hard C-Means* and with other *rough–fuzzy* clustering algorithms,



(a)

(b)

Fig. 7. ROSE Results from Kernel Set of School Buses Dataset - Last Step: (a) Lower Approx (b) Lower Approx U Boundary.

TABLE I

Spatial Outlier Detection - Quantitative Evaluation of Algorithms - Chosen Initial Centroids.

Methods	α **Index**	ρ **Index**	γ **Index**
$ROSE$	0.9836	0.0164	0.9987
$RFCM$	0.5448	0.4551	0.9250
$RPCM$	0.4725	0.5274	0.7919
$RFPCM$	0.5645	0.4354	0.9007

Legenda:
ROSE = Rough Outlier Set Extraction
RFCM = Rough Fuzzy C-Means
RPCM = Rough Possibilistic C-Means
RFPCM = Rough Fuzzy Possibilistic C-Means

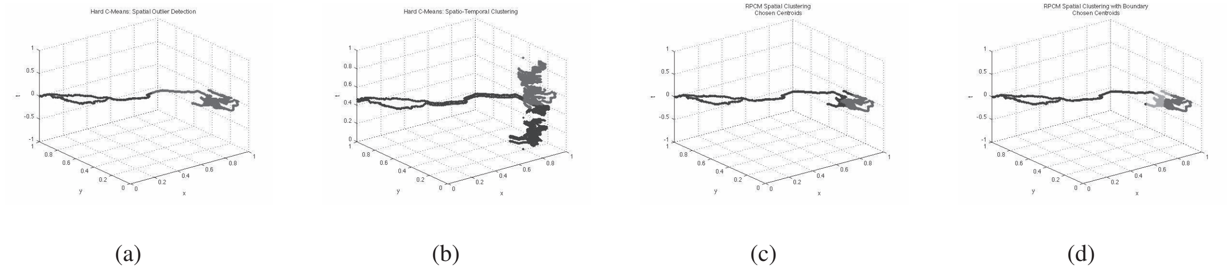
incorporating the concepts of rough sets. So, the algorithms adopted for comparison are: Hard C-Means, RFCM - Rough Fuzzy C-Means, RPCM - Rough Possibilistic C-Means, RFPCM - Rough Fuzzy Possibilistic C-Means. To analyze the performance of our proposed algorithm, tests have been performed on the School Buses dataset. Figures 8(a) and (b) show the clusters computed by Hard C-Means clustering algorithm (number of clusters set to 2) in spatial and spatio-temporal outlier detection respectively. Figures 8(c) and (d) - 9 show the results of each rough-fuzzy algorithm in spatial outlier detection. In figures 9(a) and 9(c), the two clusters are drawn with gray and black colors after the assignment of the boundary to clusters, while in the figures 9(b) and 9(d) the boundaries (before the assignment) are drawn with light gray color.

Figures 10-11 show the results of rough-fuzzy algorithms in spatio-temporal outlier detection. The parameters have been set as follows: $c = 2$ (Inlier and Outlier Cluster), ω and $\tilde{\omega}$ are equal to 0.5 in order to give the same importance to the lower approximation and to the boundary. Several runs have been done with different initializations and different parameters, related to initial centroid choice. These parameters have been maintained constant across all runs. The tests show that the best results are obtained for particular choices of initial centroids rather than

TABLE II

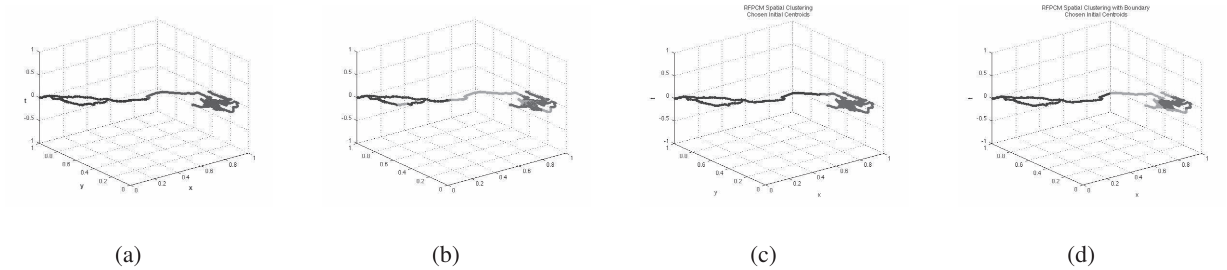
SPATIO-TEMPORAL OUTLIER DETECTION - QUANTITATIVE EVALUATION OF ALGORITHMS - CHOSEN INITIAL CENTROIDS.

Methods	α Index	ρ Index	γ Index
$ROSE$	0.8941	0.1059	0.9514
$RFCM$	0.3549	0.6450	0.6444
$RPCM$	0.3283	0.6716	0.5914
$RFPCM$	0.3651	0.6348	0.6618



(a) (b) (c) (d)

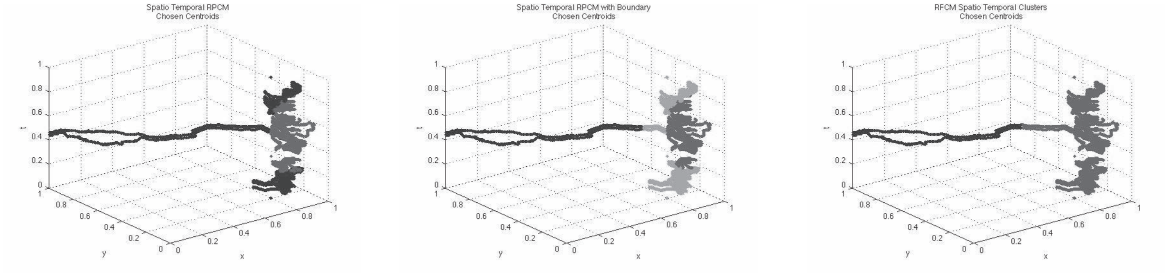
Fig. 8. Hard C-Means Clusters Results: (a) spatial outlier detection (b) spatio temporal outlier detection – Spatial Outlier Detection: (c) RPCM Clusters Results (d) RPCM Clusters Results with boundary.



(a) (b) (c) (d)

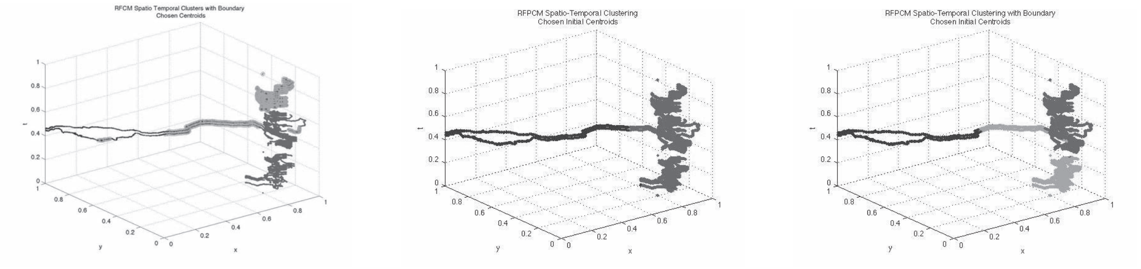
Fig. 9. Spatial Outlier Detection: (a) RFCM Clusters Results (b) RFCM Clusters Results with boundary (c) RFPCM Clusters Results (d) RFPCM Clusters Results with boundary.

for random choices of initial centroids. So, we report only the final prototypes of the best solution. Table I and Table II report the best results obtained using different algorithms for $c = 2$ in case of the same choice of initial centroids for HCM, RFCM, RPCM and RFPCM. Table I and Table II compare the performance of these different rough–fuzzy clustering algorithms with respect to α, ρ, γ in Spatial and Spatio-Temporal Outlier Detection respectively. The results reported in Tables I and II establish the fact that, although the hybridization versions of *c*–means algorithm were not designed as outlier detectors, they generate good prototypes for $c = 2$. In Spatial Outlier Detection, the RFPCM provides the best results as shown in Figure 9; the results of other two



(a) (b) (c)

Fig. 10. ST Outlier Detection: (a) RPCM Clusters Results (b) RPCM Clusters Results with boundary (c) RFCM Clusters Results.



(a) (b) (c)

Fig. 11. ST Outlier Detection: (a) RFCM Clusters Results with boundary (b) RFPCM Clusters Results (c) RFPCM Clusters Results with boundary.

versions of rough clustering are quite similar to that of the RFPCM, while in Spatio-Temporal Outlier Detection, the RPCM outperforms them as shown in Figure 10. The proposed ROSE algorithm performs better than HCM, RFCM, RPCM and RFPCM algorithms, both in terms of some qualitative measures and in terms of outliers detected, as shown in figures 6(a) and 6(b).

B. Wisconsin Breast Cancer Dataset

For the second test, the real-life dataset, named Wisconsin Breast Cancer [8] is used. The dataset is publicly available on UCI machine learning repository and consists of 699 instances with 9 continuous attributes. In order to compare our results, the experimental technique of Harkins et al. [23] by removing some malignant instances to form a very unbalanced distribution has been employed. The resultant data set had 483 instances (39 (8 %) malignant and 444 (92%) benign instances). The 9 continuous attributes are not transformed into categorical attributes.

TABLE III

ROSE RESULTS (LOWER / UPPER APPROX): COMPARISON ON WISCONSIN BREAST CANCER DATASET.

Top Ratio	**Number of rare classes included (Coverage)**						
	ROSE - Low	**ROSE - Upp**	**SEQ**	**DIS**	**NED**	**KNN**	**RNN**
1%(4)	4(10%)	6(15%)	3(8%)	4(10%)	4(10%)	3(8%)	4(10%)
2%(8)	8(20%)	11(28%)	7(18%)	5(13%)	5(13%)	6(15%)	8(21%)
4%(16)	16(41%)	22(56%)	14(36%)	11(28%)	11(28%)	11(28%)	16(41%)
6%(24)	23(59%)	28(72%)	21(54%)	18(46%)	18(46%)	18(46%)	20(51%)
8%(32)	28(72%)	35(90%)	28(72%)	24(62%)	24(62%)	25(64%)	27(69%)
10%(40)	33(85%)	37(95%)	32(82%)	29(74%)	29(74%)	30(77%)	32(82%)
12%(48)	37(95%)	38(97%)	35(90%)	36(92%)	36(92%)	35(90%)	37(95%)
14%(56)	38(97%)	39(100%)	39(100%)	39(100%)	38(97%)	36(92%)	39(100%)
16%(64)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)	36(92%)	39(100%)
18%(72)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)	38(97%)	39(100%)
20%(80)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)	38(97%)	39(100%)
28%(112)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)

1) Results and Comparison: In order to demonstrate the performance of our approach against traditional distance-based method (DIS), Neighborhood outlier detection algorithm (NED) [16], KNN algorithm [43], sequence-based outlier detection algorithm (SEQ) [28], RNN-based outlier detection method, all the other results about the *Coverage* (ratio of the number of rare classes Included to the number of objects in U belonging to that class) on this dataset can be found in the work of Harkins et al. [23] and Willams et al. [51]. Our results have been shown in Table III in the two related columns. For almost all considered *Top Ratio* values, ROSE performance, considering just the lower approximation, is higher than other methods and only sometimes equal to them. Indeed, the l.a. results go under SEQ, DIS and RNN only for *Top Ratio* equal to 14%. Instead, considering the upper approximation, i.e., the rough set contribution, the results are always higher or at least equal to all the other methods.

C. Grand St. Bernard WSN Dataset

Finally, our method has been also tested on a publicly available WSN data set named the Grand St. Bernard [26]. This dataset has been collected by a multi-hop wireless sensor network, deployed at the Grand St. Bernard pass, located between Switzerland and Italy running northeast-southwest through the Valais Alps. The deployment consists of 23 sensor nodes, measuring meteorological characteristics of the environment, during a period of two months (September-October 2007) with the sampling frequency of two minutes. The nodes are grouped in two clusters: a small cluster consists of 5 nodes and a big cluster consists of 18 nodes.

TABLE IV

ROSE RESULTS: COMPARISON ON GRAND ST. BERNARD DATASET - SPATIAL AND TEMPORAL OUTLIERS.

Methods	Running Average		Mahalanonis Dist.		Density	
	DR(%)	FPR(%)	DR(%)	FPR(%)	DR(%)	FPR(%)
TOD:	72.3	10.5	100	15.0	100	15.1
ROSE$_T$ Low:	80	1.2	96	1.0	100	0
ROSE$_T$ Upp:	87.5	1.7	100	1.2	100	0
SOD:	24.5	3.3	100	4.3	100	4.4
POD:	29.8	1.8	80	3.7	75	3.8
ROSE$_S$ Low:	96.2	1.0	92	1.0	92	0.3
ROSE$_S$ Upp:	98.1	1.2	96	1.1	100	1.1

1) Results and Comparison: This spatio temporal dataset, as most of spatio-temporal dataset, is not provided by a ground truth file. The methods TOD, SOD, POD, due to Zhang et al. [56], use this dataset labelled with three different methods, showing the different results on the basis of the three different techniques. The tests have been executed on the 30^{th} of September 2007 (06:00-14:00) and on the small cluster of 5 station (nodes: 25, 28, 29, 31, 32). The ambient temperature is the analyzed feature for each station. For temporal labelling, it was necessary to eliminate the dependency of the spatial domain, considering each sensor at a time. On the contrary for spatial labelling, all sensors (belonging to the cluster) have been considered at the same time. Table IV shows the ROSE results ($ROSE_S$ and $ROSE_T$ indicate the ROSE running for spatial/temporal outlier detection respectively) and the best tradeoff between DR% and FPR of the reported results for Zhang's TOD, SOD, POD. Concerning the temporal outlier detection, $ROSE_T$ Upp works always better than or comparable with TOD with a negligible percentage of false positives; even $ROSE_T$ Low works better than TOD on two of the three labelling techniques. Concerning the spatial outlier detection, $ROSE_S$ Upp and even $ROSE_S$ Low work always better than POD with a negligible percentage of false positives on all labelling techniques; $ROSE_S$ Upp and even $ROSE_S$ Low work highly better than SOD with a negligible percentage of false positives on running average technique and in a bit lower or comparable way than SOD on the other two labelling techniques. Globally, the achieved ROSE results outperform the compared state-of-the-art techniques on this spatio-temporal dataset.

D. Kernel Set Dimension

Experimental computations, about the dimension reduction between four analyzed datasets and their kernel sets, have been widely executed. The kernel sets dimensions, reported in the

TABLE V

KERNEL SET DIMENSION COMPUTATION ON DIFFERENT DATASETS.

Dataset Name	Universe cardinality	Kernel set cardinality	Reduction %
School Buses	30414	17101	13313 (44%)
Wisconsin Breast Cancer - Original	699	486	213 (30%)
Wisconsin Breast Cancer - Unbalanced	483	308	175 (36%)
Grand St. Bernard Dataset	2101	1535	566 (27%)

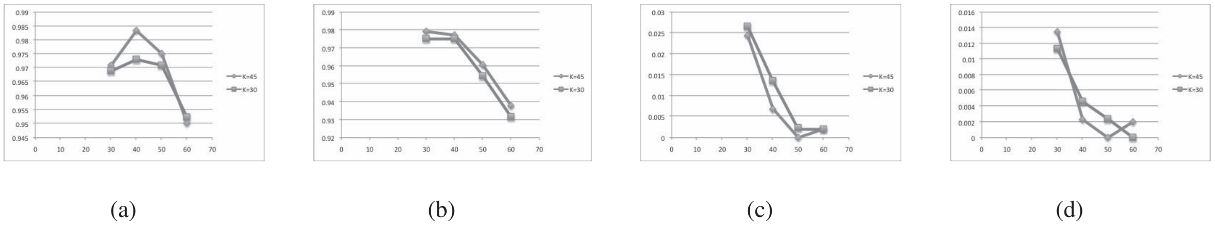
table V are the average dimensions on ten executions, varying the input parameters of the ROSE algorithm. The computed data provide an average value of reduction percentage equal to 46%. The analyzed datasets are the following: School Buses, Wisconsin Breast Cancer (original version), Wisconsin Breast Cancer (unbalanced version), Grand St. Bernard. This reduction significantly drops down the algorithm time complexity and hence its computational cost.

E. Sensitivity analysis of input parameters

This subsection ends this evaluation section and is intended to conduct a sensitivity analysis about the input parameters k and n of the algorithm in order to evaluate the algorithm behaviour. The comparison have been done on the Wisconsin Breast Cancer Dataset doing several different combinations of k and n parameters. In particular, the n and k parameters have been chosen in the following way: 1) keeping the value of n fixed (at 40, at 60) the value of k was varying at 1, 5, 9, 15, 25, 30, 45 and 2) keeping the value of k fixed (at 30, at 45) the value of n was varying at 30, 40, 50 and 60. The results have been shown in the following figures: the first couple of figures 12(a) and 12(b) shows the accuracy curves for $k = 45$ and $k = 30$ varying n; the second couple of figures 13(a) and 13(b) shows the accuracy curves for $n = 40$ and $n = 60$ varying k. Then, the figures 12(c) and 12(d) and the figures 13(c) and 13(d) show the false alarm probability curves for $k = 45$ and $k = 30$ varying n and those for $n = 40$ and $n = 60$ varying k, respectively. For $n = 50$ and $k = 45$ a reversal trend between lower and upper approximation as for $n = 40$ and a bit lower accuracy for $n = 50$ respect to $n = 40$ clearly appear. Hence, increasing too much the number n of outliers to be searched not surely improve the results. A zero false alarm probability has been reported in both cases for $k = 45$.

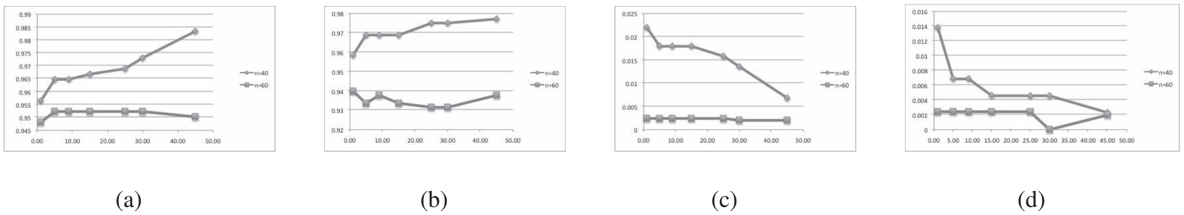
VII. CONCLUSIONS

The manuscript extends outlier detection using a new rough set approach to spatio-temporal data. Specifically, the rough set based outlier detection method has been theoretically grounded



(a) (b) (c) (d)

Fig. 12. Wisconsin Breast Cancer Dataset - for two fixed k values: (a) Accuracy: lower approximation (b) Accuracy: upper approximation (c) False Alarm Probability: lower approximation (d) False Alarm Probability: upper approximation.



(a) (b) (c) (d)

Fig. 13. Wisconsin Breast Cancer Dataset - for two fixed n values: (a) Accuracy: lower approximation (b) Accuracy: upper approximation (c) False Alarm Probability: lower approximation (d) False Alarm Probability: upper approximation.

based on a definition of outlier set as rough set. A remarkable note should be made for the definition of a new set, called kernel set, that has been demonstrated to be able to generate the ”same” output results in terms of rough outlier set with time computational benefits. The experimental results on three real world datasets prove that the performance of ROSE in detecting outliers are superior when compared to several other methods. On the real world School Buses dataset, ROSE has been compared with C-Means clustering algorithm and other *rough-fuzzy* clustering algorithms (Rough Fuzzy C-Means, Rough Possibilistic C-Means, Rough Fuzzy Possibilistic C-Means), incorporating the concepts of rough sets, producing reasonable results both in terms of quantitative and qualitative standpoints. On the benchmark Wisconsin Breast Cancer dataset, ROSE has been also compared with several state-of-the-art outlier detection methods, also rough-oriented, for general domain (SEQ, DIS, NED, KNN, RNN), demonstrating higher, and just sometimes comparable, performance. Another comparison has been made on the WSN Grand ST. Bernard dataset with spatio-temporal methods (Zhang’s TOD, SOD, POD) that use the same dataset, demonstrating the ROSE superiority even in this case. The approach is computationally less intensive compared with these approaches. The ROSE algorithm appear to consistently outperform other rough and not rough approaches in medium to large problem settings, showing to be able to do well also on datasets of varying sizes. Since spatio-temporal outlier detection might turn out to be useful in many different research fields, we hope that this work will spark further interest in such problems which are challenging and relatively unexplored.

References

[1] Aggarwal C.C. & Yu P. (2000). Finding Generalized Projected Clusters in High Dimensional Spaces. ACM SIGMOD Conference Proceedings. pp. 70–81.

[2] Aggarwal, C.C., & Yu, P. S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. The VLDB Journal, 14, pp. 211–221.

[3] Albanese A. & Petrosino A. (2011). A Non Parametric Approach to the Outlier Detection in Spatio-Temporal Data Analysis. Information Technology and Innovation Trends in Organizations, D'Atri, et al., Springer Verlag. pp. 101–108.

[4] Angiulli, F. & Pizzuti C. (2005). Outlier mining in large high-dimensional datasets. IEEE Transactions on Knowledge and Data Engineering, vol. 17 , no. 2, pp. 203–215.

[5] Angiulli, F., & Fassetti, F. (2010). Distance-based outlier queries in data streams: the novel task and algorithms. Data Mining and Knowledge Discovery, vol. 20, no. 2, pp. 290–324.

[6] Ankerst M., Breunig M. M., Kriegel H.-P. & Sander J. (1999). Optics: Ordering points to Identify the Clustering Structure, Proc. 1999 ACM SIGMOD Intl Conf. on Management of Data (SIGMOD 99), ACM Press, pp. 49–60.

[7] Barnett V. & Lewis T. (1994). Outliers in statistical data, New York John Wiley And Sons Ltd.

[8] Bay, S. D. (1999). The UCI KDD repository: http://kdd.ics.uci.edu.

[9] Birant D. & Kut A. (2006). Spatio-Temporal Outlier Detection in Large Databases. Journal of Computing and Information Technology, vol. 14, no. 4, pp. 291–297.

[10] Bittner T. (2000). Rough Sets in Spatio-temporal Data Mining. TSDM 2000, pp. 89–104.

[11] Boriah, S., Chandola, V. & Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. In SIAM International Conference on Data Mining, pp. 243–254.

[12] Breunig, M. M., Kriegel, H-P., Ng, R. T. & Sander, J.: LOF: identifying density based local outliers. In: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, Dallas, pp. 93–104.

[13] Ceglar, A., Roddick, J.F. and Powers, D.M.W. (2007). CURIO: A Fast Outlier and Outlier Cluster Detection Algorithm for Large Datasets. In Proc. 2nd Intern. Workshop on Integrating Artificial Intelligence and Data Mining, pp. 37–45.

[14] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys, vol. 41, no. 3, 15:1-15:58.

[15] Chen Y., Miao D. & Wang R. (2008). Outlier Detection Based on Granular Computing. LNCS, Springer-Verlag Berlin Heidelberg, pp. 283–292.

[16] Chen Y., Miao D. & Zhang H. (2010). Neighborhood outlier detection. Expert Systems with Applications, vol. 37, no. 12, pp. 8745–8749.

[17] Tao Cheng & Zhlin Li, (2006). A Multiscale Approach to Detect Spatio-Temporal Outliers. Transactions in GIS, vol. 10, no. 2, pp. 253–263.

[18] Das, K., & Schneider, J. (2007). Detecting anomalous records in categorical datasets. In the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, USA. pp. 220–229.

[19] Frentzos E., Gratsias K., Pelekis N. &Theodoridis Y. (2005). Nearest Neighbor Search on Moving Object Trajectories. Proc. 9th International Symposium on Spatial and Temporal Databases (SSTD'05), Angra dos Reis, Brazil, pp. 328–345.

[20] Ghosh A.K. & Chaudhuri P. (2005). On Maximum Depth Classifiers. Scandinavian J. Statistics, vol. 32, no. 2, pp. 327–350.

[21] Guha S., Rastogi R. & Shim K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. ACM SIGMOD Conference Proceedings. vol. 27, no. 2, pp. 73–84.

[22] Gutierrez, J. M. P., & Gregori, J. F. (2008). Clustering techniques applied to outlier detection of financial market series using a moving window filtering algorithm. Unpublished working paper series, No. 948, European Central Bank, Frankfurt, Germany.

[23] Harkins, S., He, H. X., Willams, G. J., Baxter, R. A. (2002). Outlier detection using replicator neural networks. In Proceedings of the 4th international conference on data warehousing and knowledge discovery, France, pp. 170–180.

[24] He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. Pattern Recognition Letters, vol. 24, pp. 1641–1650.

[25] Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, vol. 22, no. 2, 85–126.

[26] Ingelrest, F., Barrenetxea, G., Schaefer, G., Vetterli, M., Couach, O., and Parlange, M., 2010. SensorScope: application-specific sensor network for environmental monitoring. ACM Transactions on Sensor Networks, 6 (2), 1-32.

[27] Jiang F., Sui Y. & Cao C. (2006). Outlier Detection Based on Rough Membership Function. RSCTC '06, LNAI, Springer-Verlag Berlin Heidelberg, pp. 388–397.

[28] Jiang F., Sui Y. & Cao C. (2009). Some issues about outlier detection in rough set theory. in Expert Systems with Applications, vol. 36, no. 3, part 1, 4680–4687.

[29] Jornsten R. (2004). Clustering and Classification Based on the L1 Data Depth. Journal Multivariate Analysis, vol. 90, no. 1, pp. 67–89.

[30] Johnson T., Kwok I. & Ng R.T. (1998). Fast Computation of 2-Dimensional Depth Contours. KDD, pp. 224–228.

[31] Knorr, E. & Ng, R. (1998). Algorithms for Mining Distance-based Outliers in Large Datasets. In: Proc. of the 24th VLDB Conf., New York, pp. 392–403.

[32] Koufakou, A., & Georgiopoulos, M. (2010). A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery, vol. 20, no. 2, pp. 259–289.

[33] Laurikkala J., Juhola M. & Kentala E. (2000). Informal identification of outliers in medical data. In: Proceedings of IDAMAP, pp. 20–24.

[34] Liu W., Zheng Y., Chawla S., Yuan J. & Xie X., (2011). Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams. KDD11, August 21-24, San Diego, California, USA. pp. 1010–1018.

[35] Maji P. & Pal S.K. (2007). Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics. Vol. 37, n. 6, pp. 1529–1540.

[36] Markos M. & Sameer S. (2003). Novelty detection: A review part 1: statistical approaches. Signal Processing, vol. 83, no. 12, pp. 2481–2497.

[37] Muller, E., Assent, I., Steinhausen, U., & Seidl, T. (2008). OutRank: Ranking outliers in high dimensional data. In IEEE ICDE 2008 Workshops: The 3rd International Workshop on Self-managing Database Systems (SMDB), Cancun, Mexico. pp. 600–603.

[38] Ng R.T., Han J. (2002). CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering, vol. 14 , no. 5, pp. 1003–1016.

[39] Nguyen T. T. (2007). Outlier Detection: An Approximate Reasoning Approach. in Proc. of RSEISP '07, LNCS, Springer-Verlag Berlin Heidelberg, pp. 495–504.

[40] Papadimitriou S., Kitagawa H., Gibbons P.B., Faloutsos C. (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral. 19th International Conference on Data Engineering (ICDE'03). pp. 315–326.

[41] Pawlak Z., Rough Sets, Theoretical Aspects of Reasoning about data. Dordrecht, The Netherlands: Kluwer, 1991.

[42] Pawlak Z. & Skowron A. (1993). A rough set approach for decision rules generation. Proc. Workshop W12: The Management of Uncertainty in AI at 13th IJCAI.

[43] Ramaswamy S., Rastogi R. & Shim K. (2000). Efficient algorithms for mining outliers from large datasets. In Proceedings of the 2000 ACM SIGMOD Int. Conf. on Management of Data, Dallas, Texas, pp. 427–438.

[44] Ranga Suri N.N.R., Murty N., & Athithan G. (2010). Data Mining Techniques for Outlier Detection, Chapter 2. In Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications. pp. 22–38.

[45] Sander J., Ester M., Kriegel H.-P. & Xu X. (1998). Density-based Clustering in Spatial Databases: the algorithm GDBSCAN and its applications, Data Mining and Knowledge Discovery, vol. 2, pp.169–194.

[46] Sun P. & Chawla S. (2004). On Local Spatial Outliers. Proc. 4th IEEE Intl Conf. on Data Mining, pp. 209216.

[47] Tao, Y., Xiao, X., & Zhou, S. (2006). Mining distance-based outliers from large databases in any metric space. In the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia: ACM Press. pp. 394–403.

[48] Valero-Mora P.M, Young F.W. & Friendly M. (2003). Visualizing categorical data in ViSta. Computational Statistics & Data Analysis, vol. 43, pp. 495–508.

[49] Venkateswara Rao K., Govardhan A. & Chalapati Rao K.V. (2012). Spatio Temporal Data Mining: issues, Task and Applications. International Journal of Computer Science & Engineering Survey (IJCSES), vol. 3, no. 1, pp. 39–52.

[50] Wang X.R., Lizier J.T., Obst O., Prokopenko M., & Wang P. (2008). Spatiotemporal anomaly detection in gasmonitoring sensor networks. in Proceedings of the European conference on Wireless Sensor Networks (EWSN), 2008, pp. 90–105.

[51] Willams, G. J., Baxter, R. A., He, H. X., Harkins, S., Gu, L. F. (2002). A comparative study of RNN for outlier detection in data mining. In ICDM, Japan, pp. 709–712.

[52] Wu E., Liu W. & Chawla S. (2008). Spatio-Temporal Outlier Detection in Precipitation Data. In Proceedings of the second international Workshop on Knowledge Discovery From Sensor Data. SensorKDD'08, LNCS 5840. pp. 115–133.

[53] Yao Y.Y. (1996). Two views of the theory of rough sets in finite universes. International Journal of Approximate Reasoning, vol. 15, pp. 291–317.

[54] Zhang T., Ramakrishnan R., Livny M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases, ACM SIGMOD Conference Proceedings. vol. 25, no. 2, pp. 103–114.

[55] Zhang Y., Yang S., Wang Y. (2008). LDBOD: A novel local distribution based outlier detector. Pattern Recognition Letters 29(7), 967–976.

[56] Zhang Y., Hamm N.A.S., Meratnia N., Stein A., van de Voort M., Havinga P.J.M. (2012). Statistics-based outlier detection for wireless sensor networks, International Journal of Geographical Information Science, vol. 26, no. 8, pp. 1373–1392.

[57] Zhu, C., Kitagawa, H., & Faloutsos, C. (2005). Example-based robust outlier detection in high dimensional datasets. In the IEEE International Conference on Data Mining (ICDM05), pp. 829–832.

AUTHOR - BIO

Alessia Albanese received the Master Degree in Mathematics from the University of Naples Federico II and PhD Degree in Computer Science from the University of Milan. Currently, she is a senior research fellow at University of Naples Parthenope - Applied Science Dept. Her research interests are: data mining and knowledge discovery, pattern recognition, soft computing, spatio temporal data analysis. She is a student member of the IEEE Computer Society.

Sankar K. Pal is a Distinguished Scientist of the Indian Statistical Institute and its former Director. He is also a J.C. Bose Fellow of the Govt. of India. He founded the Machine Intelligence Unit and the Center for Soft Computing Research at the Institute in Calcutta which are enjoying international recognition. A PhD from Calcutta University and Imperial College, London, joined his Institute in 1975 as a CSIR Senior Research Fellow where he became a Full Professor in 1987, Distinguished Scientist in 1998 and the Director in 2005. Dr. Pal worked at the UC Berkeley and UMD, College Park; the NASA JSC, Houston, Texas; and US Naval Research Lab, Washington DC. He has been serving as a Distinguished Visitor of IEEE Computer Society since 1987 and held several visiting positions in Italy, Poland, Hong Kong and Australian universities. Prof. Pal is a Fellow of the IEEE, TWAS, IAPR, IFSA, and all the four National Academies for Science/Engineering in India. He is a co-author of seventeen books and more than three hundred research publications in the areas of Pattern Recognition and Machine Learning, Image Processing, Data Mining, Web Intelligence, Soft Computing, and Bioinformatics. He is/was in the editor boards of twenty journals including IEEE Trans.. He has received several national and international awards including the most coveted S.S. Bhatnagar Prize in India in 1990.

Alfredo Petrosino is professor of Computer Science at the University of Naples Parthenope, where he heads the research laboratory CVPRLab at University of Naples Parthenope (cvprlab.uniparthenope.it). He held positions at University of Salerno, International Institute of Advanced Scientific Studies (IIASS), National Institute for the Physics of Matter (INFM), and lastly as Researcher and Senior Researcher at National Research Council (CNR). He taught at the Universities of Salerno, Siena, Naples Federico II, Naples Parthenope. He is Senior member of the IEEE, USA, member of the International Association for Pattern Recognition, USA, International Neural Networks Society, USA. He co-edited six books and more than one hundred research publications in the areas of Computer Vision, Image and Video Analysis, Pattern Recognition, Neural Networks, Fuzzy and Rough Sets, Data Mining. He is/was an Associate Editor of Pattern Recognition journal; Member of the Editorial Board of Pattern Recognition Letters, International Journal of Knowledge Engineering and Soft Data Paradigms; Editor of IEEE Trans SMC-Part A, Fuzzy Sets and Systems, Image and Vision Computing, Parallel Computing.