Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Rough Set Based Outlier Detection Technique for Spatiotemporal Data

[1]V. Vikas, [2]Dr. A. Mary Sowjanya
[1]Student of M.Tech, [2]Assistant Professor
[1, 2] Computer Science and Systems Engineering, Andhra University College of Engineering
Andhra Pradesh, India

Abstract— High availability of data gathered from wireless sensor networks and telecommunication systems has become focus of researchers for extracting knowledge from spatiotemporal data. This is because, detecting outliers which are grossly different from or inconsistent with the remaining spatiotemporal data set is a major challenge in real-world knowledge discovery and data mining applications. We have studied and adopted Rough Outlier Set Extraction (ROSE) approach to design and implement a totally new approach called Rough Set based Outlier Detection (RSOD). ROSE is the first rough method of outlier detection technique using soft granular computing-based solution. Our ROSD method also relies on a rough set theoretic representation of the outlier set using the rough set approximations, i.e., lower and upper approximations. The experimental results on the two real-world datasets (numerical and spatiotemporal) prove that the performance of ROSE and RSOD in detecting outliers are optimal. In this paper, a rough set based outlier detection approach has been worked out on numerical data first later modified to suit to spatiotemporal data. The problem of outlier detection by exploiting the rough set approach is performed by proposing a spatiotemporal weight measure such that the outlier detection capabilities of the RSOD can be enhanced so that more reliable outlier information is obtained from the spatiotemporal data taken.

Keywords— Data mining, Outlier detection, spatiotemporal data, rough set and lower outlier set approximation and upper outlier set approximation

I. INTRODUCTION

The steady and amazing progress of computer hardware technology lead to the wide availability of huge amounts of data beyond petabytes as a result of data collected from automated data collection tools, database systems, Web, computerized society. Data mining takes up the task of turning such data into useful information and knowledge.

Spatiotemporal data is a data that contains both space and time information. Due to the high availability of data gathered from wireless sensor networks and telecommunication systems. Spatio-temporal objects can be characterized by three aspects: (1) Characterizing by what they are, e.g., to the class which they belong to; (2) Characterizing them by where they are, i.e., the spatial location of them; (3) Characterizing them by when they actually existed and when they have been, are, or will be* where, i.e., their temporal location. It has drawn the attention of researchers on the problem of extracting knowledge from spatiotemporal data. In particular outlier detection is used in many fields like changes in system behaviour, fraudulent behaviour identification, intrusion detection in computer networking, and in image processing it helps in identification of abnormal regions.

Disharmony is induced due to the presence of outliers and makes the modelling complicated. Source of outliers are Errors, that refers to a noise-related measurement coming from a faulty sensor and Events, an event is defined as a particular phenomenon that changes the real-world state, e.g., forest fire, air pollution, etc. This kind of outliers usually lasts for a relatively long period of time and changes historical pattern of sensor data. Outliers caused by errors may occur frequently.

The difficulties in Outlier Detection are as follows:

1. Estimation of every possible normal behaviour in the region.
2. Imprecise boundary between normal and outlier behaviour since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.
3. Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from malicious actions.
4. In many domains normal behaviour keeps evolving and may not be current to be a representative in the future.

Output of outlier detection is an important aspect for any outlier detection technique is the manner in which the outliers are reported. Typically, the outputs produced by outlier detection techniques can be categorised into following two types:

1. Scores: Scoring techniques give an outlier score to each instance present in the test data, which depends on the degree of outlierness of that instance. This produces outliers in a ranked list format as an output. Either a cut-off threshold or analysis of top few outliers can be chosen depending on the requirement.
2. Labels: Labels (normal or anomalous) are assigned to each instance when employing techniques under this.

The proposed method Rough Set based Outlier Detection (RSOD) upgrades the Rough Outlier Set Extraction (ROSE), which is a first rough method that improves and upgrades the "scoring methods," by exploiting the uncertainty region (boundary) to obtain more reliable results, by overcoming the dependence of ROSE over the prior assumptions of number of outliers and nearest neighbours to be considered for the k-Nearest Neighbour [1] based spatiotemporal weight calculation.

Rough-set theory (RST) [2] is a paradigm to deal with uncertainty, vagueness, and incompleteness and it is proposed for indiscernibility in classification according to some similarity. Proposed method exploits the set-oriented point of view of rough set theory to define the concept of outlier in terms of its lower and upper approximations (rough outlier set).

This paper is organized as follows: In Section 2, explains the rough set theory. Section 3 depicts related work regarding outlier detection approaches is given. Section 4 depicts RSOD approach to extract the spatiotemporal rough outlier set. Section 5 presents the results and discussions on two real world data sets. Finally conclusion remarks are given in Section6 about on-going and future work.

II. ROUGH SET THEORY

A rough set is first described by Zdzisław I. Pawlak, is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the lower and the upper approximation of the original set. While representing incomplete knowledge it can be approached as an extension to Classical Set Theory.

Let *I= (U, A)* be an information system (attribute-value system), where *U* is a non-empty set of finite objects (the universe) and *A* is a non-empty, finite set of attributes such that for every *a: U→Va* is the set of values that attribute "*a*" may take. The information table assigns a value *a(x)* from *Va* to each attribute "*a*" and object in the universe *U.*

With any $P \subseteq X$ there is an associated equivalence relation I_P:

$I_P = \{(x,y) \in U^2 \mid \forall a \in P, a(x)=a(y)\}$

The relation I_p is called a P-indiscernibility relation. If $(x,y) \in I_P$, then and are indiscernible (or indistinguishable) by attributes from P. The equivalence classes of the *P*-indiscernibility relation are denoted $[x]_{P.}$

Definition of a rough set

Let X ⊆ U be a target set that we wish to represent using attribute subset P; that is, we are told that an arbitrary set of objects X comprises a single class, and we wish to express this class (i.e., this subset) using the equivalence classes induced by attribute subset P. In general, X cannot be expressed exactly, because the set may include and exclude objects which are indistinguishable on the basis of attributes P.

However, the target set X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of X:

$$\underline{P}(X) = \{x \mid [x]_P \subseteq X\}$$

$$\bar{P}(X) = \{x \mid [x]_P \cap X \neq \emptyset\}$$

A. *Lower approximation and positive region:*

The P-lower approximation, or positive region, is the union of all equivalence classes in $[x]_P$ which are contained by (i.e., are subsets of) the target set. The lower approximation is the complete set of objects in that can be positively (i.e., unambiguously) classified as belonging to target set.

B. *Upper approximation and negative region:*

The P-upper approximation is the union of all equivalence classes in $[x]_P$ which have non-empty intersection with the target set. The upper approximation is the complete set of objects that in U/P that cannot be positively (i.e., unambiguously) classified as belonging to the complement ($\bar{X}$) of the target set X. In other words, the upper approximation is the complete set of objects that are possibly members of the target set X. The set $U – \bar{P}X$ therefore represents the negative region, containing the set of objects that can be definitely rule d out as members of the target set

III. RELATED WORK

Most of the existing works on outlier detection include:

1. Statistical Distribution-Based approach assumes a distribution or probability model for the given data set and then identifies outliers, with respect to the model, that deviates from the rest.
2. Distance-based approach is based on distance based metrics that compute the fraction of objects that are a specified distance from a particular object;
3. Depth-based techniques that classifies an outliers as an object at the outer layers when the data objects are organized in convex hull layers;
4. Density based approach classifies an outlier if it is outlying relative to its local neighbourhood, particularly with respect to the density of the neighbourhood; and
5. Deviation based outlier detection method identifies outliers by examining the main characteristics of objects in a group. Objects that deviate from this description are considered outliers.

Several surveys on the outlier detection emphasise on a specific research area or on an application field like [3], while surveys like[4],[5],[6] give contemporary and complete outlier detection techniques. They give detailed information about every technique under each category. They provide insight into the aspects like the difficulties discussed earlier to be taken into account while selecting an appropriate outlier detection methodology.

They provide information of outlier detection methodologies in a general domain. But very less number of outlier detection techniques has been proposed in the area of spatiotemporal domain. Alessia Albanese et al. [7] propose a spatiotemporal outlier detection approach called "Rough Outlier Set Extraction (ROSE)", which exploits set oriented approach of rough set theory to estimate the outliers. Another spatiotemporal data based outlier detection approach is proposed by Wang et al [8]. Rough set theory has been recently introduced into spatiotemporal domain literature for various aspects. Approximations of spatiotemporal regions and relation among the approximations are proposed by Bittner [9].

IV. RSOD APPROACH

In our approach a strict distinction between spatial and temporal components is proposed. Thus suits the datasets that are characterized by either spatial or temporal data which happens when the temporal information is implicit or not given at all in case of spatial datasets and vice versa. This helps in considering only temporal or spatial information thus saving time and space.

A. Problem Statement

Given an information system S=<U, A> with U a spatiotemporal normalized data set and A is its set of attributes. Where the universal set U is

U= {$x_i \equiv \left(a_{i1}, a_{i2,} a_{i3} \dots\dots\dots . a_{in}\right) \in [0,1]^n, i = 1,2, \dots\dots . N$}

Where x_i=1,2……..N and A={a_1,a_2,...,a_m} is the attribute set with at least three attributes i.e. spatial attributes and temporal one. The Outlier Detection problem consists of finding objects a spatiotemporal weight measure like k-nearest neighbours is used to determine the degree of dissimilarity of each object with respect to all others an outlier set is extracted.

Spatial and temporal outlier sets can be defined as set of all objects whose spatial attribute value is significantly different from those of its spatial neighbourhood objects and object whose temporal attribute value is significantly different from those of its temporal neighbourhood objects respectively. A Spatiotemporal outlier is an object that satisfies both the above conditions as spatial and temporal outliers. By using a spatiotemporal weight measure and a threshold value the lower and upper approximations are calculated and then rough outlier set is extracted for computing of any one of spatial outlier set, temporal outlier set, spatiotemporal outlier set depending upon requirement and attribute set selected.

With this goal of finding Outlier set with lower and upper approximations are computed, let S=<U, A> and $P \subseteq A$ there is an associated equivalence relation I_P:

I_P = {(x, y) ∈ U^2 |∀a ∈ P, a(x)=a(y)}

The relation I_p is called a P-indiscernibility relation. If (x, y) ∈ I_P, then and are indiscernible (or indistinguishable) by attributes from P. The equivalence classes of the *P*-indiscernibility relation are denoted $[x]_{P.}$ or $x_{k.}$

To obtain degree of outlierness a correct and appropriate measure should be chosen to associate with every object i.e. the Euclidean distance calculated between each selected object and rest of the objects in U. In this RSOD approach we propose spatiotemporal weight measure that would overcome the disadvantages of ROSE that depends on k-nearest neighbours based measure which requires prior confirmation on number of nearest neighbours to be considered and also ensures that certain number of outliers to be extracted among the given objects.

In Spatiotemporal context, the measure associated with every object is based on the difference between Mean of all the objects and minimum distance from its spatial neighbourhood and temporal neighbourhood.

Spatiotemporal weight measure = $a \cdot M_p^s(U) + b \cdot M_p^t(U)$ – (1)

Where $M_p^s(U)$ and $M_p^t(U)$ are the spatial and temporal weights respectively, which are computed as follow:

$$M_p^s(U) = \mu - Mindist_s(U)$$

$$M_p^t(U) = \mu - Mindist_t(U)$$

Where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \ \forall x_i \in$ X where X is the spatial or temporal attribute of U

Mindist_s(U) = Minimum distance from it to its nearest object.

The constants *a & b* weight such that *a + b=1*. If only spatial data or temporal data is used we could eliminate the other weight by simply choosing its corresponding constant as zero. Thus facilitating the outlier detection when spatial or temporal datasets alone are used. Lower outlier set approximation of the required outlier set is calculated using the above measure associated to each object and a threshold, such that calculated measure should be greater than threshold. Upper outlier set approximation of the required outlier set is calculated using the above measure associated to each object and previous iteration's threshold, such that calculated measure should be greater than the previous threshold. Where threshold is calculated using minimum of the weights in the upper outlier approximation set and before calculating the new threshold it is assigned to the previous threshold.

B. Existing Approach

The existing approach is Rough Outlier Set Extraction (ROSE) is a rough set oriented spatiotemporal outlier detection methodology, which depends on k-nearest neighbours based neighbourhood detection in outlier analysis by using distance to the k-nearest neighbours as a distance measure in the spatiotemporal weight calculation associated with each object in the data. The number of outliers to be detected is also fixed to detect top n outliers, which also makes it vulnerable in the absence of outliers making it to exhibit the top n objects with highest spatiotemporal weights. The proposed approach overcomes both these defects by not assuming the number of outliers to be detected in prior and also choosing of the value of ‘k’ the number of nearest neighbours to be considered for spatiotemporal weight calculation.

C. Methodology

The below figure depicts the flowchart of the RSOD approach in which at the beginning a dataset is supplied over which the algorithm would be applied.In the second step certain number of elements are extracted from the dataset for each iteration which is fixed percentage of the cardinality of the dataset. Later the spatiotemporal weight measure is calculated over the every set of extracted elements, after that comparing the weights with the thresholds updating of negative region and Upper outlier set (Uos) and Lower outlier set (Los) is done until the initial dataset has no more elements to be extracted. Finally the upper outlier set and the lower outlier set are displayed .

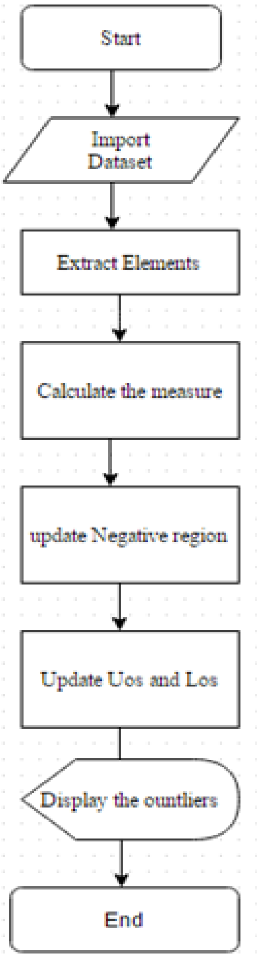


Fig. 1 Flowchart of RSOD

D. Steps

Step 1: To deal with the outliers in the spatiotemporal domain the system receives a universe of discourse U, The output would be the Rough Outlier Set (Upper, Lower Approximation).

Step 2: At each step, a chunk of objects called *Currentobjects*, from the overall data set U by *ExtractElements*is extracted, which is a fixed number.

Step 3: We calculate mean of all the objects and all the current objects extracted, we compute the Euclidean distances among the objects in the extracted once and to all the objects of U.

Step 4: UpdateUpperApprox and UpdateLowerApprox at first iteration give the same set of outliers at that step, i.e., the n objects that have their measure higher than the others. Then at next iterations, UpdateUpperApprox and UpdateLowerApprox calculate the Lower and Upper approximation of Rough Outlier Set, using the Threshold (computed by Lower Weight) and previous thresholds.

Step 5: At each iteration the Lower Weight function computes the threshold. At each iteration, the thresholds have been computed as the weight minimum value among the weight maximum n values. Thus we obtain the output the Rough Outlier Set(Upper, Lower Approximation, and Negative Region).

V. RESULTS AND DISCUSSIONS

This rough set based outlier detection approach has been worked out on numerical data first then modified to suit to spatiotemporal data. We have applied this approach on two datasets; one is on Wisconsin Breast Cancer Dataset[12] with a general domain available on UCI machine learning repository and the other on real world "trucks" dataset which is spatiotemporal in nature. Wisconsin Breast Cancer Dataset with a general domain available on UCI machine learning repository. This comprises of 699 instances with nine continuous attributes, precisely on texture attribute of the dataset. In this a few records are modified in order to represent the presence of outliers. The below figure depicts the scatter plot of the result produced by the ROSE method on the Wisconsin Breast Cancer Dataset.

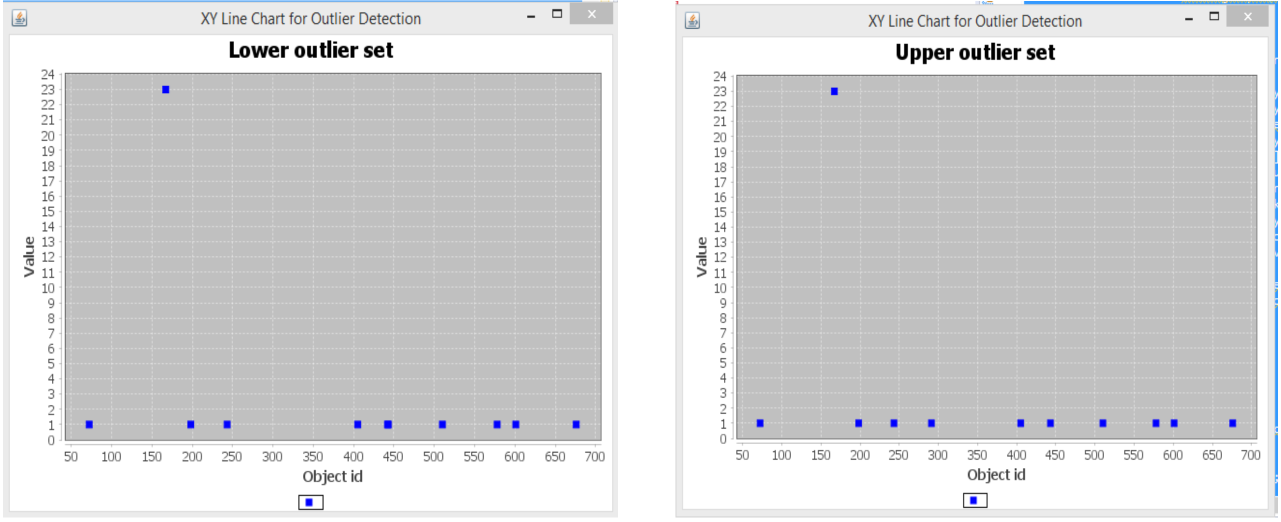
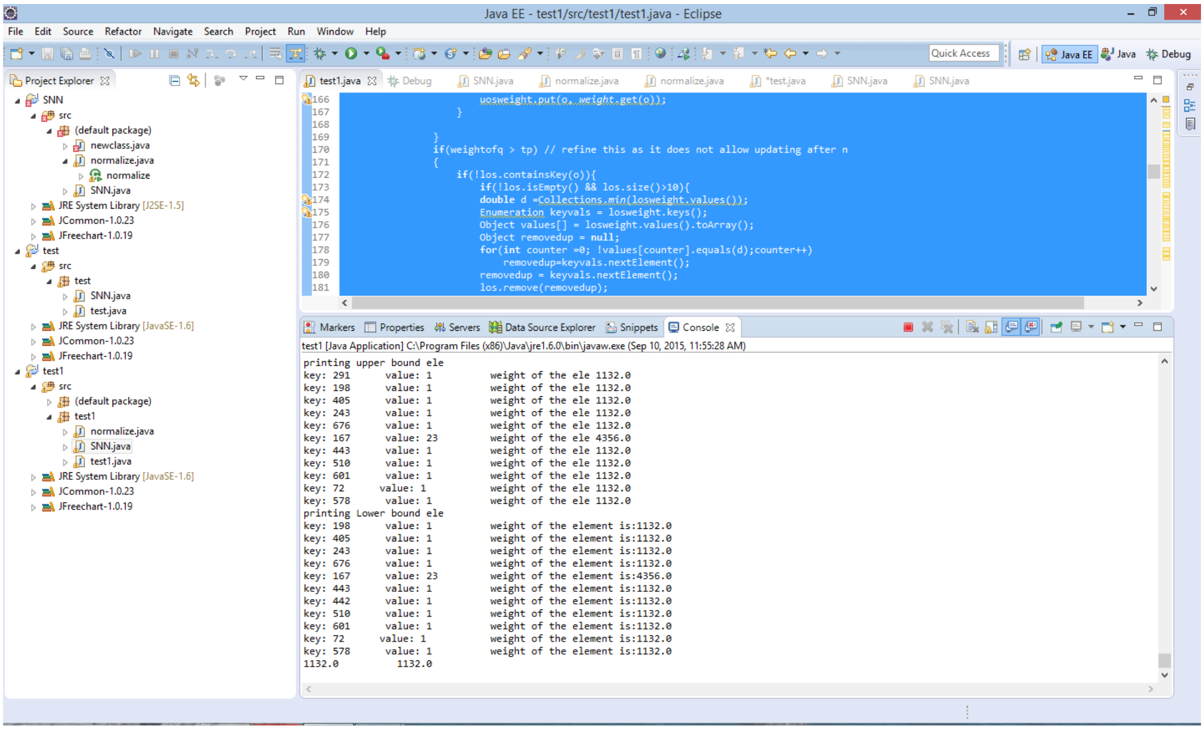


Fig. 2. Lower and Upper outlier sets for Wisconsin Breast Cancer Dataset on texture field

The bellow figure depicts the results produced by ROSE on the Wisconsin breast cancer dataset on texture field which is earlier displayed using the scatter plot method.



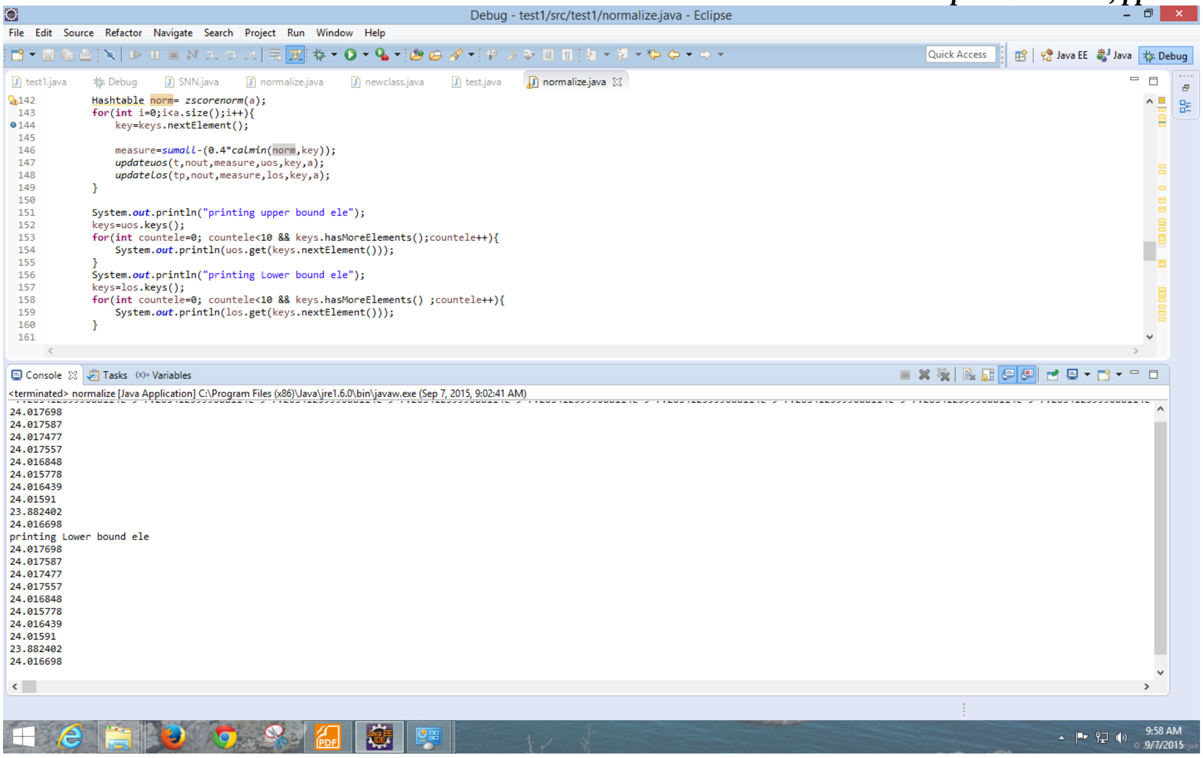
```
                uosweight.put(o, weight.get(o));
            }

        }
        if(weightofq > tp) // refine this as it does not allow updating after n
        {
            if(!los.containsKey(o)){
                if(!los.isEmpty() && los.size()>10){
                double d =Collections.min(losweight.values());
                Enumeration keyvals = losweight.keys();
                Object values[] = losweight.values().toArray();
                Object removedup = null;
                for(int counter =0; !values[counter].equals(d);counter++)
                    removedup=keyvals.nextElement();
                removedup = keyvals.nextElement();
                los.remove(removedup);
```

test1 [Java Application] C:\Program Files (x86)\Java\jre1.6.0\bin\javaw.exe (Sep 10, 2015, 11:55:28 AM)

```
printing upper bound ele
key: 291      value: 1         weight of the ele 1132.0
key: 198      value: 1         weight of the ele 1132.0
key: 405      value: 1         weight of the ele 1132.0
key: 243      value: 1         weight of the ele 1132.0
key: 676      value: 1         weight of the ele 1132.0
key: 167      value: 23        weight of the ele 4356.0
key: 443      value: 1         weight of the ele 1132.0
key: 510      value: 1         weight of the ele 1132.0
key: 601      value: 1         weight of the ele 1132.0
key: 72      value: 1          weight of the ele 1132.0
key: 578      value: 1         weight of the ele 1132.0
printing Lower bound ele
key: 198      value: 1         weight of the element is:1132.0
key: 405      value: 1         weight of the element is:1132.0
key: 243      value: 1         weight of the element is:1132.0
key: 676      value: 1         weight of the element is:1132.0
key: 167      value: 23        weight of the element is:4356.0
key: 443      value: 1         weight of the element is:1132.0
key: 442      value: 1         weight of the element is:1132.0
key: 510      value: 1         weight of the element is:1132.0
key: 601      value: 1         weight of the element is:1132.0
key: 72      value: 1          weight of the element is:1132.0
key: 578      value: 1         weight of the element is:1132.0
1132.0         1132.0
```

Fig. 3. Result for Wisconsin Breast Cancer Dataset on texture field by ROSE in Eclipse

For the next test, we have used a real world dataset named Trucks [13] dataset. The dataset is publicly available and consists of 112203 instances of 50 trucks. The structure of each record in the dataset is as follows: {obj id; traj id;date; time; lat; lon; x; y} where obj id is the truck identification, traj id is the unique trajectory identification, the date and time are the sampling time stamps (date in dd=mm=yyyy format and time in hh:mm:ss format), the (lat; lon) and (x; y) are the truck location, in WGS84 and in GGRS87 reference systems, respectively. In our case, the obj id and traj id are not considered, we chose latitude attribute to be tested. In the following figure we have taken nearly one third of the above dataset instances (i.e. 30,000 instances). The below figure depicts the Lower and Upper outlier set approximations of the outlier set detected from the dataset by RSOD approach.



```
        Hashtable norm= zscorenorm(a);
        for(int i=0;i<a.size();i++){
            key=keys.nextElement();

            measure=sumall-(0.4*calmin(norm,key));
            updateuos(t,nout,measure,uos,key,a);
            updatelos(tp,nout,measure,los,key,a);
        }

        System.out.println("printing upper bound ele");
        keys=uos.keys();
        for(int countele=0; countele<10 && keys.hasMoreElements();countele++){
            System.out.println(uos.get(keys.nextElement()));
        }
        System.out.println("printing Lower bound ele");
        keys=los.keys();
        for(int countele=0; countele<10 && keys.hasMoreElements() ;countele++){
            System.out.println(los.get(keys.nextElement()));
        }
```

```
<terminated> normalize [Java Application] C:\Program Files (x86)\Java\jre1.6.0\bin\javaw.exe (Sep 7, 2015, 9:02:41 AM)
24.017698
24.017587
24.017477
24.017557
24.016848
24.015778
24.016439
24.01591
23.882402
24.016698
printing Lower bound ele
24.017698
24.017587
24.017477
24.017557
24.016848
24.015778
24.016439
24.01591
23.882402
24.016698
```

Fig 4 Result for trucks dataset on Latitude field by RSOD

VI. CONCLUSIONS

In this paper, a rough set based outlier detection approach has been worked out on numerical data first then modified to suit to spatiotemporal data. Specifically, the rough set based outlier detection method has been theoretically grounded based on a definition of outlier set as rough set. The implemented approach overcomes the disadvantages of the ROSE such as dependence of ROSE over user to input number of outliers to be detected and the number of nearest neighbours to be considered for spatiotemporal weight thus making it more efficient. Experimental results on real-world data sets namely trucks dataset from chorochronos data stories and cancer dataset from UCI repository demonstrate the superiority of ROSE outliers detected, over those obtained by Inter quartile range oriented outlier detection. In future, this algorithm can be applied on some other suitable spatiotemporal series datasets like earth quake, hurricanes, road traffic jam, road accidents and weather etc. Using this approach in combination with clustering algorithms could be analyzed for more efficient outlier detection and better computational benefits.

.

REFERENCES

[1] S. Ramaswamy, R. Rastogi, and K. Shim, ―Efficient Algorithms for Mining Outliers from Large Data Sets,‖ Proc. ACM SIGMOD Int‗l Conf. Management Data, pp. 427-438, 2000.

[2] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, 1991

[3] NilamUpasani and Dr.Hari Om, "Outlier Detection: A Survey on Techniques Involving Fuzzy and/or Neural Approaches," IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions, July 2013.

[4] Victoria J. Hodge and Jim Austina, "Survey of Outlier Detection Methodologies" Kluwer Academic Publishers,Artificial Intelligence Review 22: 85–126, 2004

[5] Irad Ben-Gal, "Outlier Detection" Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

[6] K. VenkateswaraRao, A. Govardhan, and K.V. ChalapatiRao,"Spatio Temporal Data Mining: Issues, Task and Applications,"Int'l J. Computer Science Eng. Survey, vol. 3, no. 1, pp. 39-52, 2012

[7] Alessia Albanese, Sankar K. Pal and Alfredo Petrosino, ―Rough Sets, Kernel Set, and Spatiotemporal Outlier Detection‖ vol. 26, no. 1, January 2014.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, ―Efficient Algorithms for Mining Outliers from Large Data Sets,‖ Proc. ACM SIGMOD Int‗l Conf. Management Data, pp. 427-438, 2000.

[9] X.R. Wang, J.T. Lizier, O. Obst, M. Prokopenko, and P. Wang, "Spatiotemporal Anomaly Detection in Gasmonitoring Sensor Networks," Proc. European Conf. Wireless Sensor Networks (EWSN), pp. 90-105, 2008.

[10] T. Bittner, "Rough Sets in Spatio-Temporal Data Mining," Proc. First Int'l Workshop Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers (TSDM '00), pp. 89-104, 2000.

[11] Data Mining – Concepts and Techniques by Han and Kamber, Second Edition, Morgan Kaufmann publishers, 2006

[12] S.D.Bay, "The UCI KDD Repository," http://kdd.ics.uci.edu,1999.

[13] Nikos Pelekis, "collection of moving object databases", http://chorochronos.datastories.org, 2011