Rough Sets in Spatiotemporal, Multimedia Outlier Detection

C . Leela Krishna
Dept. of CSE,
Sree Vidyanikethan Engg. College,
Tirupati, A.P.

Mr . Shaik Salam
Associate Professor, Dept. of CSE,
Sree Vidyanikethan Engg. College,
Tirupati, A.P.

Dr . T . V . Rao
Head, Dept. of CSE,
PVP Siddhartha Inst. of Tech.,
Vijayawada, A.P.

Abstract—**The major issue that has drawn the attention of researchers on extracting knowledge from spatiotemporal data and multimedia data is the high availability of data gathered from multiple sources. Spatiotemporal data can be gathered from wireless sensor networks and other environmental monitoring systems. Multimedia data can be gathered from various audio, video files. The data collected from various data sources contain outliers, which are grossly different or exceptional when compared with others. Detection of outliers in the process of knowledge discovery is found to be challenging. In this paper, we are dealing the outlier detection problem to find the top outliers from both spatiotemporal and multimedia data by exploiting rough outlier set extraction (ROSE) method, which is based on the rough set theory in its lower and upper approximations, even with uncertainties in data sets.**

Index Terms—**Spatiotemporal data, Multimedia data, Outliers, Rough sets, Uncertainty.**

I.   INRTODUCTION

Data mining can be referred to a process of   discovering knowledge from data where the main emphasis is on uncovering interesting data patterns that are hidden in large data sets. Spatiotemporal and multimedia data mining are growing research areas involving the discovery of hidden knowledge in large spatiotemporal and multimedia databases, mainly by detecting periodic and frequent patterns. A database may contain data objects that do not comply with the general behavior or model of the data [4]. These objects are known as outliers. The major applications of outlier detection include credit card fraud detection, intrusion detection in computer networks, and detection of abnormal regions or detecting motion in images. The presence of outliers makes modeling and mining difficult due to the discordance of outliers that are introduced into the data [1]. The most important existing approaches for outlier detection are:

1.   Distribution-based approaches [7] that make use of statistical distributions are used to model the data and
     outliers deviate from the   model.
2.   Depth-based approaches [8] that evaluate different layers of convex hulls are computed and the objects belonging to the outer layers are declared as outliers.
3.   Distance-based approaches [9] that compute distances of every object from a target object.
4.   Density-based approaches [10] that assign a weight to each sample based on the neighbourhood density.

The proposed method involves detection of outliers from both spatiotemporal and multimedia data through rough sets known as Rough Outlier Set Extraction (ROSE) [1], where rough set theory is used to define outliers in terms of its lower and upper approximations. This paper is organized as follows: Section 2 describes spatiotemporal data mining and Section 3 describes multimedia data mining. Section 4 gives the major concepts of rough set theory that are related to the work. Section 5 introduces the problem definitions. Section 6 gives the ROSE approach to extract spatiotemporal and multimedia outliers. Finally Section 7 concludes the ongoing work.

II.   SPATIOTEMPORAL DATA MINING

A spatial data set is one that has its attributes containing land information or geographical locations on the earth. A temporal data set is one that has its attributes containing several events that are ordered by one or more dimensions of time. Data sets with attributes containing both location information constrained by time information are referred to as spatiotemporal data sets. Data relating to space and time together, having spatial extension along with temporal duration is also referred to as spatiotemporal data. Similarly, a spatial database that stores spatial objects that change with time is called a spatiotemporal database, from which interesting information can be mined [4]. Extraction of implicit knowledge, spatial and temporal relationships stored in spatiotemporal databases is also known as spatiotemporal data mining. For example, to group the trends of moving objects and identify some strangely moving vehicles, or distinguish a bioterrorist attack from a normal outbreak or the flu based on the geographic spread of a disease with time [4]. Spatiotemporal uncertainty is the lack of, or the error in, knowledge about either an object's position on the earth with respect to the latitudes and longitudes or an object's motion at various time intervals.

III.   MULTIMEDIA DATA MINING

Multimedia is defined as a combination of more than one media [2]. Generally, the two types of media are static and

dynamic media [5]. Static media include text, graphics, and images. Dynamic media include animation, audio, video, etc. A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text mark-ups, and linkages [4]. Extraction of implicit knowledge stored in multimedia databases is also known as multimedia data mining. It is applicable in picture content-based retrieval, voice-mail systems, video-on-demand systems, and speech-based user interfaces to recognize spoken commands [4].

IV. ROUGH SET THEORY

Rough set theory [3] was first proposed by Z. Pawlak, which is used to deal with uncertain, vague and incomplete data. The main idea is based on the indiscernibility relation that describes the indistinguishability of objects. If one cannot represent an object in the form of a crisp set, then such object can be represented by a rough set with the approximate representation of knowledge. A rough set gives the lower and upper approximation of the set that has incompletely specified knowledge.

A. Information System

An information system is a table consisting of rows and columns, where each row represents an object and each column represents an attribute. Hence, the information system is shown as IS = (U, A) where U is the universe of discourse with a finite set of objects, $U = \{O_1, O_2, O_3 ... O_n\}$ and A is a finite and nonempty set of attributes $A = \{a_1, a_2, a_3... a_n\}$.

B. Indiscernibility Relation

The relation between two or more objects where all or some of the values of its attributes taken into consideration are identical is known as an indiscernibility relation. This relation gives the indistinguishability of objects through an equivalence relationship. A relation among various objects is said to be equivalence iff they are reflexive, symmetric and transitive. For B as a subset of A, $B \subseteq A$, the indiscernibility relation IS_B on U is given as

$$I_B = \{(x, y) \in U \times U : \forall a \in B, a(x) = a(y)\}$$
(1)

C. Lower and Upper Approximations

Approximations are fundamental to rough set theory, as every set is represented by means of its lower and upper approximations. The equivalence classes induced by the B-indiscernibility relation is denoted by $[x]_B$. For any set X, that is a subset of the universe U, $X \subseteq U$,
the lower approximation is defined as follows:

$$\underline{B}(X) = \{x \in U : [x]_B \subseteq X\},$$
(2)

the upper approximation is defined as follows:

$$\overline{B}(X) = \{x \in U : [x]_B \cap X \neq \emptyset \},$$
(3)

The lower approximation or positive region is the union of all equivalence classes [12] in $[x]_B$ which are already contained by the target set, where the objects excluding the target set cannot be included here. The upper approximation or the negative region is the union of all equivalence classes in $[x]_B$ that have non-empty intersection [12] with the target set, where the objects excluding the target set can also be included here. The objects in the lower approximation can be certainly classified as members of X on the basis of knowledge in B, and the objects in the upper approximation can only be classified as possible members of X on the basis of B. The boundary region consists of objects that can neither be ruled in nor ruled out [12] as members of the target set X.

$$BND(X) = \overline{B}(X) - \underline{B}(X)$$
(4)

D. Example

The following table shows an information system containing a collection of multimedia data as a set of objects. The universe of discourse is $U = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7\}$. The set of attributes is represented as A = {Illustration, Timeline, Movement}. The possible values for both Illustration and Timeline are Yes, No and for Movement the possible values are Static and Dynamic [2].

TABLE I

Information System

Object	Condition Attribute		
	Illustration	*Timeline*	*Movement*
O_1	Yes	Yes	Dynamic
O_2	Yes	No	Static
O_3	Yes	No	Dynamic
O_4	No	Yes	Static
O_5	Yes	Yes	Static
O_6	No	Yes	Dynamic
O_7	No	Yes	Dynamic

From the above table, it can be observed that for the *Illustration* attribute, the set regarding the indiscernibility relation is $\{O_1, O_2, O_3, O_5\}$, since the objects in the set are indistinguishable to each other as they possess same value of the mentioned attribute. Similarly, for the *Timeline* and *Movement* attributes, the sets showing indiscernibility relations are $\{O_1, O_4, O_5, O_6, O_7\}$ and $\{O_1, O_3, O_6, O_7\}$ for 'Dynamic' and $\{O_2, O_4, O_5\}$ for 'Static'.

For the subset B = {*Illustration, Timeline*}, the set with indiscernibility relation is {{O₁, O₅}, {O₂, O₃}, {O₄, O₆, O₇}}, since the objects in the set have same attribute values based on the knowledge in B. If the target set is X = {O₁, O₂, O₃}, then the corresponding lower and upper approximations are $\underline{B}(X)$ = {O₂, O₃} and $\overline{B}(X)$ = {O₁, O₂, O₃, O₅} and the boundary region is BND(X) = {O₁, O₅}, which is evaluated as the set difference between upper and lower approximations. This region specifies the data objects that can neither be ruled in nor ruled out [12].

V. OUTLIER DETECTION PROBLEM

Data objects that are grossly different from others or inconsistent with the remaining objects in the data set are known as outliers [4]. Outliers can be caused by measurement or execution error [4]. The process of extraction of outliers from a data set is known as outlier mining. Outliers present in spatiotemporal data and multimedia data are to be detected by means of rough set theory. The proposed approach finds both spatiotemporal and multimedia outliers.

A. Problem Definitions

Consider an information system IS = (U, A) with U being the universe of discourse with a data set containing normalized values (between 0 and 1) for all of its objects, and A containing the set of attributes. U can be written as follows:

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

(5)

where p_i represents a set of data objects with i=1,2,...,N is a m-dimensional vector and A = {a₁, a₂, ..., aₘ} is the attribute set. When the universe of discourse is given, for an integer n>0 the number of outliers to be evaluated, the degree of dissimilarity for each object with others for every $p_i \in U$, given as $D_{p_i}(U)$ is to be evaluated.

The outlier detection problem consists of finding $\bar{n} \geq n$ objects $p_1, p_2, \dots p_n, p_{n+1}, \dots p_{\bar{n}}$ such that

$$D_{p_1}(U) \geq D_{p_2}(U) \geq \cdots \geq D_{p_n}(U) D_{p_{n+1}}(U) \dots = D_{p_{\bar{n}}}(U) > D_{p_j}(U), \forall j = \bar{n} + 1 .. N$$

(6)

An n-outlier set consists of $O \subseteq U$ with the set of objects

$$O = \{p_1, p_2, \dots p_n, p_{n+1}, \dots p_{\bar{n}} U / D_{p_1}(U) \geq D_{p_2}(U) \geq \cdots \geq D_{p_n}(U) D_{p_{n+1}}(U) \dots D_{p_{\bar{n}}}(U) > D_{p_j}(U), \forall j = \bar{n} + 1 \dots N\}$$

(7)

From the above formulae, the outlierness threshold is considered as $\tau = D_{p_n}(U)$ which is defined as the minimum value among the n maximum measures computed in U, that belongs to the n-outlier set.

The threshold value for the n-outlier set is

$$\tau = \min\{max_1(D_p(U), D_q(U)), max_2(D_p(U), D_q(U), \dots, max_n(D_p(U), D_q(U))\}, \forall p, q \in U.$$

(8)

The following definitions are applicable in the outlier detection problem from both spatiotemporal and multimedia along with their respective notations given below:

Definition 1. A spatial outlier is an object whose spatial attribute value is different from its closer objects in its neighbourhood.

Definition 2. An object that belongs to the spatial n-outlier set is referred to as a Spatial Outlier (Oₛ).

Definition 3. A Temporal outlier is an object whose temporal attribute value is different from its closer objects in its neighbourhood.

Definition 4. An object that belongs to the temporal n-outlier set is referred to as a Temporal Outlier (Oₜ).

Definition 5. A Spatiotemporal Outlier is an object that contains both spatial and temporal outliers (Oₛ,ₜ).

Definition 6. A Multimedia Outlier is an object whose multimedia attribute value is significantly different from its closer objects in the neighbourhood.

Definition 7. A Multimedia Outlier is an object that belongs to the multimedia n-outlier set (Oₘ).

An object p ∈ U is said to be an outlier and belongs to n-outlier set iff $d^t_p(U) \geq \tau$. To obtain the degree of dissimilarity, an appropriate measure used is the euclidean distance computed between each object and all the others belonging to U. Euclidean distance between two data objects O₁(x₁,x₂,..,xₙ) and O₂(y₁,y₂,...yₙ) in an n-dimensional space is calculated by the following formula:

$$D(O_1, O_2) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2}$$

(9)

The real world applications with huge amount of data is found to be unfeasible due to its high computational complexity $O(N^2)$ with N = |U| [1], the cardinality of the data set which gives the number of objects in the data set, as it is obvious from (9). To reduce the complexity, two strategies can be approached by making use of:

1. k-nearest neighbors [11] rather than the entire N objects where k<<N.
2. a pruning strategy to remove all objects that cannot be in the n-outlier set.

B. Spatiotemporal Outlier Detection

In a spatiotemporal context, the measure associated with each object is based on the distances from its spatial k-nearest neighbors and its temporal k-nearest neighbors which can be given as:

$$D^{s,t}_p(U) = \alpha D^s_p(U) + \beta D^t_p(U)$$

(10)

where

$$D^s{}_p(U) = \sum_{j=1}^{k} D^s(p, NN^s(p, p_j))$$
(11)

$$D^t{}_p(U) = \sum_{j=1}^{k} D^t(p, NN^t(p, p_j))$$
(12)

where $k > 0$ is the number of nearest neighbors and $NN^s(p, p_j)$ specifies the jth nearest spatial neighbor and $NN^t(p, p_j)$ specifies the jth nearest temporal neighbor. α and β are chosen in such that $\alpha + \beta = 1$. Definition 3 defines the spatiotemporal outlier detection problem by evaluating a measure in (10).

C. Multimedia Outlier Detection

In a multimedia context, the measure associated with each

object is based on the distances from its multimedia k-nearest neighbors which can be given as:

$$D^m{}_p(U) = \sum_{j=1}^{k} D^m(p, NN^m(p, p_j))$$
(13)

where $k > 0$ is the number of nearest neighbors and $NN^m(p, p_j)$ specifies the jth nearest multimedia neighbor. Definition 6 defines the multimedia outlier detection problem by evaluating a measure in (13).

VI. ROUGH OUTLIER SET EXTRACTION

A. Theory

Exploitation of the rough set theory in defining the Outlier Set is referred to as Rough Outlier Set. For an information system $IS = (U, A)$ with U being a normalized data set consisting of data objects and A being the attribute set. If $n > 0$ are the required outliers, then an n-Outlier Set $O \subseteq U$ is described as $(\underline{B}(O), \overline{B}(O))$ with the B-Lower and B-Upper approximations respectively. The B-Lower approximation $\underline{B}(O)$ defines the set of objects that can be certainly classified as members of the set O, on the basis of knowledge in B. The B-Upper approximation $\overline{B}(O)$ defines the set of possible members of O, on the basis of knowledge in B. The formulae for evaluating both lower and upper approximations are already mentioned in (2) and (3).

To better illustrate the definitions given in Section 5, an example spatiotemporal data set is taken from [6]. It contains $U = \{ p_i \equiv (z_{i1}, z_{i2}, z_{i3}) \in [0,1]^3, i = 1, \ldots, 18 \}$ where p_i is a normalized three dimensional data set with $A = \{a_1, a2, a3\}$, with a_1, a_2 are the spatial attributes and a_3 being the temporal attribute. U contains 18 data objects as plotted in Fig. 1. When the values of k and n are fixed as 3 and 4, the outlier sets are computed based on euclidean distance and the resultant outlier sets are computed as follows:

A *4-Spatial Outlier Set O_s $\subseteq$ U* gives the set of objects that deviate from the rest with respect to spatial component only and the result is $O_s = \{p_7, p_8, p_{17}, p_{18}\}$.

A *4-Temporal Outlier Set O_t $\subseteq$ U* gives the set of objects that deviate from the rest with respect to temporal component only and the result is $O_t = \{ p_1, p_2, p_{17}, p_{18} \}$.

A *4-Spatiotemporal Outlier Set $O_{s,t}$ $\subseteq$ U* gives the set of objects that deviate from the rest with respect to both spatial and temporal components and the result is $O_{s,t} = \{ p_7, p_8, p_{17}, p_{18} \}$.
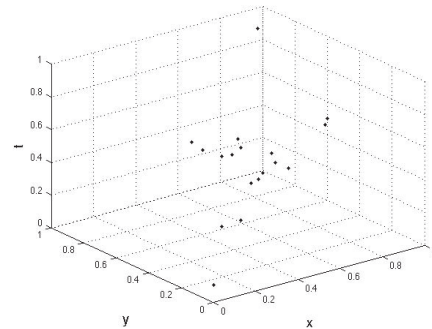

Fig. 1. The example data set U

Spatial Outliers can be evaluated by making reduction based on temporal attribute, then the B-indiscernibility relation of U with $B = \{z_{i3}\}$ will contain the following indiscernibility relation [6]:

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

From the *B*-indiscernibility relation all the equivalent objects have been obtained and the *B-lower approximation* of the *Spatial Outlier Set O_s* is given as $\underline{B}(O_s) = \{\{p_7, p_8\}, \{p_{17}\},$ $\{p_{18}\}\}$ and the *B-upper approximation* of the *Spatial Outlier Set O_s* is given as $\overline{B}(O_s) = \{\{p_7, p_8\}, \{p_{17}\}, \{p_{18}\}\}$.

Temporal Outliers can be evaluated by making reduction based on spatial attribute, then the B-indiscernibility relation of U with $B = \{z_{i1}, z_{i2}\}$ will contain the following indiscernibility relation [6]:

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

From the *B*-indiscernibility relation all the equivalent objects have been obtained and the *B-lower approximation* of the *Temporal Outlier Set O_t* is given as $\underline{B}(O_t) = \{\{p_{17}\}, \{p_{18}\}\}$ and the *B-upper approximation* of the *Temporal Outlier Set O_t* is given as $\overline{B}(O_t) = \{\{p_1, p_{12}\}, \{p_2, p_{13}\}, \{p_{17}\}, \{p_{18}\}\}$.

The following figure shows the spatial, temporal and spatiotemporal outliers detected from the example data set U [6]:
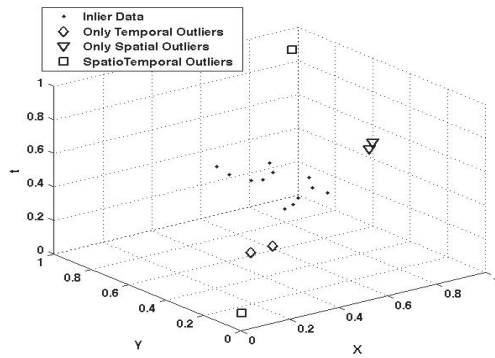
Fig. 2. Detected outlier sets

B. Algorithm

```
begin ROSE(U, n, k)
UpperSet = null; LowerSet = null;
weight(q)=0;
τ=0; τ_prev=0;
WorkingSet = ExtractElements(U);
while (WorkingSet ! = null) do
  for p∈U do
    for q∈WorkingSet do
      if (LowerSet == null and UpperSet == null) or
         (weight(q)>= τ_prev)then
              ds(p,q) = CalculateSpDistance(p,q)
              dt(p,q) =CalculateTempDistance(p,q)
              EvaluateKNN(p,q,ds,dt,k)
      else
              AddNegativeRegion(p)
      end if
    end for
  end for
  for q∈WorkingSet do
    weight(q) = CalculateWeight(q)
    UpperSet = UpdUpperApp(τ_prev,n,weight(q))
    LowerSet = UpdLowerApp(τ,n,weight(q))
  end for
  τ= LowerWeight(UpperSet)
  if(τ!=0) then
    τ_prev= τ;
  end if
  U = U-WorkingSet
  WorkingSet = ExtractElements(U)
end while
return Rough Outlier Set
end ROSE()
```

The Rough Outlier Set Extraction Algorithm receives 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 with its Upper, Lower Approximation, and Negative Region. The algorithm selects a small subset of objects, called WorkingSet, from the overall data set U. For all selected objects, the algorithm computes 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 depending upon the chosen attribute subset B with respect to the Rough Outlier Set. UpdUpperApp and UpdLowerApp at first iteration create the same set of n top outliers at that step, i.e., the n objects that have a measure higher than the others. Then, at next iterations, UpdUpperApp and UpdLowerApp compute the Lower and Upper approximation of Rough Outlier Set, using the τ (computed by LowerWeight) threshold. At each iteration i, the pruning strategy selects objects from U with their measure under the computed threshold to build the Negative Region. The LowerWeight function computes the τ threshold, which is computed as the weight minimum value among the weight maximum n values.

VII. CONCLUSIONS

This paper provides a rough set approach as a solution to the outlier detection problem in both spatiotemporal and multimedia data. The detected outlier set through rough sets is named as rough outlier set. The computational benefit of using euclidean distance on only some of the nearest neighbours instead of all data objects is also considered. We hope this work will attain further interest in various problems.

REFERENCES

[1] A. Albanese, S. K. Pal, A. Petrosino, "Rough Sets, Kernel Set, and Spatiotemporal Outlier Detection," IEEE Trans. Knowledge and Data Eng., vol. 26, no. 1, pp. 194-207, Jan. 2014.
[2] M. A. Rahman, M. Lazim, F. Mohamed, "Applying Rough Set Theory in Multimedia Data Classification," Int'l Journal on New Computer Architectures and Their Applications, pp 633-693, 2011.
[3] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data.* Kluwer, 1991.
[4] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Elsevier.
[5] J. Griffioen, B. Seales, R. Yavatkar, "Content Based Multimedia Data Management and Efficient Remote Access," Extrait de la Revue Informatique et Statistique dens les Sciences Humanies. Vol 1(4), pp. 213-233 (1997).
[6] The Computer Society Digital Library, http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.234
[7] V. Barnett and T. Lewis, *Outliers in Statistical Data.* John Wiley & Sons, 1994.
[8] M. Breunig, H-P. Kriegel, R.T. Ng, and J. Sander,"LOF: Identifying Density Based Local Outliers," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 93-104, 2000.
[9] A. K. Ghosh and P. Chaudhuri, "On Maximum Depth Classifiers," *Scandinavian J. Statistics*, vol. 32, no. 2, pp. 73-84, 1998.
[10] E. Knorr and R. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24th Int'l Conf. Knowledge Discovery and Data Mining, pp. 224-228, 1998.
[11] 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.
[12] Rough set, http://en.wikipedia.org/wiki/Rough_set