RESEARCH ARTICLE OPEN ACCESS

Spatio-Temporal Outlier Analysis and Detection using K-medoids with SVM

M. Naveena Priya M.Sc., (M.Phil) [1],
Mrs. P.Anitha, M.C.A., M. Phil.,(B.Ed.,) [2]
Department of Computer Science [1], Assistant Professor [2]
Department of Computer Applications [2]
Vellalar College for Women (Autonomous) Erode
Tamil Nadu - India

ABSTRACT

Spatio – temporal methods is the process of innovations and finding the patterns from the knowledge representations through outliers. This kind of data representing the (i) the states of an object (ii) position or event in space at a particular period of time. It refers to the Objects whose attribute values are entirely different from its neighborhood. Always their locations are different even the nodes from the entire population are unique. Outlier Detection is the most important techniques in data mining, which is useful for identifying several activities from the huge data set. This Project is deals with the identification of Breast cancer. Here we are comparing the accuracy and performance with the previous technology, as expected Our proposed algorithm using k-medoids – support vector machine is more accurate then the Rough Outlier set Extraction mode.

Keywords:- K-medoids Support vector machine, Rough set Extraction, Spatio-temporal Outliers.

I. INTRODUCTION

The main objectives are to be mentioned in many different ways. First, supports an intelligent rule management agent for checking and enforcing spatio-temporal constraints. Moreover, a separate agent called spatio-temporal information agent has been proposed and implemented to manage the spatio-temporal constraints in order to provide effective access control for web databases. A generalized K-medoids model provides the access control by considering the status level of the user and spatio-temporal constraints to provide better access control by restricting unauthorized users. In addition, this proposed system provides a WPBC named as Wisconsin Pattern Breast Cancer Dataset to analyze breast cancer symptoms that uses rough set based decision tree algorithm and an outlier detection algorithm for effective classification. In order to reduce the false alarm rate the system proposed work supports concept of outlier detection along with classification techniques. To achieve this goal, a K-Support vector Machine Based rough set based outlier detection algorithm have been proposed and implemented.

The breast cancer data objects are frequently change their paths over time in the spatio temporal data mining, which will describes the existing research in Rough Outlier Set Extraction named as ROSE. Our proposed methods by using K-medoids for outlier detection exploit rough theory to define new rough weights as degree of outliers. While comparing the accuracy and performance, Our proposed algorithm, k-medoids is more accurate then the Rough Outlier set Extraction mode. The Proof for the same is to be demonstrated in this research paper. And also, we can able to use all the data set by means of k-medoids without any wastage of memory space.

II. PROPOSED RESEARCH METHODOLOGY

The proposed research methodology is system provides a WPBC (Wisconsin Pattern Breast Cancer) Dataset to analyze breast cancer symptoms that uses K- medoids with Support Vector Machine rough set based outlier detection algorithm for effective classification and it has been implemented. The tool used in this proposed system is Java.

Spatio temporal outlier detection

Our method WPBC (Wisconsin Pattern Breast Cancer) data set is too loaded in order to get the outliers. First of all needs to collect the errors, missing data to be found and removed. Then the second process is the Data Preparation. Based on the data preparation, the patterns related to time factor has to be resolved.

In the Rough set Outlier Extraction, there are two methods adopted as lower and upper approximations. The Original data set is having in the kernel which in turn describes the experimental results. Our proposed algorithm, Support vector machine in K-medoids is used to detect more outliers than the ROSE, which is the major contribution by the classification technique using as a advantage of this model.

2.1. Temporal Outlier Detection using Patterns

In order to identify temporal outliers, are utilized to obtain temporal patterns and differences of the real-life data set, named Wisconsin Breast Cancer is used. The data set is publicly available on UCI (University of California, Irvine) machine learning repository and consists of 699 instances with nine continuous attributes. To compare the results, the experimental technique of removing some malignant instances to form a very unbalanced distribution has been employed. The resultant data set had 8 percent and malignant 92 percent benign minimum possibility instances. The nine continuous attributes are not transformed into categorical attributes.

2.2. Pattern Formation

The Formation values cannot address the outlier detection in this analysis and patterns are thus compared to identify outliers. Behavior of the attribute in a specific time interval is called a pattern. The patterns are obtained by conjoining the values of cancer attributes in consecutive time slices. In order to, form patterns Outlier measurements and WPBC Dataset products should be in a similar time scales. While the Outlier data were observed with two minutes frequency, MF's (matched filter) are hourly instantaneous values and MO's (matched outlier) are hourly mean values. WPBC Dataset forecasts are derived from instantaneous values at the full hour HH:00 and forecast model time step is 60s only and thus the real time resolution is about five minutes (WPBC Dataset, 2013). Subsequently, every hour of instantaneous Outlier measurements are aggregated with observations of five minutes before and after and will be monitored.

2.3. Proposed Algorithm

The *K-medoids* algorithm is based on the a clustering algorithm in which it is related to the *K-means* algorithm and the Medoids shift algorithm.

It can be run in multiple iterations where the SVM (Support Vector Machine) learner initialization is performed by using the clustering. In the first iteration, it runs standard K-medoids algorithm to yield a clustering based on the primary space *X*. This iteration has two purposes. First, it uses the clustering result from this step as a baseline for comparison. It generates the initialization set defined by the label for the SVM experiments of K-medoids. In the beginning of an iteration it is *t* + 1.

To look at each cluster *πt i* generated in the previous run and select *m* objects closest to the centroid of *πt i* and use their associated *ui* for SVM initialization. The one-against rest classification is used in the SVMs,. the function *xi*'s that are closest to the centroid are more unique to be correctly assigned to their correct clusters whereas the incorrect assignments tend to appear towards the boundaries of the clusters. One thing to note about K-medoids is that the objective function in equation is guaranteed to converge to a local maxima since K-medoids s, as in the case of K-means, it will reassign or reallocate an object from an old cluster *πi* to a new cluster *πj one and* only if the object is more similar to *πj*'s centroid than *πi*'s centroids on spatio temporal outliers.

Definition:

x_i : Objects to be clustered

d_{ij} : distance of object x_i to cluster π_j

$m(i)$: assigned cluster of x_i

$l(\pi_i)$: SVM learner of cluster π_i

$\hat{y}(u,\pi) = \sum_{z=1}^{n} \alpha_z^{\pi} K^{\pi}(u, u_z^{\pi}) + b^{\pi}$ SVM decision value u for cluster π

Λ: Penalty term

```
For each $x_i \in X$
    d= $x_i \cdot c_{m(i)}$ , $\pi_i \leftarrow m(i)$
    s= $\sum_{\forall uk, Ttk = 1} \hat{y}(u_k, \pi_i)$
     For each cluster $\pi_j$, i≠j
       $\hat{d} = x_i \cdot c_j$
       $\hat{s} = \sum_{\forall uk, Ttk = 1} \hat{y}(u_k, \pi_j)$
        If  $(\hat{d} < d \; and \; s < 0 \; and \; \hat{s} > 0)$  or $\hat{d} \cdot \lambda < d \; and \; \hat{s} < 0)$
            Remove $u's$ related with $x_i$ from $l(\pi_i)$
            Insert $u's$ related with $x_i$ to $l(\pi_j)$ as +1
            Insert $u's$ related with $x_i$ to $l(\pi_p)$ as -1,
j≠p
            $m(x_i) \leftarrow \pi_j$
        End
    End
End
```

2.4. Outlier Identification Using K-Medoids:

Our proposed algorithm comprises two stages:

Concept-independent pre-processing:- The step one is concept-independent, thus allowing for learning different concepts at a later stage.

Concept-specific pre-processing:-

The second step is concerned with the selection of appropriate subset of the instances to be defined. Following the selection of the subset, the relevant learning algorithm (K-medoids) is applied to obtain the classifier.

Assume the availability of a large collection of training instances for pre-processing. This does not prevent the addition or removal of instances from the dataset. The same dataset is used for learning multiple concepts defined by varying labels. A classifier must be learned experimentally to classify unseen instances for each and every clusters.

Thus, the use SVMs as the classifier algorithm, multiple hyper planes would need to be learned, one for each class. Essentially, the second step represents the reuse of the training dataset with different labels. Such a scenario is common in data repositories where the data is available for pre-processing with periodic updates adding/removing instances.

III. SYSTEM ARCHITECTURE

3.1. Architecture

Outlier detection is termed as data preprocessing job as well as data post processing task, where abnormal facilities and features are identified for analysis of useful tasks.

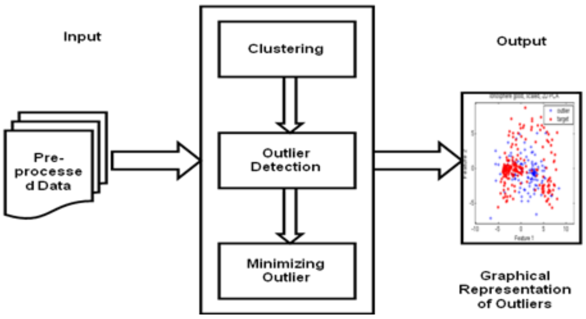


Fig 3.1 System architecture for outlier detection

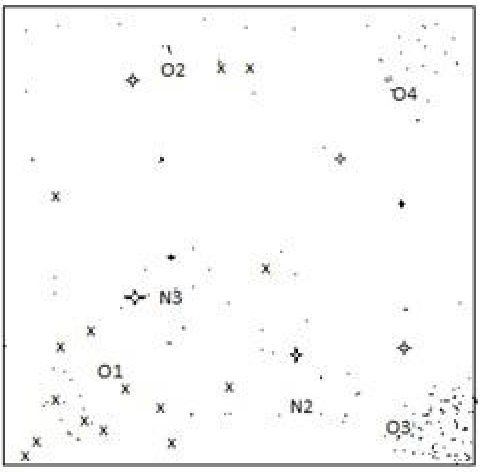


Fig 3.2 Outliers in 2-dimentaional data

3.2. System Design Of Outlier Detection In Breast Cancer Dataset

The proposed system design is based on the following four phases.

(i) Concept-independent preprocessing
(ii) Concept-specific sampling
(iii) Similarity assessment
(iv) Temporal outlier detection

3.2.1. Concept-Independent Preprocessing

The first preprocessing step is performed only once for the entire dataset. This concept-independent stage determines the set of SVM (Support vector Machine) of each instance. At the end of this stage, it have a data structure containing the indexes and distances of the K-medoids with SVM of every instance in the dataset.

Insertion of a new instance into the dataset involves computing the buckets to which it hashes. Deletion of instances is noted by setting flag.

3.2.2. Concept-Specific Sampling

The second step is the concept-dependent stage where, given the class labels of the instances, a subset of the instances is to be randomly selected by the user and it is to be used as input for the learning methodology. Here it computes the distance between the data points to define outliers

3.2.3. Similarity Assessment

Similarity assessment was performed by calculating the cancer radius. For the similarity assessment data points of Outlier measurements and cancers are compared. The data points for Outlier measurements are calculated from differences between a current mean value and previous mean value within the same time period.

So each attributes of (k) value are independent that expect the region to contain a fraction f^k of the records should be with in a range of region between data point (p) to Kth distance vector value (d).

Using the similarity assessment can measure as follows,

$$S(D)=\frac{n(D)-N.f^k}{\sqrt{N}.p^k.(1-f^k)}$$

This data set is taken from fine human breast tissue was generally used among researchers who use machine learning (ML) methods for breast cancer classification by our proposed work. The WBCD was used and analyzed.

3.2.4. Temporal Outlier Detection

The temporal outliers are those observations of outlier on a specific time slice where patterns of outlier and WPBC dataset show differences. Patterns are considered similar when the data points of the lower upper and kernel set are placed within the tolerance of the outlier data points.

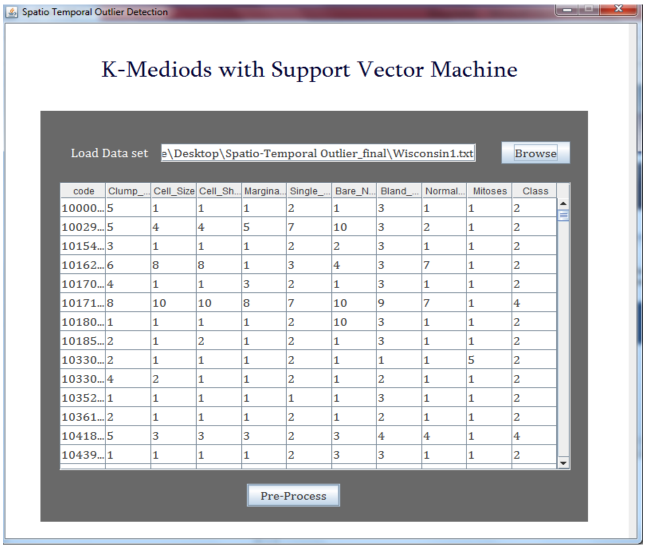
In other words, if the data points of the WPBC dataset at specific hour lay outside the outlier data point's tolerance, a subsequent relevant outlier observation of that hour is labeled as a kernel set outlier, otherwise labeled as normal.

The tolerance is defined based on the accuracy of the wireless sensor device, presented in table 1. Based on the accuracy of the device, the maximum and minimum range for each observation can be estimated. Thus, the tolerances are defined based on possible minimum and maximum variation for each outlier sequence.

3.3 System Analysis

3.3.1. Loading breast cancer dataset

In this form to load the Wisconsin Pattern Breast Cancer (WPBC) dataset and to pre-processing the all data.



K-Mediods with Support Vector Machine

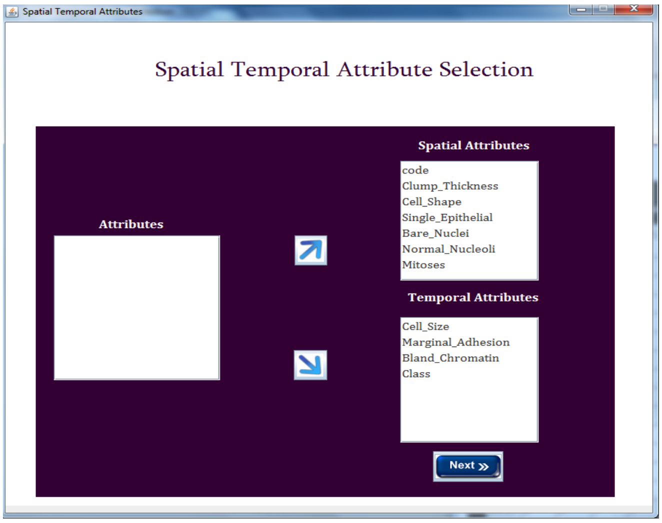
Load Data set e\Desktop\Spatio-Temporal Outlier_final\Wisconsin1.txt

code	Clump_...	Cell_Size	Cell_Sh...	Margina...	Single_...	Bare_N...	Bland_...	Normal...	Mitoses	Class
10000...	5	1	1	1	2	1	3	1	1	2
10029...	5	4	4	5	7	10	3	2	1	2
10154...	3	1	1	1	2	2	3	1	1	2
10162...	6	8	8	1	3	4	3	7	1	2
10170...	4	1	1	3	2	1	3	1	1	2
10171...	8	10	10	8	7	10	9	7	1	4
10180...	1	1	1	1	2	10	3	1	1	2
10185...	2	1	2	1	2	1	3	1	1	2
10330...	2	1	1	1	2	1	1	1	5	2
10330...	4	2	1	1	2	1	2	1	1	2
10352...	1	1	1	1	1	1	3	1	1	2
10361...	2	1	1	1	2	1	2	1	1	2
10418...	5	3	3	3	2	3	4	4	1	4
10439...	1	1	1	1	2	3	3	1	1	2

Fig 3.4 Loading Dataset

3.3.2 Spatial and Temporal Attribute Selection

To select the spatial and temporal attributes in the breast cancer attributes. This is find the more outlier and improve the accuracy.



Spatial Temporal Attribute Selection

Spatial Attributes

code
Clump_Thickness
Cell_Shape
Single_Epithelial
Bare_Nuclei
Normal_Nucleoli
Mitoses

Temporal Attributes

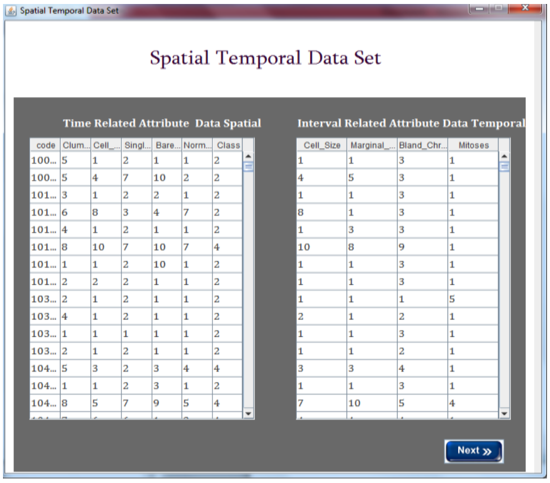
Cell_Size
Marginal_Adhesion
Bland_Chromatin
Class

Fig 3.5 Spatial and Temporal Attribute Selection

The Code, Clump_Thickness, Cell_Shape, Single_Epithelial, Bare_Nucleoli, Mithoses are the spatial attributes. The spatial is to find the location of attribute. The Ceel_Size, Marginal_Adhesion, Bland_Chromatin, Class are the temporal attributes. The temporal is to find the timing oriented attributes.

3.3.3. Clustering the Spatial and Temporal Dataset

To clustering the Spatial and Temporal breast cancer dataset. This is to find the more outliers and improve the accuracy.



Spatial Temporal Data Set

Time Related Attribute Data Spatial

code	Clum...	Cell_...	Singl...	Bare...	Norm...	Class
100...	5	1	2	1	1	2
100...	5	4	7	10	2	2
101...	3	1	2	2	1	2
101...	6	8	3	4	7	2
101...	4	1	2	1	1	2
101...	8	10	7	10	7	4
101...	1	1	2	10	1	2
101...	2	2	2	1	1	2
103...	2	1	2	1	1	2
103...	4	1	2	1	1	2
103...	1	1	1	1	1	2
103...	2	1	2	1	1	2
104...	5	3	2	3	4	4
104...	1	1	2	3	1	2
104...	8	5	7	9	5	4

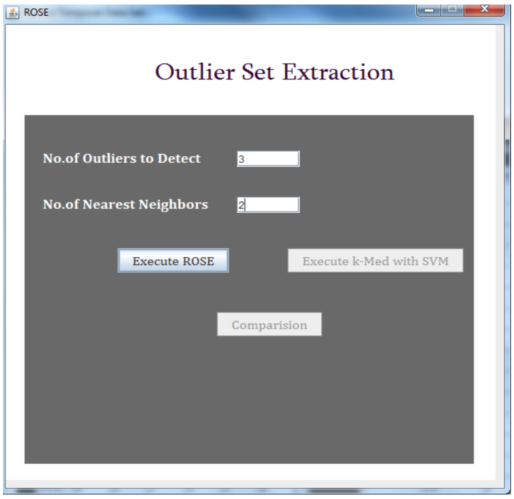
Interval Related Attribute Data Temporal

Cell_Size	Marginal_...	Bland_Chr...	Mitoses
1	1	3	1
4	5	3	1
1	1	3	1
8	1	3	1
1	3	3	1
10	8	9	1
1	1	3	1
1	1	3	1
1	1	1	5
2	1	2	1
1	1	3	1
1	1	2	1
3	3	4	1
1	1	3	1
7	10	5	4

Fig 3.6 Clustering the Spatial and Temporal Dataset

3.3.4. Outlier Set Extraction

In this form to extract the outlier set. The outliers and nearest neighbors values are given for the user need and get the result.



No.of Outliers to Detect: 3

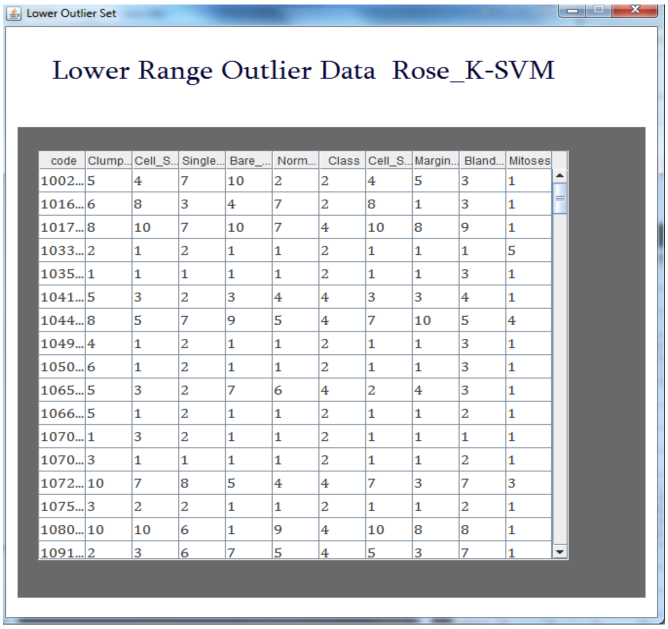
No.of Nearest Neighbors: 2

Fig 3.7 Outlier Set Extraction

In this form execute the K-medoids algorithm. It clustering the data and detect the more outliers. And also compare the existing rough outlier set extraction (ROSE), K-medoids with support vector machine.

3.3.5. Lower Range Outlier Data

The Lower possibility of outlier data will be displayed. The Lower possibilities of cancer patient details are listed.

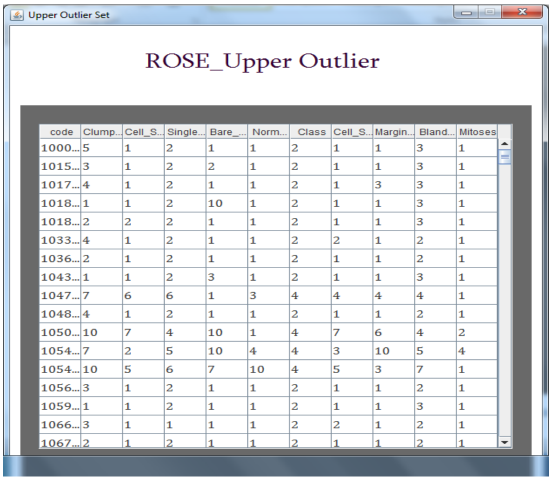


code	Clump...	Cell_S...	Single...	Bare_...	Norm...	Class	Cell_S...	Margin...	Bland...	Mitoses
1002...	5	4	7	10	2	2	4	5	3	1
1016...	6	8	3	4	7	2	8	1	3	1
1017...	8	10	7	10	7	4	10	8	9	1
1033...	2	1	2	1	1	2	1	1	1	5
1035...	1	1	1	1	1	2	1	1	3	1
1041...	5	3	2	3	4	4	3	3	4	1
1044...	8	5	7	9	5	4	7	10	5	4
1049...	4	1	2	1	1	2	1	1	3	1
1050...	6	1	2	1	1	2	1	1	3	1
1065...	5	3	2	7	6	4	2	4	3	1
1066...	5	1	2	1	1	2	1	1	2	1
1070...	1	3	2	1	1	2	1	1	1	1
1070...	3	1	1	1	1	2	1	1	2	1
1072...	10	7	8	5	4	4	7	3	7	3
1075...	3	2	2	1	1	2	1	1	2	1
1080...	10	10	6	1	9	4	10	8	8	1
1091...	2	3	6	7	5	4	5	3	7	1

Fig 3.8 Lower Range Outlier Data

3.3.6. Upper Range Outlier Data

The Upper possibility of outlier data will be displayed. The upper possibilities of cancer patient details are listed.

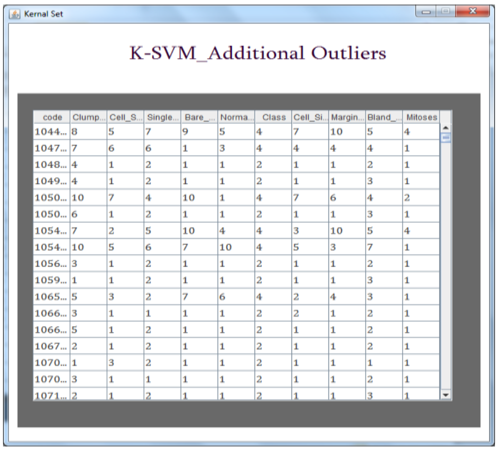


code	Clump...	Cell_S...	Single...	Bare_...	Norm...	Class	Cell_S...	Margin...	Bland...	Mitoses
1000...	5	1	2	1	1	2	1	1	3	1
1015...	3	1	2	2	1	2	1	1	3	1
1017...	4	1	2	1	1	2	1	3	3	1
1018...	1	1	2	10	1	2	1	1	3	1
1018...	2	2	2	1	1	2	1	1	3	1
1033...	4	1	2	1	1	2	2	1	2	1
1036...	2	1	2	1	1	2	1	1	2	1
1043...	1	1	2	3	1	2	1	1	3	1
1047...	7	6	6	1	3	4	4	4	4	1
1048...	4	1	2	1	1	2	1	1	2	1
1050...	10	7	4	10	1	4	7	6	4	2
1054...	7	2	5	10	4	4	3	10	5	4
1054...	10	5	6	7	10	4	5	3	7	1
1056...	3	1	2	1	1	2	1	1	2	1
1059...	1	1	2	1	1	2	1	1	3	1
1066...	3	1	1	1	1	2	2	1	2	1
1067...	2	1	2	1	1	2	1	1	2	1

Fig 3.9 Upper Range Outlier Data

3.3.7. Kernel set of Outlier data

The Upper possibility of outlier data will be displayed. And the new set addition upper possibilities of cancer patient details also listed.



K-SVM_Additional Outliers

code	Clump...	Cell_S...	Single...	Bare_...	Norma...	Class	Cell_Si...	Margin...	Bland_...	Mitoses
1044...	8	5	7	9	5	4	7	10	5	4
1047...	7	6	6	1	3	4	4	4	4	1
1048...	4	1	2	1	1	2	1	1	2	1
1049...	4	1	2	1	1	2	1	1	3	1
1050...	10	7	4	10	1	4	7	6	4	2
1050...	6	1	2	1	1	2	1	1	3	1
1054...	7	2	5	10	4	4	3	10	5	4
1054...	10	5	6	7	10	4	5	3	7	1
1056...	3	1	2	1	1	2	1	1	2	1
1059...	1	1	2	1	1	2	1	1	3	1
1065...	5	3	2	7	6	4	2	4	3	1
1066...	3	1	1	1	1	2	2	1	2	1
1066...	5	1	2	1	1	2	1	1	2	1
1067...	2	1	2	1	1	2	1	1	2	1
1070...	1	3	2	1	1	2	1	1	1	1
1070...	3	1	1	1	1	2	1	1	2	1
1071...	2	1	2	1	1	2	1	1	3	1

Fig 3.10 Kernel set of Outlier data

IV. PROPOSED RESEARCH NETWORK ENVIRONMENT

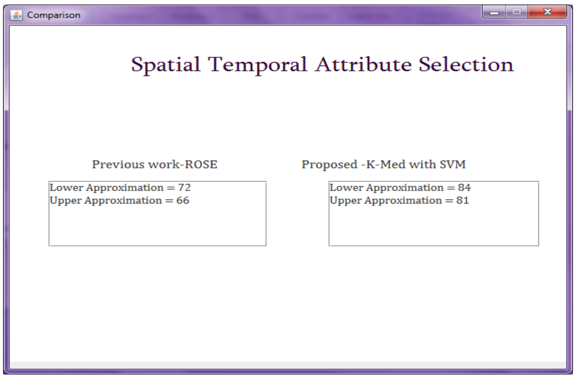
The K-medoids running for spatial outlier detection result and the best tradeoff between better outlier detection percent and FPR(False Positive Result) of the reported results are more expected than ROSE existing work. Globally, but all the way achieved K-medoids results outperform the compared state-of-the-art techniques on this spatiotemporal data set.

4.1.1. Cluster Data Analysis

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group.

The analysis are generated data set with two clusters 1015 total data points where 15 of them were generated as global outliers. After applying K-medoids technique with 0.05 as a probable vale, all the expected global outliers were detected and all the additional points were detected where some of those can be considered as local outlier's accuracy with existing rose.

K-medoids algorithms approach also was able to detect all above labeled outliers correctly producing all K-medoids values corresponding to outliers greater than existing system.



Spatial Temporal Attribute Selection

Previous work-ROSE

Lower Approximation = 72
Upper Approximation = 66

Proposed -K-Med with SVM

Lower Approximation = 84
Upper Approximation = 81

Fig 4.1 Outlier detection: K-medoids Technique – with ROSE for 699 data points

4.1.2. Path Data Analysis

Path analysis is a method employed to determine whether or not a multivariate set of non-experimental data fits well with a particular causal model. Path analysis allows specifying a model and relationships between variables. The data sets generated with different behaviors. Here the set of points that is located on curved paths and some deviated points as well. This set consists with 1000 normal data and 23 significantly deviated points. The represents output results of outlier detection using the proposed method. Generating equivalent results to K-medoids approach, in this K-medoids classical partitioning technique also detected 22 outliers with values greater than existing work.

4.2. Experimental Performance Results

The spatiotemporal data set labeled with three different methods, showing the different results (lower, upper, and kernel) on the basis of techniques. The ambient attribute is the analyzed feature for each station.

This experimental setup which was done in java with net beans 8.0. It works under 699 records in for WPBC dataset. That describes experiments and results with synthetic data sets followed by how the data was generated. It ran the experiments where time (t) was taken as 1.05. i.e., these experimental results are with 95% confidence. To cover the broad range of applications generated two main categories Time and Interval paths.

It use probabilistic distribution based data generation which takes user inputs to decide parameters of the data pattern, i.e., identify variables and then use a probabilistic model to generate the required number of data points and outliers. Clusters are the baseline choice of experimentation, and have been the focus of outlier detection algorithms.

The rigorous set of tests to path data to the understand strength (or weakness) of the method.

After generating data, each set of data points with feature scaling was tested both with proposed outlier detection method and with the K-medoids classical partitioning technique. Since it gives a degree of being an outlier of a point, there is no clear cutoff value differentiating normal points from outliers. For comparison and calculation purposes, we considered a data point with K-medoids value greater than existing work as an outlier.

The outlier detection iteration and calculation process is running and finding the lower, upper and kernel set in breast cancer dataset.

4.2.1. Performance of Rough Outlier Set Extraction (ROSE) and K-medoids with Support Vector Machine (SVM)

The K-medoids algorithm is gives the better performance. The maximum number of outliers is detected. The Existing algorithm (ROSE) is detected the lower approximation '72', and upper approximation '66'. The proposed method (K-SVM) is detected the lower approximation '84', and upper approximation '81'.

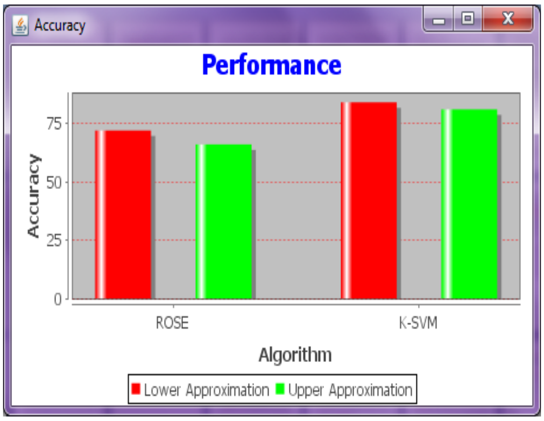


Fig 4.3 Performance of lower and upper approximation in ROSE-KSVM

4.2.2. Comparison of Rough Outlier Set Extraction (ROSE) and K-medoids with Support Vector Machine (SVM)

The existing Rough Set Outlier Extraction is compared with the proposed K-medoids SVM. To compare the performance evaluation of True Positive, False Positive, False Measure, Error-Reduction-Ratio, Time, Accuracy Low Outlier, and Accuracy Upper Outlier are gives the better result of K-medoids algorithm.

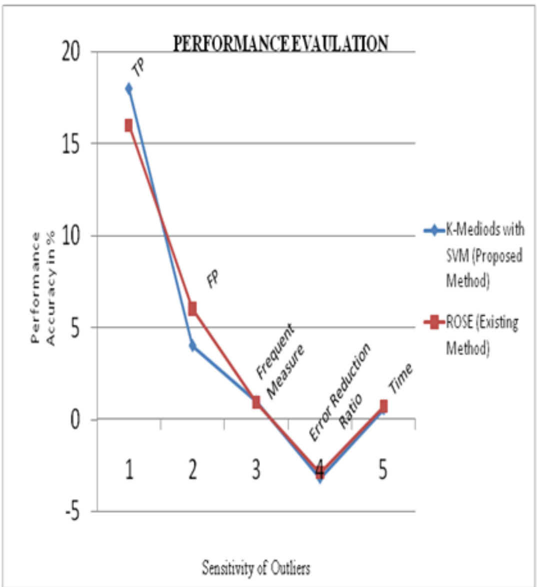


Fig 4.4 Comparison graph of ROSE-KSVM

4.3. RESULT

The simulation results demonstrate that the K-medoids has reduced time and error ratio and increased outliers. It performs much better compared to ROSE.

V. CONCLUSION

The K-medoids algorithm, which is designed to perform clustering on rich structured multivariate datasets. The applicability of Support Vector Machines (SVM) is not limited to classification problems and K-medoids SVM clustering can affect the performance of clustering algorithms for multivariate datasets.

Even in the absence of labeled training instances for support vector machine taken initiative and effectively increases clustering performance.

From the experimental results on the integration of authorship analysis with topical clustering of documents show significant improvements over ROSE (Rough Outlier Set Extraction) algorithm and confirms that there is added advantage for incorporating such type. Since spatial-temporal outlier detection might turn out to be useful in many different research fields, it will spark further interest in such problems that are challenging and relatively unexplored.

REFERENCES

[1] Alessia Albanese, Sankar K. Pal and Alfredo Petrosino, "Rough Sets, Kernel Set, and Spatiotemporal Outlier

Detection," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO.1, JAN. 2014

[2] Aggarwal. C. C, and Yu. P, "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 70-81, 2000.

[3] Aggarwal. C. C, and Yu.P.S, "An Effective and Efficient Algorithm for High-Dimensional Outlier Detection," VLDB J., vol. 14, pp. 211-221, 2005.