Five Steps of Achieving a Good Quality Clustering

Nidhi Sharma[1], Sachin Yele[2]

[1]*M.Tech Research Scholar Sanghvi Institute of Management and Science, Indore, India*

[2]*Assistant Professor Sanghavi Institute of Management and Science, Indore, India*

Abstract— Clustering is a manner of organizing the data into similar groups. The grouping of data can be performed by similarity among the properties or attributes. In the data mining processes clustering is a process of data analysis under unsupervised manner. The technique which works unsupervised need to not to have the class labels with the attributes or set of data, the algorithms evaluate each object of data and conclude the group according to the selected attributes. Most of the time the clustering is used to analyse a significant amount of data as compared to classification techniques, therefore the clustering algorithms has their own contribution and importance in data mining. In this presented work the clustering algorithms are explored and studied. Thus a number of clustering algorithms are observed. During study most of the techniques are suffered from the outlier issues and empty cluster issues. Therefore need to develop and design a technique that provides an optimal solution for recovering the addressed issues. The proposed solution includes the five stage solution for improving the data quality and obtaining the good clustering outcomes. The data pre-processing, normalization, outlier detection, centroid estimation and data integration are the basic steps of the proposed solution. The proposed solution provides a way by which the data analysis time and accuracy of clustering can be improved. The implementation of the proposed technique is performed using the JAVA development environment. After implementation the performance of the algorithm is estimated and compared with the traditionally available ECLARNS algorithm. The comparative analysis is made on the basis of accuracy, error rate, memory consumption and time complexity. The results demonstrate the effective and enhanced performance of the proposed algorithm as compared to traditional algorithm.

Keywords—clustering, data mining, unsupervised learning, data quality improvement, outlier detection.

I. INTRODUCTION

Data mining is a technique of analysing data and extraction of valuable patterns by which the recognition, prediction and other work is performed with human interaction. This automated technique of data evaluation need to include the computer based algorithms that are evaluating the data and producing the outcomes for the similar outcomes. There are two main kinds of computer based techniques are available namely supervised technique of data analysis frequently called the classification algorithms and the unsupervised learning techniques that are termed as cluster analysis of data.

According to the applications both kinds of the algorithms has their own contributions. When the data is available in small amount and the training samples are available with the pre-evaluated class labels then the classification algorithms are used for data analysis and when the data is available in significant amount and class label development is complex then the clustering algorithms are providing support for data evaluation and categorization.

A number of clustering algorithms and techniques are available among them two most frequently used techniques are hierarchical clustering and second are partition based clustering. The hierarchical clustering algorithm is used when the high accurate results are required because the amount of time consumption is higher in this kind of data evaluation. On the other hand the partition based clustering algorithms are efficient and consumes less time during the data analysis and the accuracy of the algorithm is also less.

In this context the partition based clustering techniques are suffers from the outlier points. These outlier points are misguiding the clustering algorithms and produce the deficiency in the clustering performance. Therefore the proposed is concentrated on finding the better approach of outlier detection and removal first by which the input data is amplified and performance of the clustering algorithm becomes improvable. This section provides the basic understanding of the proposed working domain for improving the clustering performance by reducing the outlier points in learning datasets.

II. PROPOSED WORK

The proposed method is based on the assumption of dataset enhancement and computation of the suitable centroid for obtaining dense and clear clusters from the input datasets. In order to find the proposed enhanced clustering technique the following process is followed.

1. Data pre-processing: the data by nature found in noisy nature, therefore the removal of the noise is necessary from

the data before making use of the data. Therefore the missing data is removed from the data, in addition of that the statically values and continuous values are also required to remove from the data which is used to make the cluster. To make the noise free data the following procedure is called as listed using the table 1.

Data pre-processing

Input: dataset D
Output: Pre-processed Dataset
Process:
• [Row, Col] = ReadData(D)
• for (i=1; i ≤ Row; i++)
  • for (j=1; j ≤ Col; j++)
  • if D(i,j)==null
  • Remove (D(i).row)
  • end if
  • end for
• end for
• for (k=1; k ≤ Col; k++)
  • if unique (D(k))==1
  • Remove (D(k))
  • End if
  • if unique (D(k))==Row
  • Remove (D(k))
  • End if
• End for
• Return D_p=D

Table 1 data pre-processing

2. Data normalization: as studied before the datasets are combination of different attributes sets thus the different attributes can have the different scales to be measured. Therefore in order to measure all the scaled data into a similar scale need to normalize the values of datasets. To perform the data normalization the min-max normalization technique is followed thus the following formula can be used for normalization process.

Therefore the following process is followed:

Data normalization

Input: pre-processed data
Output: Normalized Data
Process:
• [nRow, nCol] = readData()
• for (i=1; i ≤ nCol; i++)
• x_{max} = max(D_p(i))
• .x_{min} =min(D_p(i))
  • for (j=1; j ≤ nRow; j++)
  • x = D_p (i,j)
  • x' = (x-x_{min})/(x_{max}-x_{min})
  • Replace(x, x')
  • End for
• End for

Table 2 data normalization

3. Outlier detection: There are a number of techniques available for outlier detection and approximation, one of the effective methods is modified Thompson Tau test. That is used to determine outlier in a data set. The method provides a statistically rejection zone for estimating the outlier point in dataset. To compute the outlier first, data set's average is determined in further the absolute deviation between each data point and the average are determined and finally a rejection region is determined using the formula:

$$RR = \frac{T_{\frac{\alpha}{2}}(n-1)}{\sqrt{n}\sqrt{n-2+T_{\frac{\alpha}{2}}^2}}$$

Where

$T_{\frac{\alpha}{2}}$ = the critical value obtained in dataset t distributions,

n is training set size

s is the sample standard deviation.

To determine an outlier: need to compute

$$\delta = \left|\frac{X - mean(X)}{s}\right|$$

If δ > RR, then the data point is an outlier. If δ ≤ RR, the data point is not an outlier.

4. Centroid selection: the finally computed and refined data from all the step of solutions above described is used for cluster analysis. In first step the optimal cluster head or centroid is selected by evaluation of data, in this process the following procedure is taken place.

Centroid selection

Input: refined , number of clusters K
Output: K data points
Process:
• [Row, Col] = readData()
• for (i=1; j ≤ Row; i++)
  • S_{data} = R_{data}(i)
  • TempW=0

- for (j=1; j ≤ Row; j++)
  - $C_{data} = R_{data}$ (j)

$$TempW = TempW + \sqrt{\sum_{j=1}^{N} (C_{data}^{j} - S_{data}^{i})^2}$$

  - End for
  - Weight[i] = TempW
- End for
- FinalW[Row] = sort(Weight[Row])
- select k index from FinalW[Row]
- return k points from dataset

Table 3 Centroid selection

5. Data integration: the selected k points from the previous phase of data analysis are used for obtaining the final clusters of the data. That is just derived by comparing the estimated weights against the selected centroid points.

The proposed technique of improved data cluster formation is described in detail in this section in the next section results analysis is reported.

III. RESULTS ANALYSIS

The given section provides the results evaluation and the performance analysis of the proposed and traditional ECLARNS algorithm. Therefore to compare the performance essential performance factors are evaluated and their results are reported.

A. Accuracy

The accuracy is the amount of data that are accurately clustered during evaluation of data using the implemented clustering algorithm. That can also be evaluated using the following algorithm. The computed accuracy of both the algorithms is given using figure 1 for performing the comparative results evaluation.

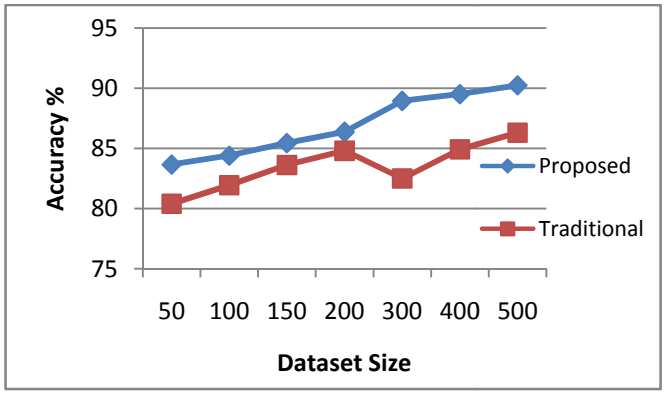

Figure 1 accuracy

Dataset size	Proposed algorithm	Traditional algorithm
50	83.67	80.41
100	84.41	81.96
150	85.46	83.64
200	86.39	84.82
300	88.96	82.51
400	89.52	84.94
500	90.25	86.32

Table 4 accuracy

The comparative performance of both the algorithms is given using figure 1 and table 4. In this diagram the X axis demonstrate the dataset size and the Y axis shows the obtained accuracy of the algorithms. According to the obtained results the proposed algorithm replicate the higher accurate cluster formation of the proposed algorithm as compared to the traditional algorithm. In addition of that for demonstrating the more clear difference among the performance of both the implemented unsupervised learners. The mean accuracy is reported using the figure 2. According to the obtained mean performance of the algorithms the proposed algorithm demonstrates much higher accurate performance as compared to the traditional algorithm. The proposed algorithm shows approximately 80-90% of accuracy of the different data sets in addition of that the traditional algorithm shows the 80-85% accurate results over the different kinds of dataset evaluation. Therefore the proposed algorithm is adoptable and efficient for performing the clustering over the different data.

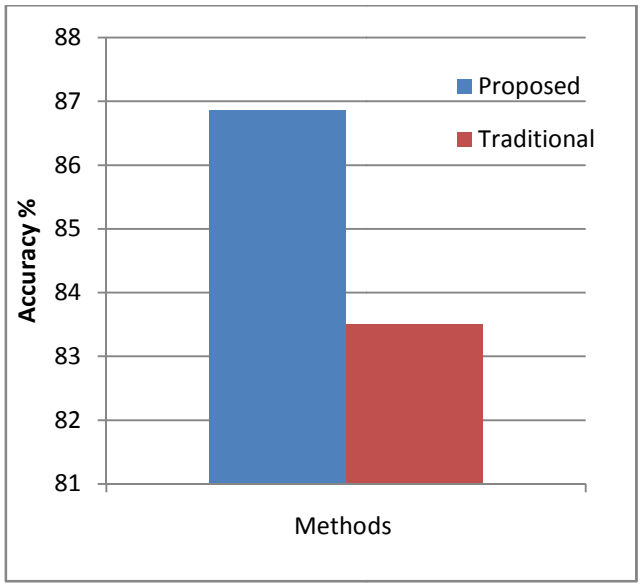

Figure 2 mean accuracy

B. Error rate

The amount of data misclassified using the implemented algorithms are termed as the error rate of algorithm. That can be evaluated using the following formula

.Or

The comparative error rate in terms of percentage error rate is given using the figure 3 and the table 5. The given results are reported according to the increasing size of data sets. In addition of that for representing the performance of the algorithm X axis contains the different size of experimental dataset and the Y axis shows the percentage error rate obtained during the experimentation. According to the obtained results the proposed algorithm produces less error rate as compared to the traditional algorithm. Thus the proposed algorithm is more adoptable then the traditional algorithm. In order to demonstrate clearer comparative analysis the mean error rate percentage is evaluated. According to the obtained performance the proposed algorithm produces approximately 14% of error rate and the traditional algorithm produces 17% of error rate. Therefore the proposed algorithm improves the performance about 4%.

Dataset size	Proposed algorithm	Traditional algorithm
50	16.33	19.59
100	15.59	18.04
150	14.54	16.36
200	13.61	15.18
300	11.04	17.49
400	10.48	15.06
500	9.75	13.68

Table 5 error rate

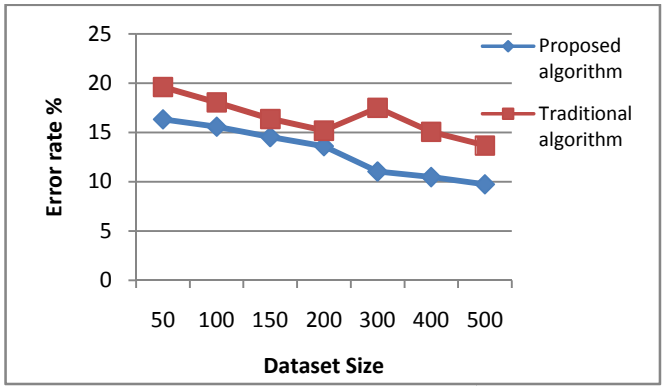

Figure 3 error rate

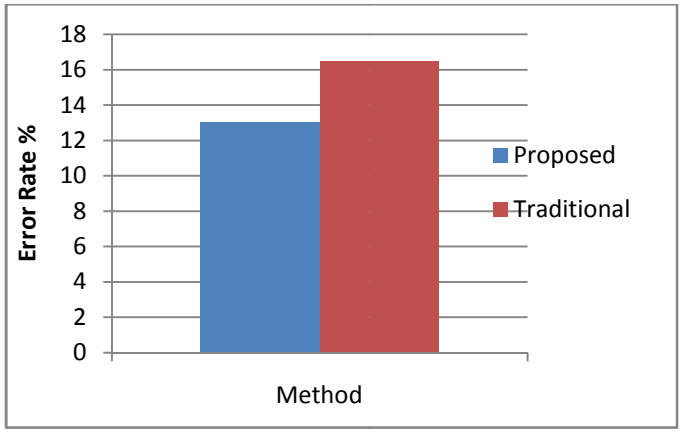

Figure 4 mean error rate

C. Memory usages

The amount of main memory required to execute the algorithm for obtaining the patterns from the input size of data is termed as the memory usage or the space complexity. The figure 5 and table 6 shows the comparative performance of the algorithms.

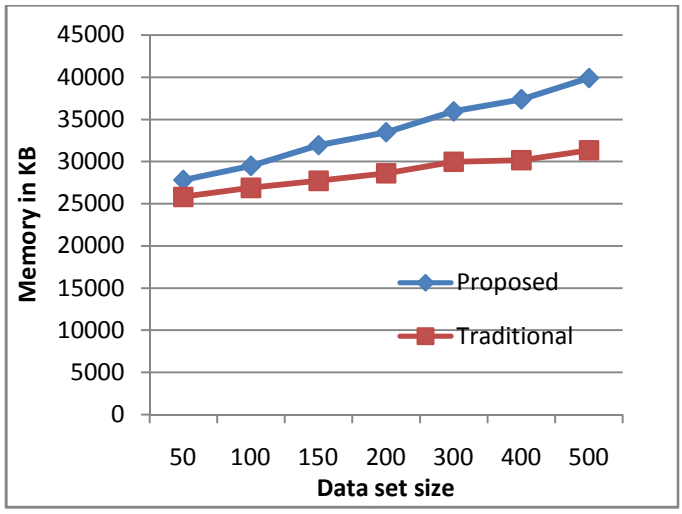

Figure 5 memory consumption

Dataset size	Proposed	Traditional
50	27817	25818
100	29488	26891
150	31938	27716
200	33462	28612
300	35938	29981
400	37362	30164
500	39882	31332

Table 6 memory consumption

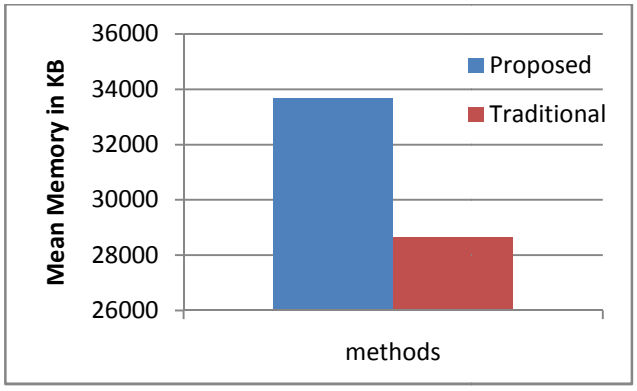

Figure 6 mean memory consumption

In this diagram the X axis contains the increasing size of experimental datasets and the Y axis shows the amount of main memory consumed in terms of KB (kilobytes). According to the given comparative results the performance of the proposed algorithm is lacked as compared to the traditional algorithm. Because the proposed algorithm needs to load most of the data into main memory for evaluation of data thus the traditional algorithm perform more efficiently as compared to the proposed algorithm. In order to show the clear memory consumption the mean memory consumption is demonstrated using the figure 6. According to the given results the proposed algorithm produces higher memory consumption as compare to traditional algorithm for data clustering.

D. Time consumption

The amount of time required to perform the clustering is termed here as the time consumption of the algorithms. The comparative time consumption of the algorithms is given using figure 7 and table 7. In this diagram the amount of dataset instances are given in X axis and the Y axis shows the amount of time consumed in terms of milliseconds. According to the obtained performance the proposed clustering schemes improve the time consumption more effectively as compared to the traditional approach of clustering. In this diagram the amount of time in traditional scheme is increase more rapidly as compared to the proposed algorithm as the amount of data for clustering is increases. For representing more relevant comparative outcomes the figure 8 shows the mean time consumption of both the algorithms this also demonstrates the effective performance of the proposed algorithm and reduces significant amount of time consumption.

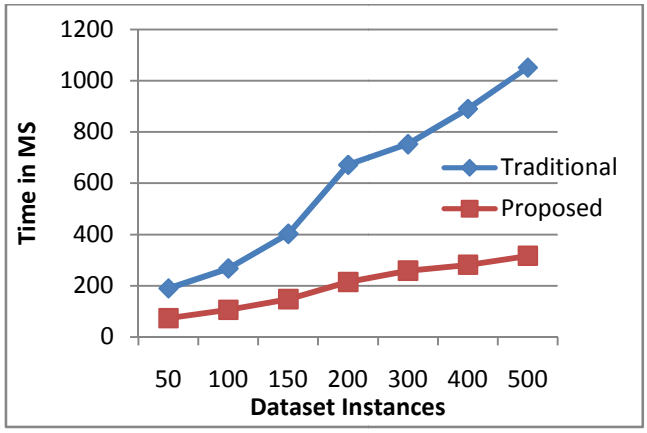

Figure 7 time consumption

Dataset size	Traditional	Proposed
50	189	73
100	267	105
150	402	147
200	671	214
300	752	258
400	890	281
500	1051	316

Table 7 time consumption

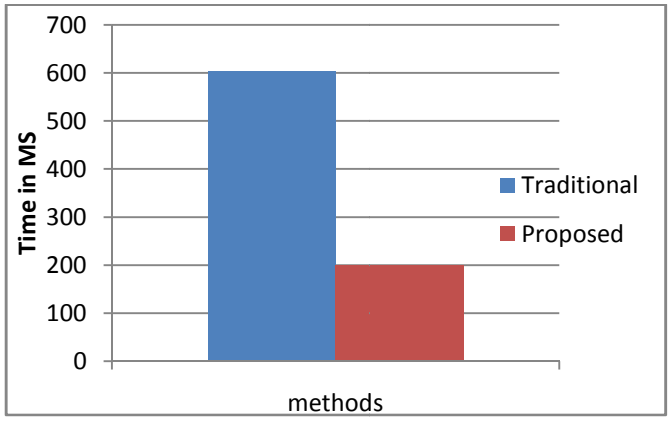

Figure 8 mean time consumption

IV. CONCLUSIONS

The proposed work for obtaining the efficient and accurate cluster formation technique by improving the traditional methods of clustering is performed successfully and their performance is also evaluated with different kinds of datasets. The given chapter includes the summarized facts obtained during the experimentation and design of the proposed algorithm. In addition of that the future extension of the proposed technique is also included.

A. Conclusion

The data mining is a broad view for utilizing the same data in different aspects and the applications. The similar data can be used for obtaining different kinds of patterns and information. In this context two different kinds of techniques are frequently used first classification and secondly the cluster analysis of data. The classification is a supervised approach of data analysis and the clustering is an unsupervised approach for pattern recovery. In most of the cases the supervised learning approach is used which small amount of data and pre-determined patterns labels. But in the cluster analysis the data can be any size and type and also need not to be labeled. Therefore the clustering is an effective process of data analysis.

In this presented work the cluster analysis techniques are investigated in detail. Therefore a number of clustering approaches are studied, during investigation of the different approaches of cluster analysis a common problem is identified in most of the clustering algorithms more specifically, empty clustering issue and outlier points. Therefore the key focus of the study is placed on the empty cluster problem and outlier detection techniques. In further for optimizing the addressed issues a new clustering scheme is proposed. In this approach there are five phases of the data analysis. In first three phases the data is pre-processed, cleaned and the outlier points are identified for improving the quality of data. Finally the weighted clustering scheme is utilized for identifying the best centroids. And using the selected centroids the clustering of data is performed.

The implementation of the proposed technique is performed using the JAVA development technology and their performance is evaluated. The detailed analysis of results shows the effective performance of the proposed technique and also reduces the time of data evaluation. The performance summary of the proposed technique and traditional technique is given using table 8.

S. No.	Parameters	Proposed	Traditional
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory usages	High	Low
4	Time consumption	Low	High

Table 8 performance summary

According to the obtained performance the proposed technique is efficient and effective for numerical data analysis.

B. Future extension

The proposed technique is an efficient and adoptable for all the evaluated performance factors. But the given method has some limitations to evaluate the data such as large memory consumption and limit to use for numerical data analysis. In near future both the issues are required to resolve before utilizing with the real world data.

V. REFERENCES

[1] Shruti Aggrwal, Prabhdip Kaur, “Survey of Partition Based Clustering Algorithm Used for Outlier Detection”, International Journal for Advance Research in Engineering and Technology, Volume 1, Issue V, June 2013

[2] Sivaram K, Saveetha D, “An Effective Algorithm for Outlier Detection”, Global Journal of Advanced Engineering Technologies, Vol2-Issue1-2013 ISSN: 2277-6370

[3] Data mining Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber, http://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts. and .Techniques.2nd.Ed.pdf

[4] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[5] Mahak Chowdhary, Shrutika Suri and Mansi Bhutani, “Comparative Study of Intrusion Detection System”, 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

[6] Mrs. Pradnya Muley, Dr. Anniruddha Joshi, “Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)

[7] Ghazaleh Khodabandelou, Charlotte Hug, Rebecca Deneckere, Camille Salinesi, “Supervised vs. Unsupervised Learning for Intentional Process Model Discovery”, Business Process Modeling, Development, and Support (BPMDS), Jun 2014, Thessalonique, Greece. pp.1-15, 2014

[8] Importance of Predictive Analytics in Business, http://www.orchestrate.com/blog/importance-of-predictive-analytics-in-business/

[9] David A. Dickey, N. Carolina State U., Raleigh, NC, “Introduction to Predictive Modeling with Examples”, Statistics and Data Analysis, SAS Global Forum 2012

[10] Hand, Manilla, & Smyth, “Descriptive Modeling”, http://www.stat.columbia.edu/~madigan/DM08/descriptive.ppt.pdf

[11] K. Jayavani, "Statistical Classification in Machine Intelligence", ISRJournals and Publications, Volume: 1 Issue: 1 18-Jul-2014, I

[12] Data Mining - Cluster Analysis, http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[13] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Carlos A. Coello Coello, "Survey of Multi-Objective Evolutionary Algorithms for Data Mining: Part-II",

[14] M.I. López, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums", Proceedings of the 5th International Conference on Educational Data Mining

[15] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, 2012 Elsevier Ltd. All rights reserved.

[16] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur, "Efficient K-Means Clustering Algorithm Using Ranking Method in Data Mining", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[17] Dr. S. Vijayarani, Ms. P. Jothi, "An Efficient Clustering Algorithm for Outlier Detection in Data Streams", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013

[18] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, Arthur Zimek, "Outlier Detection in Arbitrarily Oriented Subspaces", Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, 2012.

[19] Alessia Albanese, Sankar K. Pal, and Alfredo Petrosino, "Rough Sets, Kernel Set and Spatio-Temporal Outlier Detection", IEEE Transections of Knowledge and Data Engineering, VOL, NO , DECEMBER 2011, November 18, 2012

[20] Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi, "An Outlier Detection Method based on Clustering", 2011 Second International Conference on Emerging Applications of Information Technology

[21] Manish Gupta, Jing Gao, Yizhou Sun, Jiawei Han, "Integrating Community Matching and Outlier Detection for Mining Evolutionary Community Outliers", permission and/or a fee. KDD'12, August 12–16, 2012, Beijing, China Copyright 2012 ACM 978-1-4503-1462-6 /12/08

[22] Fuyuan Cao, Jiye Liang, Deyu Li, Liang Bai, Chuangyin Dang, "A dissimilarity measure for the k-Modes clustering algorithm", Knowledge-Based Systems, 2011 Elsevier B.V. All rights reserved.