Improved Hybrid Clustering and Distance-based Technique for Outlier Removal

P. Murugavel,
Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Dr. M. Punithavalli,
Research Supervisor, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Abstract-Outliers detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. It has many uses in applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. Using clustering algorithms for outlier detection is a technique that is frequently used. The clustering algorithms consider outlier detection only to the point they do not interfere with the clustering process. In these algorithms, outliers are only by-products of clustering algorithms and they cannot rank the priority of outliers. In this paper, three partition-based algorithms, PAM, CLARA and CLARANs are combined with k-medoid distance based outlier detection to improve the outlier detection and removal process. The experimental results prove that CLARANS clustering algorithm when combined with medoid distance based outlier detection improves the accuracy of detection and increases the time efficiency.

Keywords : CLARA, CLARANS, Cluster, Medoid-based Clustering, Outlier Detection, PAM.

I. INTRODUCTION

The current era of information explosion utilizes powerful data mining techniques that can efficiently analyze, interpret and extract valuable knowledge. The rapid growth in the number and size of databases, dimension and complexity of data has made it necessary to automate the analysis process, whose results can then be used by decision-making processes. The techniques used for this purpose can be grouped into four main categories. They are clustering, classification, dependency detection and outlier detection. Out of the four techniques, outlier detection is considered crucial in many research areas and application domains. The reason behind such popularity is that the failure to detect or incorrect treatment of outliers has a direct impact on the validity of the knowledge discovered during mining process. It is therefore important to identify them prior to modeling and analysis ([25], [16]).

According to [5], outlier detection is a task that finds objects that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data. Outlier detection has wide applications which include data analysis, financial fraud detection, network intrusion detection and clinical diagnosis of diseases. In data analysis applications, outliers are often considered as error or noise and are removed once detected. Examples include skewed data values resulting from measurement error, or erroneous values resulting from data entry mistakes. Approaches to detect and remove outliers have been studied by several researchers. Some techniques are developed for certain application domains, while others are more generic [7]. These approaches can be classified into Distribution-based approaches, Depth-based approaches, Clustering-based approaches, Distance-based approaches and Density-based approaches. Each of these methods has its own advantages and disadvantages. In general, in all these methods, the approach to detect outliers consists of two steps. The first identifies a profile around a data set using a set of inliers (normal data). In the second step, a data instance is analyzed and is identified as outlier when its attributes are different from the attributes of inliers. All these methods assume that all normal instances will be similar, while the anomalies will be different.

When the number of normal attributes is more than the abnormal behavioural attributes, a clustering-based approach to outlier detection provides more positive results. In these situations, the key assumption made here is that large and dense clusters have normal data and the data which do not belong to any cluster or small clusters (low dense clusters) are considered as outliers. Cluster-based methods either belong to semi-supervised or supervised categories. In semi-supervised techniques, the normal data is clustered to create modes of normal behaviour and instances which are not close to any clusters are identified as outliers. In unsupervised techniques, a post-processing step is included after clustering to determine the size of the clusters. The distance from the clusters is then calculated, using which the outliers are detected. Furthermore, depending on the method adopted to define clusters, the techniques can be further grouped as partitional clustering, hierarchical clustering, density-based clustering and grid-based clustering [21]. All these algorithms use the distance measure between two objects and clustering is based by grouping objects which have minimum distance from the centre of the cluster. The advantage of using cluster-based algorithm is that they are easily adaptable to incremental mode suitable for anomaly detection from temporal data. On the other hand, they are computationally expensive and large/dense clusters frequently have both inliers and outliers.

To solve this problem, recently, amalgamation of techniques for outlier detection is proposed and has gained more attention in recent years. These hybrid approaches combine techniques for efficient anomaly detection. In this paper, the partition clustering algorithm and distance-based outlier detection method are combined for efficient clustering and outlier detection. The main objective is to detect outliers while simultaneously perform clustering operation.

The authors of [15] initialized the concept of distance-based outlier, which defines an object 'O' being an outlier, if at most 'p' objects are within the distance 'd' of 'O'. In the distance-based approach, outliers are detected as follows. Given a distance measure on a feature space, a point q in a data set is an outlier with respect to the parameters M and d, if there are less than M points within the distance d from q, where the values of M and d are decided by the user. The problem with this approach is that it is difficult to determine the values of M and d. While considering medoid, the distance is calculated using Absolute Distances between Medoids.

Al-Zoubi [3] proposed a hybrid cluster-based outlier detection system, which used Partitioning Around Medoid (PAM) clustering algorithm and Absolute Distance between Medoid (ADMP) for distance-based outlier detection. This method produced good results with small datasets, but the performance degraded with large datasets. The reason for this degradation while scaling up the size of dataset was because of PAM clustering algorithm. To solve this, the present work considers two other algorithms, namely, CLARA (Clustering around LARge Applications) and CLARANS (Clustering Large Applications based on RAndomized Search) for clustering the data. After clustering, small clusters are removed as outliers. The outliers in the large clusters are then detected using a modified distance-based detection approach proposed by [3]. The paper compares the performance of all the four algorithms on outlier detection efficiency.

The Paper is organized as follows. Section II provides a brief discussion on the previous works related to the topic. Section III explains the working of PAM, CLARA and CLARANS algorithms. Section IV discusses the distance based techniques. The proposed methodology is presented in Section V, while Section VI presents the experimental results. A brief summary along with future research directions is given Section VII.

II. RELATED STUDIES

Many data mining algorithms in the literature find outliers as a side-product of clustering algorithms. In early days, a fuzzy based clustering approach was used by [8]. Ester et al. [9] proposed a density based clustering algorithms to discover outliers in large spatial databases. A similar work was proposed by [24] and [29]. Techniques that define outliers as data points outside clusters are presented by [22], [28], [26], [27], [6].

A two-phase method to detect outliers was proposed by [10]. The first phase used a modified k-means clustering algorithm, which clustered the data using the heuristic 'if one pattern is far from all the cluster center, then treat it as a new cluster centre'. Using the heuristic, successfully divided the data into cluster that contained either all outliers or all inliers. In the second phase, a minimum spanning tree (MST) was used to construct tree. From this tree, the subtree with minimum number of nodes was treated as outliers and was removed.

In 2004, [17] used the result of hierarchical clusters as indicators for the presence of outliers. Later, this method was enhanced by [2]. Recently, [23] has reviewed partition based clustering algorithms in the field of data mining.

An algorithm called PAMST, that combined Partitioning clustering algorithm (PAM) and Separation technique was introduced by [1]. The separation technique was calculated based on the dissimilarity between two data objects. Data objects with large separation value were considered as outliers. Jiang and An [11] proposed a clustering-based outlier detection method (CBOD). The CBOD used a two-step process to detect outliers. The first stage used a one-pass clustering algorithm to cluster the dataset and removed all small clusters as outliers. The second stage used a outlier factor to determine the outliers in the large clusters. Al-Zoubi *et al*. [4] presented a fuzzy clustering method to detect outliers. The rest of the outliers were then determined by computing the difference between the objective function values and when a noticeable change is noticed, the points are considered as outliers.

III. PARTITIONING ALGORITHMS

The procedure followed by partitioning algorithms can be stated as follows : "Given n objects, these methods construct k partitions of the data, by assigning objects to groups, with each partition representing a cluster. Generally, each cluster must contain at least one object; and each object may belong to one and only one cluster, although this can be relaxed". The present study analyzes the use of CLARA and CLARANS clustering algorithms for outlier detection. The performance of all these algorithms is compared with the PAM and this section provides the working behind these four algorithms.

A. Partitioning Around Medoid (PAM)

Kaufman *et al*. in 1987 [14] developed a k-medoids-based clustering called PAM. A medoid is defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set. PAM procedure is given in Figure 1, where k is the number of

clusters, n is the number of objects in the datasets, S is the set of objects to be clustered, s_j is an object $\in$ S, R is the set of objects $\in$ S selected of medoids, r_j, $r_c \in$ R, d is the dissimilarity function. In the present study the Minkowski distance metric (Equation 1) is used.

$$\cos t(x, c) = \sum_{i=1}^{d} | x - c | \qquad (1)$$

where x is any data object, c is the mediod and d is the dimension of the object. Pam is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean.

The PAM algorithm forms clusters by examining all objects that are not medoids. This imposes an expensive computation cost of $O(k(n-k)^2)$ in each iteration [19]. This indicates that cost and size of the datasets are directly proportional, that is small datasets have low computation cost, as only a few iterations are required and makes PAM an impractical solution for large datasets.

B. Clustering LARge Applications (CLARA)

To eliminate the computational complexity problem of PAM algorithm, another partition based clustering algorithm called CLARA was introduced by [13]. The algorithm is outlined in Figure 2 [20]. This procedure, considers small samples of the actual data as a representatives of the data. PAM algorithm is used to identify the medoids for each of these samples. Then each object of the entire dataset is assigned to the resulting medoids. Similar to PAM, the objective function is computed to select the best set of medoids as output. Experiments described in [13] indicated that 5 samples of size 40+ 2k give satisfactory results. The computational complexity of each iteration of CLARA is of $O(ks^2 + k(n-k))$, where s is the size of the sample.

1. **Build Phase**: Randomly select two initial data points as medoids. The selection is made in such a way that the dissimilarity to all other data objects is minimal. The main objective of this step is to decrease the objective function.
2. **Swap Phase**: The Swap phase computes the total cost 'T' for all pairs of objects r_i and s_h, where ri $\in$ R is currently selected and sh $\in$ S is not.
3. **Selection Phase:** This phase selects the pair (r_i, s_h) which minimizes 'T'. If the minimum T is negative, the swap is carried out and the algorithm reiterates Step 2. Otherwise, for each non-selected object, the most similar medoid is found and the algorithm stops.

Figure 1 : PAM Procedure

1. For i=1 to 5, repeat Steps 2 to 5.
2. Draw a sample of 40 + 2k objects randomly from the entire data set and call PAM algorithm to find k medoids of the sample.
3. For each object O in the entire data set, determine k-medoids which is most similar to O.
4. Calculate average dissimilarity of the clusters obtained from Step 3. If this value is less than current minimum, use the new value as current minimum and retain the k medoids found in Step 2 as the best set of medoids obtained so far.
5. Return to Step 1 to start the next iteration.

Figure 2 : CLARA Procedure

C. Clustering Large Applications based on RAndomized Search (CLARANS)

CLARANS, another portioning algorithm, was developed by [19], as an improvement to CLARA to form clusters with minimum number of searches. The algorithm is given in Figure 3. CLARANS, similar to CLARA, does not check all nodes' neighbor. But, unlike CLARA, it does not restrict its search to a particular subgraph, but it searches the original graph. One key difference between CLARANS and PAM is that the former only checks a sample of the neighbors of a node. But, unlike CLARA, each sample is drawn dynamically in the sense that no nodes corresponding to particular objects are eliminated outright. In other words, while CLARA draws a sample of nodes at the beginning of a search, CLARANS draws a sample of neighbors in each step of a search. This has the benefit of not confining a search to a localized area

The CLARANS procedure depends on two parameters, namely, maxneighbor and numlocal. Maxneighbour is the maximum number of neighbors examined and numlocal is the number of local minima obtained. The higher the value of maxneighbor, the closer is CLARANS to PAM, and the longer is each search of a local minima. But, the quality of such a local minima is higher and fewer local minima need to be obtained. In the procedure, steps 3 to 6 search for nodes with progressively lower costs. But, if the current node has already been compared with the maximum number of the neighbors of the node (specified by maxneighbor) and is still of the lowest cost, the current node is declared to be a "local" minimum. Then, in Step 7, the cost of this local minimum is compared with the lowest cost obtained so far. The lower of the two costs above is stored in mincost. Algorithm CLARANS then repeats to search for other local minima, until numlocal of them have been found. The computational complexity is O(N2) where N is the number of objects [19].

1. Input parameters numlocal and maxneighbor. Initialize i to 1, and mincost to a large number.
2. Set current to an arbitrary node in Gn;k.
3. Set j to 1.
4. Consider a random neighbor S of current, and based on 5, calculate the cost differential of the two nodes.
5. If S has a lower cost, set current to S, and go to Step 3.
6. Otherwise, increment j by 1. If j maxneighbor, go to Step 4.
7. Otherwise, when j > maxneighbor, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set bestnode to current.
8. Increment i by 1. If i > numlocal, output bestnode and halt. Otherwise, go to Step 2.

Figure 3 : CLARANS Procedure

IV. PROPOSED METHODOLOGY

As understood from the literature study, clustering algorithms consider outlier detection but only to the point they do not interfere with the clustering process. In these algorithms, outliers are only by-products of clustering algorithms and they cannot rank the priority of outliers. Furthermore, algorithms that combine and compare the performance of using partition clustering algorithm combined with distance based outlier detection is not available. Most of the studies compare the clustering performance of these three algorithms but have not combined it with outlier detection. In the field of outliers, [18] studied the performance of their proposed fuzzy based outlier detection algorithm with CLARANS. A similar study was performed by [12] to propose an outlier detection technique that used CLARANS for detecting and reducing the outliers for clustering problems. In this study, the two algorithms CLARA and CLARANS are combined with a distance based outlier detection algorithm, to efficient detect outliers in large and small datasets. The proposed methodology is shown in Figure 4.

The algorithm first performs partition clustering using one of the algorithms PAM/CLARA/CLARANS. The algorithm produces a set of clusters and a set of medoids (cluster centers). In the next step, the average number of points in 'k' cluster is calculated (AKN) and the clusters are segregated as small and large clusters. All those clusters which have less than half of AKN are declared as small cluster. These small clusters are removed from the datasets as outliers or noise. The outliers in the large clusters are then detected using the following procedure. First, the Absolute Distances between the Medoid (μ) (ADMP) of the current cluster and each one of the points (p_i) is calculated using Equation 2. A threshold value is calculated as the average of all ADMP values of the same cluster multiplied by 1.5. When the ADMP value of a cluster is greater than T, then it is an outlier, else it is an inlier.

$$ADMP = |p_i - \mu| \quad (2)$$

EXPERIMENTAL RESULTS

The effectiveness of the clustering algorithm when combined with distance based outlier detection is presented in this section. Two benchmarked datasets, namely, iris dataset with four dimensions and three classes and Bupa dataset with six dimensions and two classes are used during experimentation. The iris dataset has three classes of Iris plants (Iris setosa, Iris versicolor and Iris viginica and has four variables/dimensions). It is previous established that there are ten outliers in class 3 (Iris viginica) in iris dataset [1]. The BUPA dataset has 2 classes and six dimensions. Again, [1] have found that there are 22 outliers in class1 and 26 doubtful outliers in class2, totaling to 48 outliers. Table I shows the effectiveness the algorithms in terms of correct identification of outliers in Iris and BUPA datasets.

TABLE I : Number of Outliers Detected

Dataset	PAM	CLARA	CLARAN
IRIS	7	7	9
BUPA (Class 1)	18	19	20
BUPA (Class 2)	17	18	22

It could be seen that the CLARANS method is better when compared with PAM and CLARA. While considering Iris dataset, CLARANS could identify 9 out of the 10 outliers correctly while it 7 out of 10 by PAM and CLARA. The performance of both PAM and CLARA are the same while using Iris dataset. The same trend was found with BUPA dataset also. The PAM algorithm detected 35 outliers, CLARA detected 37 outliers and the CLARANS method found 42 outliers respectively. Thus, it could be deduced that the CLARANS algorithm is accurately detecting the outliers from both the datasets.

The second performance parameter considered the time efficiency of the algorithms (Figure 5).

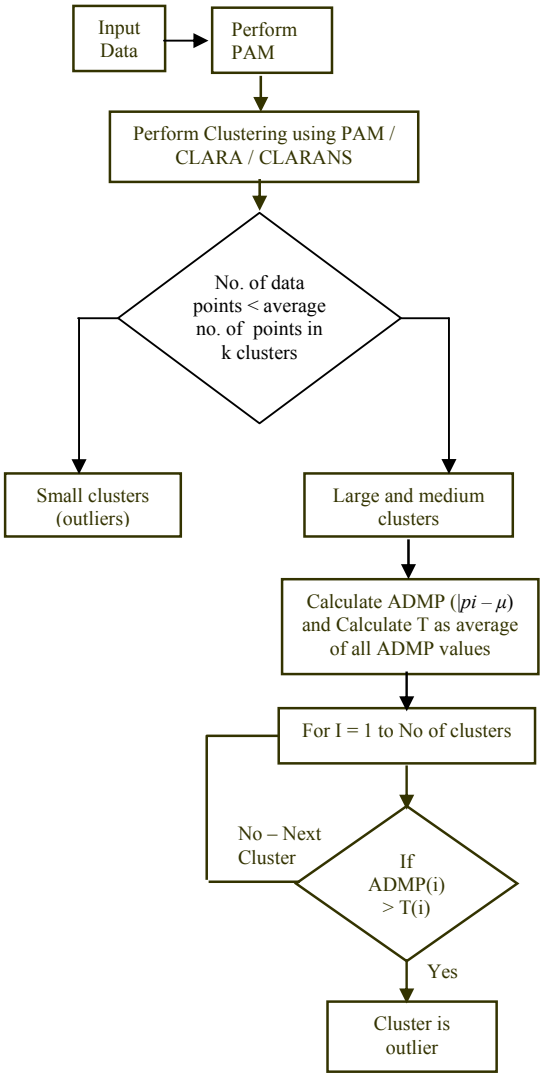

Figure 4 : Proposed Algorithm

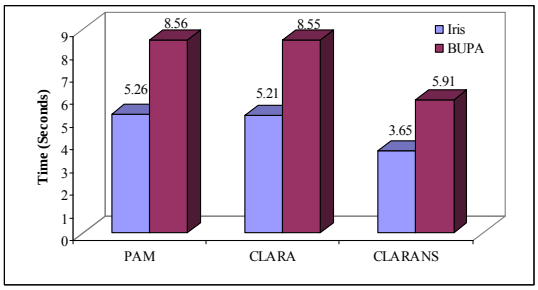

Figure 5 : Time Efficiency

From the figure, it could be seen the fastest algorithm is CLARANS, followed by CLARA AND PAM. The performance of CLARA and PAM is very close. According to [19], for small data sets, CLARANS is a few times faster than PAM; the performance gap for larger data sets is even larger. When compared with CLARA, CLARANS has the advantage that the search space is not localized to a specific subgraph chosen a priori, as in the case of CLARA. Consequently, when given the same amount of runtime, CLARANS can produce clusters that are of much better quality than those generated by CLARA.

CONCLUSION

This paper considered the use of three partition algorithms (PAM, CLARA and CLARANS) combined with distance based method for outlier detection. The main advantages of all these approaches is that they are all unsupervised methods, which means new data can be added to the database can be tested for outliers in future in an efficient manner. Experiments showed that CLARANS is the best candidate while considering outlier detection, followed by CLARA and PAM. The distance-based algorithm used requires distance calculation between each and every data point in the large clusters. This increases the time complexity of the detection process. The time complexity can be reduced by considering some shape based outlier detection can be combined with the present system and its effect on outlier detection can be studied.

REFERENCES

[1] Acuna E. and Rodriguez C. (2004) A Meta Analysis Study of Outlier Detection Methods in Classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, academic.uprm.edu/~eacuna/paperout.pdf, Proceedings IPSI 2004, Venice.

[2] Almeida, J., Barbosa, L., Pais, A. and Formosinho, S. (2007) Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering, Chemometrics and Intelligent Laboratory Systems, Vol. 87, Pp. 208–217.

[3] Al-Zoubi, M. (2009) An Effective Clustering-Based Approach for Outlier Detection, European Journal of Scientific Research, Vol.28, No.2,Pp. 310-316.

[4] Al-Zoubi, M., Al-Dahoud, A. and Yahya, A.A. (2010) New Outlier Detection Method Based on Fuzzy Clustering, WSEAS Transactions on Information Science and Applications, Vol. 7, Issue 5, Pp.681-690.

[5] Angiulli, F. (2009) Outlier Detection Techniques for Data Mining, John Wang (Ed.), Encyclopedia of Data Warehousing and Mining, Second Edition, Pp. 1483-1488

[6] Carvalho, R., and Costa, H. (2007) Application of an integrated decision support process for supplier selection, Enterprise Information Systems, Vol. 1, No. 2, Pp.197–216.

[7] Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection: A survey, ACM Computing Surveys (CSUR), ACM Digital Library,Vol. 41, Issue 3, Pp.1-58.

[8] Cutsem, B and Gath, I. (1993) Detection of Outliers and Robust Estimation using Fuzzy Clustering, Computational Statistics & Data Analyses, Vol. 15, Issue 1, Pp. 47-61.

[9] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noises, Proc. 2nd int. conf. on knowledge discovery and data mining, AAAI Press, Portland, Pp. 226–231.

[10] Jiang, M., Tseng, S. and Su, C. (2001) Two-phase Clustering Process for Outlier Detection, Pattern Recognition Letters, Vol. 22, Pp. 691-700.

[11] Jiang, S. and An, Q. (2008) Clustering-Based Outlier Detection Method, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 2, Pp.429-433.

[12] Karmaker, A. and Rahman, S. (2009) Outlier Detection in Spatial Databases Using Clustering Data Mining, Sixth International Conference on Information Technology: New Generations, Pp.1657-1658.

[13] Kaufman, L. and Rousseeuw, P. (1990) Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, New York.

[14] Kaufmann L, Rousseeuw P. Clustering by means of medoids (1987) Dodge Y, (ed.). Statistical Data Analysis Based on the L1 Norm and Related Methods, Elsevier Science, Pp. 405–416.

[15] Knorr, E.M. and Ng, R.T. (1998) Algorithms for mining Distance-based outliers in Large Datasets, VLDB

[16] Liu, H., Shah, S. and Jiang, W. (2004) On-line outlier detection and data cleaning, Computers and Chemical Engineering, Vol. 28, Pp. 1635-1647.

[17] Loureiro,A., Torgo, L. and Soares, C. (2004) Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[18] Mahfouz, M.A. and Ismail, M.A. (2009) Fuzzy relatives of the CLARANS algorithm with application to text clustering, World Academy of Science, Engineering and Technology, Vol. 49, Pp. 334-341.

[19] Ng, R. and Han, J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Conf. Very Large Databases, Pp. 144–155.

[20] Ng, R. and Han, J. (2002) CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering. Vol.14, No.5.

[21] Osama, A., Herna, V., Eric, P. and Marc, R. (2004) Exploring anthropometric data through cluster analysis, Society of Automotive Engineers, New York, NY, ETATS-UNIS (1927-200), Vol. 113, No. 1, Pp. 241-244.

[22] Qiu, G., Li, H., Xu, L., & Zhang, W. (2003) A knowledge processing method for intelligent systems based on inclusion degree, Expert Systems, Vol. 20, No.4, Pp. 187–195.

[23] Velmurugan, T. and Santhanam, T. (2011) A survey of partition based clustering algorithms in data mining: An experimental approach, Inform. Technol. J., Vol.10, Pp. 478-484.

[24] Wang, W., Yang, J. and Muntz, R. (1997) Sting: a Statistical Information Grid Approach to Spatial Data Mining, Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB), Pp. 186-195.

[25] Williams, G.J., Baxter, R.A., He, H.X., Hawkins, S. and Gu, L. (2002) A comparative study of RNN for outlier detection in data mining, IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan.

[26] Xu, L. (2006) Advances in intelligent information processing, Expert Systems, Vo. 23, No. 5, Pp.249–250

[27] Xu, L., Liang, N., & Gao, Q. (2008) An integrated approach for agricultural ecosystem management, IEEE Transactions on Systems Man and Cybernetics, Part C, Vol.38, No.3, Pp. 1-8.

[28] Zhang, M., Xu, L., Zhang, W. and Li, H. (2003) A rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory, Expert Systems, Vol.20, No.5, Pp.298-304.

[29] Zhang, T., Ramakrishnan, R. and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases, J. Widom (Ed.), Proceedings of the 1996 ACM SIGMOD international conference on management of data, SIGMOD'96 Montreal, Quebec, Canada, ACM Press, New York, Pp. 103–114. .