Clustering Of Datasets By Using K-Means & C-Means (Fuzzy) Methodology

Mr.Niraj N Kasliwal,
M-Tech, IT
RKDF
Bhopal,(India)

Prof Shrikant Lade,
HOD,IT
RKDF
Bhopal,(India)

Abstract

The integrated data mining processing technique to find appropriate initial centroids and Vectors in data clustering process by K-means and C-means algorithm. The processes include data cleansing, preprocessing, and finding features relation with Apriori algorithm to get appropriate features.Used K means and C means model that represents the processes for finding appropriate initial clustering centroids , vectors and selecting the most relevant features from large datasets. We can get better clustering result with k-means and C-means clustering methodology. The experimental result shows the differences in the working of both clustering methodology.

Index Terms*—Data mining, Apriori algorithm, K-means clustering, C- means (Fuzzy) clustering.*

1. Introduction

The data mining [1.19] is the automatic process of searching or finding useful knowledge. The process extracts data from large database with mathematics-based algorithm and statistic methodology to reveal the unknown data patterns that can be useful information. The information got from data mining process is very important knowledge that help user in decision making concerned business strategies. These processes are also called Knowledge Discovery in Database (KDD) in that knowledge discovery and analysis can be performed from many information and raw data in databases . The knowledge can be used in decision support system or used to predict customer's behavior or predict product sale rate in the future.

Fig . DATA MINING PROCESS

This paper studies various techniques to improve the data clustering by using K-means & C means clustering methods. The problems in data clustering with k-means are the selection of initial centroids that can effect to SSE in each cluster of data. Poor selection results in more time processing. But in case of C means its exactly opposite i.e good selection result in less time. The research has focused on the working of both clustering methodology. In this paper, the main idea of data mining technique in data clustering from raw data with appropriate initial centroids selection is presented. The techniques used in this paper are Apriori algorithm for feature selection process and clustering data with k-means & C means (Fuzzy) clustering methodology.

A. Apriori-based Algorithm

The association rules [1] are one of popular data mining techniques employed by several enterprise sectors, especially in retailing business. The association rules are to be used to analyze the sale rate and sold related goods in store. Entrepreneurs can predict and arrange the shelf of products that customers usually bought together. These rules represent in the format of "If…Then" rule that does not like other rules in data mining techniques for example clustering and classification. Mining for

association rules have many processes and use more time processing in finding related features in data groups . The result of association rule mining shows many rules which are combination of related features that users have to analysis and select usable set of features. This technique quite differs from classification technique because classification methodology shows the results that are specific to some class of data.

The combination of item sets, or features, from the result of association rule mining has many patterns with several groups of items. Users have to set threshold of minimum support value to limit the result that shows only groups of item sets related to the specified criteria. The results are also filtered by minimum confidence value that is to be specified by user corresponding to user's requirement and usage. Association results include group of related features called "item set" that are considered in each frequent item set, for example, examine two related features in co-occurrence type is called two-item set.

Although the association rules [2] are very effective to find relevant features, the method requires much time to process and analyze every possible item sets. This is due to the process that each item set will be considered and rules are to be generated in each group of item set. Thus association mining has many techniques to speed up time processing in the consideration of item sets . One of those techniques is Apriori-based algorithm. Apriori is a structure to count candidate item sets efficiently. It generates candidate item sets of length k from the k-1 item sets and avoids expanding all the item set's graph. Then it prunes the candidates which have an infrequent sub pattern. The candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. With Apriori technique the algorithm can decrease time processing in generating fewer groups of item sets and avoid infrequent candidate item sets expansion.

B. K-means Clustering Methodology

The data clustering [19] is processing of raw data to find clusters or groups of similar data. In each cluster, members have some similarity in type of data. The principles of data clustering are finding value of score in similarity, and assigning each member to be in the same group of other members that have similar or same score.

The data mining technique in finding data clusters is different from data classification in that user does not have to specify target feature for assigning each data record to the appropriate cluster. Data clustering is thus an unsupervised learning method. The clustering method relies on the similarity measurement to automatically from groups of relevant or similar data members as visually shown in figure. After the clustering process, user can apply some classification algorithm to extract data pattern in each cluster for a better understanding of cluster model.

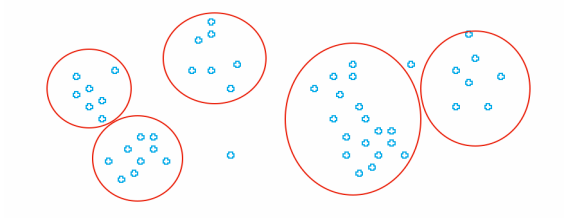

Fig: Clustering Visualization

K-means clustering algorithm is the most selected technique to cluster data. K-means is a nonhierarchical clustering and use looping to group data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The steps of k-means [19] algorithm can be summarized as the following:

1) Specify group number and select initial centroid of each group.
2) Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.
3) Calculate distance's mean of every data member and own centroid to define new centroid in each group.
4) Repeat steps 2 and 3 until each group has stable centroid or same centroid.

C. C-means Clustering Methodology

Fuzzy c-means (FCM) [7,8] is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by Dunn in 1973 and improved by Bezdek in 1981 is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

,

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. The algorithm is composed of the following steps:

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

4. *If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.*

2. Research Methodology

A . K-means Clustering Methodology

It provides a Processing K-means Algorithm model that shows the processes of preparation the suitable data, selection the good centroids, and achieving better clustering performance with their processing. The processes include preprocessing steps that cleansing raw data from KDD Cup data and select suitable features by association analysis with Apriori-based algorithm. After that clustering subset of train data by random selection of data representatives to get initial centroids for clustering on all training data to provide good data cluster. With better initial centroids from sampled data and improved algorithm by concurrent processing, our model will get better results in processing.

The research processes follow the model. The model includes the data set selection from KDD Cup website[21] . These data has complicated and varied type of categories suitable for data mining task to find some knowledge patterns. These data sets will be selected for interesting subjects related to the research's objective.

The objective of this research is to study features related to packets . All features selected from KDD Cup database will be filtered by association technique with Apriori-based algorithm to generate relevant features data sets.

The gained related feature data set will be clustered by K-means clustering technique and improved with the concurrent processing methodology. In k-means clustering, we will compare the result of clustering with clustering technique that got initial centroid from sampled data against the sequential data records without centroid selection technique.

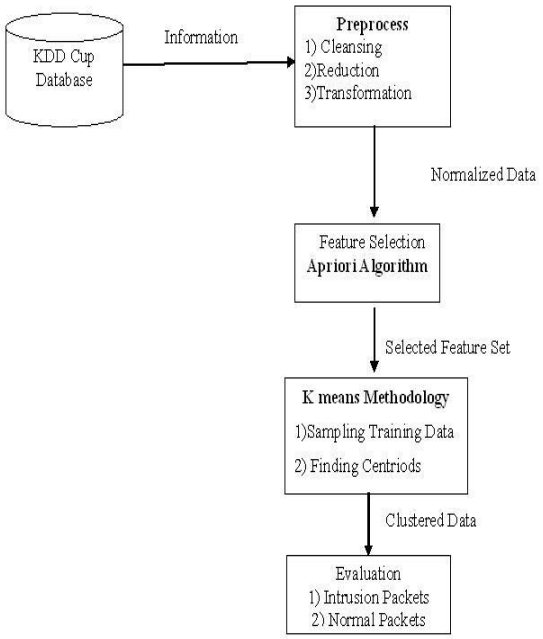

Fig 4.1: State Diagram For K means Methodology

Data sets and Implementations

1 Large Data Set Selection

This research uses large data sets from KDD cup database .The above state diagram i.e K-means Algorithm model that contain 6 features related to packets . The selected features are shown in table I.

TABLE I

SELECTED DATA FEATURES FROM KDD CUP

No.	Attribute	Description
1	PacketPort	Port of Packet
2	PacketBaseIP	Base IP of Packet
3	PacketIP	IP of Packet
4	PacketProtocol	Type of Protocol for Packet
5	PacketType	Type of Packet
6	PacketSize	Size of Packet

2 Data Preprocessing

The first process for data mining is data preprocessing that researchers will prepare the data sets for use in data mining processes. Researchers have to set clear criteria to filter all data sets suitable to the research objectives. The first step in data preprocessing is the data cleansing process that gets rid of noise and outlier. Then data has been reduced and transformed into the format that is appropriate for data mining software to analyze and clustering.

3 Feature Selection with Apriori-based Algorithm

In the continuing step of Processing model is feature selection processing. The feature selection will use the association technique with Apriori-based algorithm to generate the sets of feature relation rules. With Apriori-based algorithm used to analyze and generate features that are related and affect to other features in the group, more effective action in association technique is required. We have to filter the rules that appropriate to research objective. In this research our aim is to finding features that affect performance of packets. So that all features from KDD Cup selection will be used to calculate the associate rules with Apriori-based to get related features for clustering.

Criteria for *Feature Selection with Apriori-based Algorithm in K-means*

1) The size of minimum packet is 20 and maximum is 5000 above this would be treat as a false packets.
2) The minimum port for sending the packets is 200 below ports would be discarded.
3) If rejected protocol = “udp domain ” then it would be treat as a rejected protocol
4) If the given port = {246, 219, 324} then it would be a normal packets .
5) The number of iterations performed =4.
6) The maximum intrusion is 1000 above that not count.

4 Clustering by K-means Clustering Methodology

The main process in the model is clustering selected data with the k-means clustering method. We implement the k-means clustering algorithm with the Erlang programming language. Finding the initial centriod from the given packets first and then applying k means algorithms on same datasets. After doing these things we are getting these packets in two clusters like intrusion and normal packets.

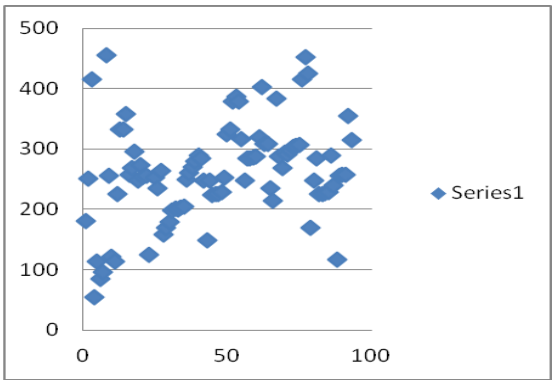

Fig 4.2 : Before Clustering

When by applying K means algorithm to the given files , the rejected protocols are automatically rejected from the files because of apriori algorithm , at the end got two clusters like intrusion and normal packets . Table No.1 shows the amount of data sets after filtering with all criteria.

Table number 1

File Size (kb)	35	40	53	41	55	54	51	118	142	209	124	146	205	149	234	158
Intrusion K	7	7	6	5	14	8	7	35	20	143	30	10	57	20	14	45
Normal	187	168	220	200	263	252	230	529	641	654	584	674	927	664	1118	643

Now, on the basis of these figures we plot the graph of intrusion and the normal packets of the given datasets.

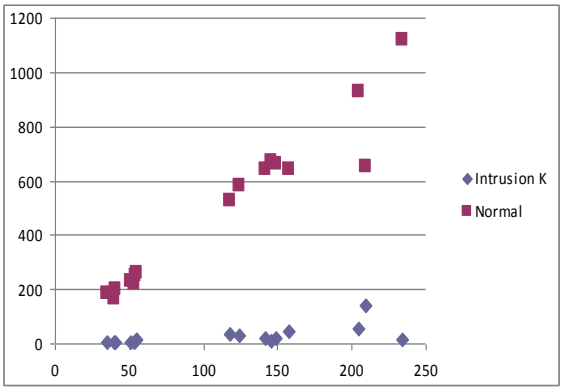

Fig 4.3 After applying K means Clustering

B. C-means Clustering Methodology

In **fuzzy clustering** (also referred to as **soft clustering**), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

It provides a Processing model that shows the processes of preparation the suitable data, selection the good centroids, and achieving better clustering performance with their processing. The processes include preprocessing steps that cleansing raw data from KDD Cup data and select suitable features by association analysis with Apriori-based algorithm. After that clustering subset of train data by random selection of data representatives to get initial center vectors for clustering on all training data to provide good data cluster. With better initial center vectors from sampled data and improved algorithm by concurrent processing, our model will get better results in and time processing.

The research processes follow the model. The model includes the data set selection from KDD Cup website . These data has complicated and varied type of categories suitable for data mining task to find some knowledge patterns. These data sets will be selected for interesting subjects related to the research's objective.

The gained related feature data set will be clustered by C-means clustering technique and improved with the concurrent processing methodology. In C-means clustering, we will compare the result of clustering with clustering technique that got initial center vectors from sampled data against the sequential data records without center vectors selection technique.

In the last process of model we will evaluate and compare results of clustering data. Clustering results are evaluated by comparing the intrusion and normal packets from clustering . All processes about C means are shown in figure .

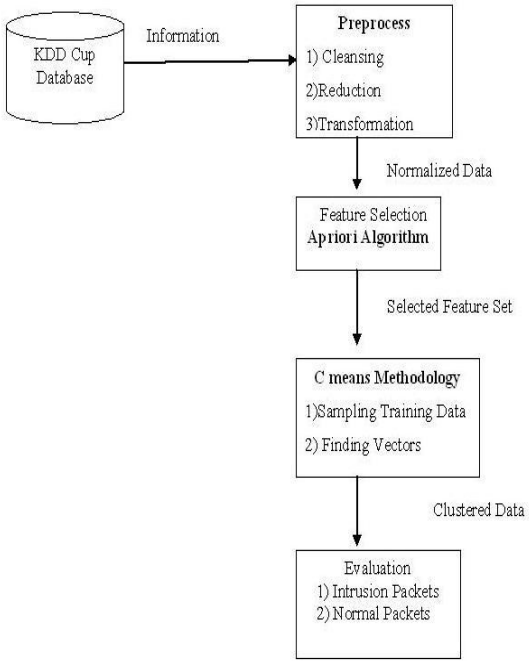

Fig 4.4 State Diagram for C means Methodology

Data sets and Implementations

1 Large Data Set Selection

This research uses large data sets from KDD cup database i.e the same example which is already used for K means. The above state diagram i.e C-means Algorithm model that contain 6 features related to packets . The selected features are shown in table II.

TABLE II

SELECTED DATA FEATURES FROM KDD CUP

No.	Attribute	Description
1	PacketPort	Port of Packet
2	PacketBaseIP	Base IP of Packet
3	PacketIP	IP of Packet
4	PacketProtocol	Type of Protocol for Packet
5	PacketType	Type of Packet
6	PacketSize	Size of Packet

2 Data Preprocessing

The first process for data mining is data preprocessing that researchers will prepare the data sets for use in data mining processes. Researchers have to set clear criteria to filter all data sets suitable to the research objectives. The first step in data preprocessing is the data cleansing process that gets rid of noise and outlier. Then data has been reduced and transformed into the format that is appropriate for data mining software to analyze and clustering

3 Feature Selection with Apriori-based Algorithm

In the continuing step of Processing model is feature selection processing. The feature selection will use the association technique with Apriori-based algorithm to generate the sets of feature relation rules. With Apriori-based algorithm used to analyze and generate features that are related and affect to other features in the group, more effective action in association technique is required. We have to filter the rules that appropriate to research objective. In this research our aim is to finding features that affect performance of packets. So all features from KDD Cup selection will be used to calculate the associate rules with Apriori-based to get related features for clustering.

Criteria for Feature Selection with Apriori-based Algorithm in C-means.

1) The size of minimum packet is 20 and maximum is 5000 above this would be treat as a false packets.

2) The minimum port for sending the packets is 200 below ports would be discarded.
3) If rejected protocol = “udp domain ” then it would be treat as a rejected protocol
4) If the given port = {246, 219, 324} then it would be a normal packets .
5) The number of iterations performed =4

4 Clustering by K-means Clustering Methodology

The main process in the model is clustering selected data with the C-means clustering method. We implement the C-means clustering algorithm with the Erlang programming language. Finding the initial center vectors from the given packets first and then applying C means on same datasets. After doing these things we are getting these packets in two clusters like intrusion and normal packets.

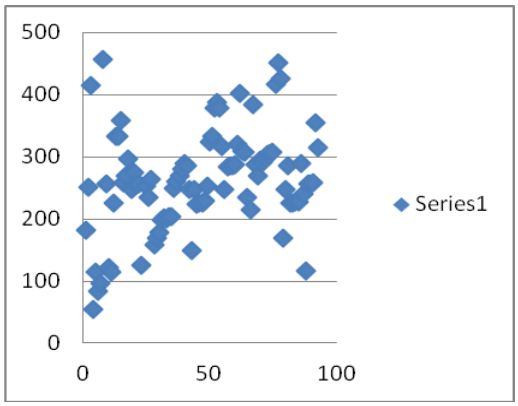

Fig 4.5 : Before Clustering

When by applying C means algorithm to the given files , the rejected protocols are automatically rejected from the files because of apriori algorithm , at the end got two clusters like intrusion and normal packets .

Table No. 2 shows the amount of data sets after filtering with all criteria.

File Size (kb)	35	40	53	41	55	54	51	118	142	209	124	146	205	149	234	158
Intrusion C	103	98	118	105	144	124	113	295	342	378	298	318	517	331	557	343
Normal	91	77	108	100	133	136	124	269	319	419	316	366	467	353	575	345

Table number 2

Now, on the basis of these figures we plot the graph of intrusion and the normal packets of the given datasets.

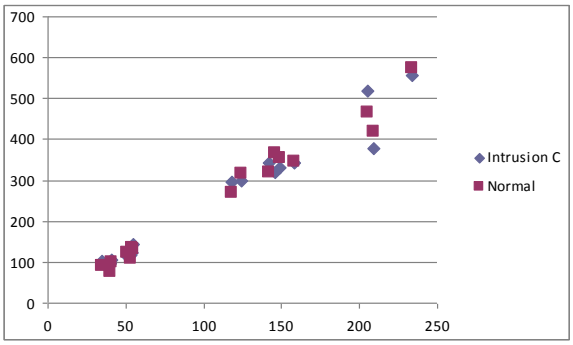

Fig 4.6 after applying C means Clustering

3. EXPERIMENTAL RESULTS:

The experimental results shows the difference between the K means and the C means clustering algorithms .If you refer the below table (table number 3) knows the differences between the working of both the clustering algorithms.

File Size (KB)	**Intrusion K**	**Normal (by using K means)**	**Intrusion C**	**Normal (by using C means)**
35	7	187	103	91
40	7	168	98	77
53	6	220	118	108
41	5	200	105	100
55	14	263	144	133
54	8	252	124	136
51	7	230	113	124
118	35	529	295	269
142	20	641	342	319
209	143	654	378	419
124	30	584	298	316
146	10	674	318	366
205	57	927	517	467
149	20	664	331	353
234	14	1118	557	575
158	45	643	343	345

Table number 3

Now, on the basis of these figures we plot the graph of intrusion and the normal packets of k means and c means together from the given datasets.

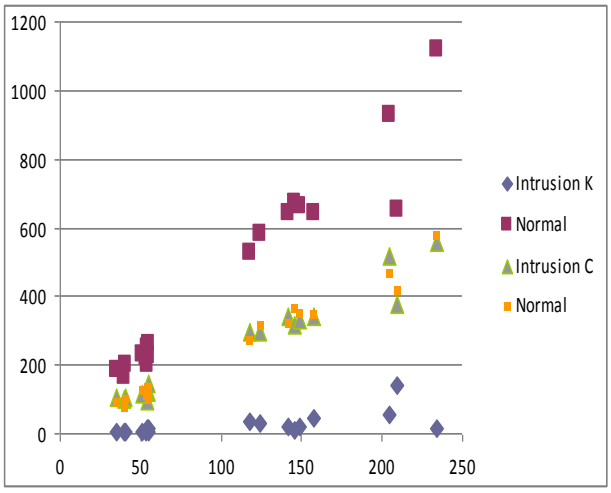

Fig 5.1 K-means Vs C-means

The experimental results show the difference between the K means and the C means clustering algorithms. With the help of these differences, knows the process of both.

4. CONCLUSION

In this algorithm I tried to improve the efficiency of K means and the C means clustering methodology with apriori algorithm. The data mining has many techniques available for users to apply suitable data types and usage. From this research we present one of unsupervised data mining technique called data clustering that integrated other mining technique and concurrent processing. The preprocessing added the association rules obtained from Apriori-based algorithm in feature selection to get better feature set. In the clustering process we used random data to generate initial centroids and vectors that works better for data sets. The performance can be measure on the number of intrusion packets in a less time. The experimental results show the difference between the K means and the C means clustering algorithms .

5.REFERENCES

[1] Han J. and Kamber M.(2001). Data Mining : concept and techniques CA: Academic Press.
[2] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", in: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 1994, pp. 487–499.
[3]Hung,Ming-Chuan Dept. of Inf. Eng., Feng Chia Univ.,Taichung,Taiwan Yang, Don-Lin L. on "An efficient Fuzzy C-Means clustering algorithm " Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference , 2001
[4] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE VOL. 24, NO. 7, JULY 2002
[5] Tan , Steinbach and Kumar , on "Introduction to Data Mining" Apr 2004.
[6]Data mining concept A.K. Pujari.
[7] Cebron, Nicolas Dept. of Comput. & Inf. Sci., Konstanz Univ. Berthold,MichaelR. on "Adaptive Fuzzy Clustering " Fuzzy Information Processing Society, 2006. NAFIPS 2006, 3-6 June 2006
[8]Wang,Weina Dept. of Appl. Math., Dalian Maritime Univ. Zhang, Yun-Jie; Li, Yi; Zhang,Xiaona on "The Global Fuzzy C-Means Clustering Algorithm Intelligent Control and Automation, 2006. WCICA 2006,
[9]Tan,Steinbach and Ghosh Kumar ,on " Top Ten Data Mining Algorithms" Dec 2006.
[10] Lawrence K.D., Kudyba S. and Klimberg R.K., Data mining methods and applications. USA: Auerbach Publications, 2008, pp. 83-104.
[11] Rahman H., Data mining applications for empowering knowledge societies. USA: Information Science Reference, 2009, pp. 43-54.
[12] Taniar D., Data mining and knowledge discovery technologies. USA: IGI Pub, 2008, pp. 118-142.
[13]Chen Zhang Sch. of Comput. Sci. & Technol., China Univ. of Min. & Technol., Xuzhou Shixiong Xia , on "K-means Clustering Algorithm with Improved Initial Center" Knowledge Discovery and Data Mining, 2009. WKDD 2009. 23-25 Jan. 2009
[14]Wang J., Data warehousing and mining : concepts, methodologies, tools, and applications. USA: Information Science Reference, 2009, pp. 303-335.
[15] Wu X. and Kumar V., The top ten algorithms in data mining. USA: CRC Press, 2009, p. 21, p. 93.
[16].,Lu,Lin Sch. of Inf. & Control Eng., Xi"an Univ. of Archit. & Technol Xi"an, China
Liu, Pei-qi , on "Stduy on an improved apriori algorithm and its application in supermarket "Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference , 23-25 June 2010
[17]Pradeep Rai & Shubha Singh," A Survey of Clustering Techniques",IJCA, Volume 7– No.12, October 2010
[18] S. Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama on "A Survey on Partition Clustering Algorithms" IJECBS Vol. 1 Issue 1 January 2011
[19] Noppol Thangsupachai, Phichayasini Kitwatthanathawon, Supachanun Wanapu, and Nittaya Kerdprasop , on " Clustering Large Datasets with Apriori-based Algorithm and Concurrent Processing" ,IMECS 2011, March 16-18 , 2011,Hong Kong.
[20]www.faculty.uscupstate.edu/atzacheva/SHIM450/KMeansExample.doc
[21]http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html