Comparison of Clustering Algorithm in Automated Inventory System

Akshay A. Bandivadekar , Harshal D. Gadhia ,V.Mukhilan,

Abstract— **Patterns and classification of stock or inventory data is very important for decision making and business support. In proposed system an algorithm for mining patterns of huge stock data to predict factors affecting the sale of products, Identification of sales patterns from inventory data indicate the market trends which can further be used for forecasting, decision making and strategic planning. The objective is to get better decision making for improving sales, services and quality as to identify the reasons for dead stock, slow moving and fast moving stock. The system proposes two phases in which first phase includes initial clustering which is performed on the database with the help of a clustering algorithm. In the second phase the system uses most frequent pattern, MFP algorithm to find the frequencies of property values of the items. The existing system uses k-means clustering algorithm along with MFP for mining patterns. In order to improve the execution time the proposed system uses efficient methods for clustering which includes Partitioning Around Medoids, PAM and Balanced Iterative Reducing and Clustering using Hierarchies BIRCH along with MFP.The most efficient iterative clustering approach called as PAM is used for initial clustering and is then combined with frequent pattern mining algorithm. In order to meet the memory requirements, an incremental clustering algorithm BIRCH is also used for mining frequent patterns. So, the evaluation of these clustering algorithms along with MFP is made with respect to the execution times.**

Keywords: Data mining, Clustering, K-means, PAM, MFP, computational complexity

I. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics

Association algorithms are designed to operate on databases containing transactions. As is common in association rule mining, given a set of item sets, the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

The relationship among the large amount of biological data has become a hot research topic. It is desirable to have clustering methods to group similar data together so that, when a lot of data is needed, all data are easily found in close proximity to some search result.

Clustering techniques have a wide use and importance nowadays. This importance tends to increase as the amount of data grows and the processing power of the computers increases. Clustering applications are used extensively in various fields such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing.

The main purpose of clustering techniques is to partitionate a set of entities into different groups, called clusters. These groups may be consistent in terms of similarity of its members. As the name suggests, the representative-based clustering techniques uses some form of representation for each cluster. Thus, every group has a member that represents it. The motivation to use such clustering techniques is the fact that, besides reducing the cost of the algorithm, the use of representatives makes the process easier to understand. There are many decisions that have to be made in order to use the strategy of representative-based clustering. For example, there is an obvious trade-off between the number of clusters and the internal cohesion of them. If there are few clusters, the internal cohesion tends to be small. Otherwise, a large number of clusters makes them very close, so that there is little difference between adjacent groups. Another decision is whether the clusters should be mutually exclusive or not, that is, if an entity can co-exist in more than one cluster at the same time.

II. BACKGROUND AND RELATED WORK

For the transportation industry hugely contributes to the economy of India.For transportation purpose, the development and management of highway is a must. So tries to cluster research objects (namely chosen hub city) using clustering analysis , and then sort them according to some rules in order to make sure different layers of highway transportation cities sets[1]. Data is very important for every organization and business. Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables. The algorithms used are K-means, MFP(Most Frequent Pattern), K-Medoid, Birch

algorithm [13].To formulates, simulates and assess an improved data clustering algorithm for mining web documents with a view to preserving their conceptual similarities and eliminating problem of speed while increasing accuracy. The proposed algorithm was simulated using the fuzzy logic and statistical toolbox in Matlab7.0 [4] . To represent a data mining approach for inventory forecasting and planning a Bill of Materials in a highly competitive environment such as an Italian car racing team. By exploiting clustering algorithms and by using statistical techniques to identify the optimal number of clusters this work presents a method to optimally cluster a multi-year dataset containing the products used in car revision after each rally competition during a three year period [9] .

Stock market produces huge datasets that deals enormously complex and dynamic problems with data mining tool. Data mining is the emerging methodology used in stock market, finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decision. [12] .To obtained the frequent patterns from the stock data. Hybrid clustering association mining approach is proposed to classify stock data and find compact form of associated patterns of sale . From the experimental results it is clear that proposed approach is very efficient for mining patterns of stock data with less computational time than the proposed approach .By these patterns we may predict the factors affecting the sales. In future we may try to implement the same process in document classification and may even try to have better computational efficiency by efficient algorithms [17] .

III. EXISTING STSTEM

In paper [17], an algorithm is used for mining patterns of huge stock data to predict factors affecting the sale of products**.** To achieve these goals, we need to fully exploit this data by extracting all the useful information from it.

The algorithms used are –1. K-Mean algorithm.
2. K- Medoids algorithm (PAM)
3. Birch Algorithm
4. Most Frequent Pattern (MFP)

In this paper, for clustering K-Mean algorithm, K- Medoids algorithm (PAM), Birch Algorithm are used and Association is done by using Most Frequent Pattern (MFP) algorithm.

In this paper, comparison done by using Number of Iteration

This paper shows that Birch with MFP algorithm shows best result as compared to K-Mean with MFP algorithm and K-Medoids with MFP algorithm.

IV. PROPOSED SYSTEM

Only through data mining it is possible to extract useful pattern and association from the stock data.

Algorithm are used:-

1)Clustering Algorithm:
a) K-Mean algorithm.
b) K- Medoids algorithm (PAM)
c) Birch Algorithm
2)Association Rule Mining Algorithm:
a) Most Frequent Pattern (MFP) algorithm.
b) Apriori Algorithm

Comparison will be done by using two parameters Number of Iterations and size of dataset handled by each algorithm.

Our aim is to find out which algorithm is best and time efficient among clustering algorithms and which algorithm works best with which association algorithm.

V. METHODOLOGY

1)Clustering Algorithm:
A) K-Mean Algorithm
- 3 categories
- Dead-Stock (DS).
- Slow-Moving (SM).
- Fast-Moving (FM) .

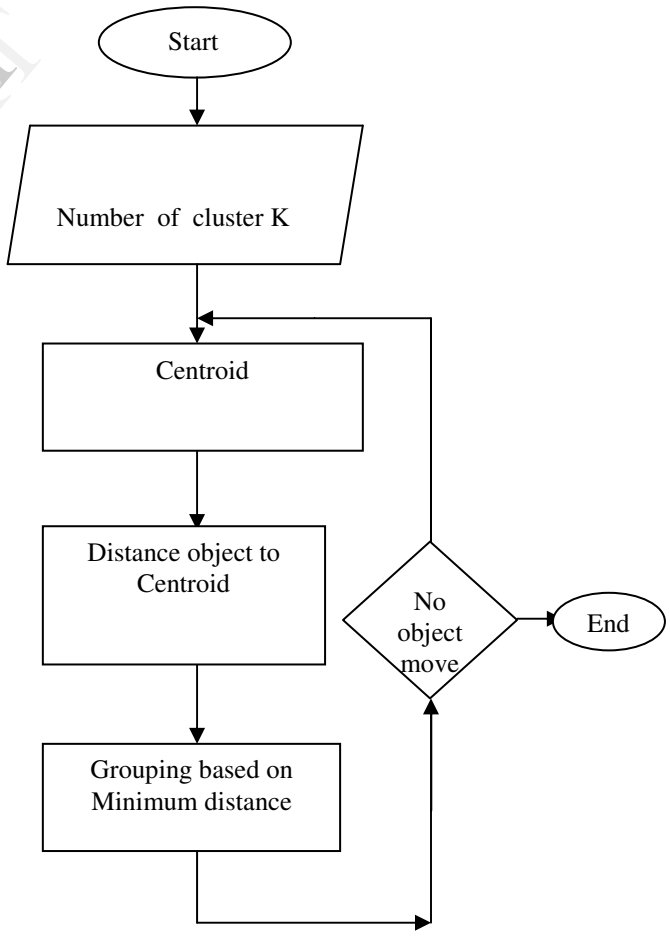


Fig.K-Mean Algorithm

Steps for the K-Mean algorithm

- Step 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Step 2: Assign each object to the group that has the closest centroid.
- Step 3: When all objects have been assigned, recalculate the positions of the K centroids.
- Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

2) The PAM Clustering Algorithm

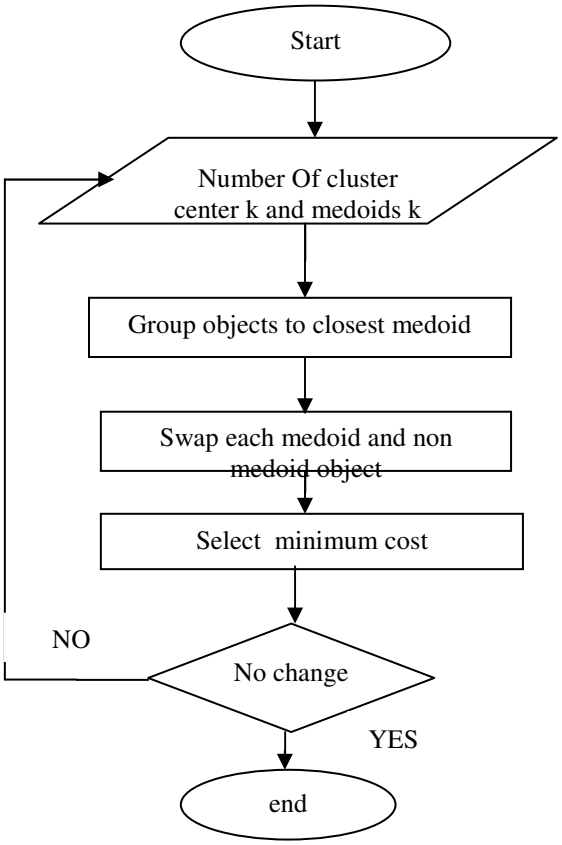


Figure 1: Flowchart of PAM algorithm

PAM (Detail Algorithm given by Margaret H.Dunham):-
Arbitrarily select k medoids from D;
Repeat
for each th not a medoid do
for each medoidt,do
calculateTCih;
Find i,h where TCih is the smallest;
If TCih<0 then
replacemedoidti with th;
Until Tcih<=0;
for each ti belonging to D do
assignti to kj where dis(ti,tj) is smallest over all medoids

3)BIRCH: An Efficient Data Clustering Method for Very Large Database
Introduction:
Definition of Data clustering
Given the desired number of clusters K and a distance-based measurement function, we are asked to find a partition of the dataset that minimizes the value of the measurement function.database-oriented constraint:
The amount of memory available is limited and we want to minimize the time required for I/O.

Insertion into a CF Tree
We now present the algorithm for inserting an entry into a CF tree.Given entry "Ent", it proceeds as below:

1. Identifying the appropriate leaf: Starting from the root, according to a chosen distance metric D0 to D4 as defined before, it recursively descends the CF tree by choosing the closest child node .
2. Modifying the leaf: When it reaches a leaf node, it finds the closest leaf entry, and tests whether the node can absorb it without violating the threshold condition.

- If so, the CF vector for the node is updated to reflect this.
- If not, a new entry for it is added to the leaf.
- If there is space on the leaf for this new entry, we are done.

3. Modifying the path to the leaf: After inserting "Ent" into a leaf, we must update the CF information for each nonleaf entry on the path to the leaf.

- In the absence of a split, this simply involves adding CF vectors to reflect the additions of "Ent".
- A leaf split requires us to insert a new nonleaf entry into the parent node, to describe the newly created leaf.
- If the parent has space for this entry, at all higher levels, we only need to update the CF vectors to reflect the addition of "Ent".
- Otherwise, we may have to split the parent as well, and so on up to the root.

4. Merging Refinement: In the presence of skewed data input order, split can affect the clustering quality, and also reduce space utilization. A simple additional merging often helps ameliorate these problems: suppose the propagation of one split stops at some nonleaf node N_j, i.e., N_j can accommodate the additional entry resulting from the split.

- Scan N_j to find the two closest entries.
- If they are not the pair corresponding to the split, merge them.
- If there are more entries than one page can hold, split it again.
- During the resplitting, one of the seeds attracts enough merged entries, the other receives the rest entries.

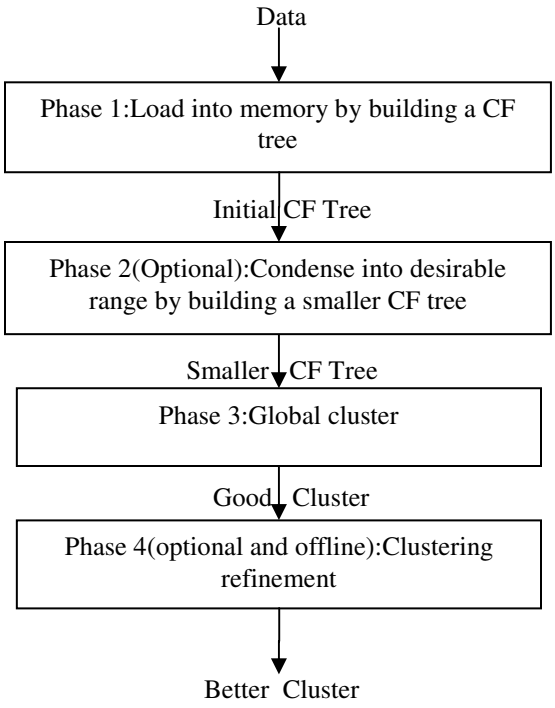


Figure:Flowchart of Birch Algorithm

BIRCH Clustering Algorithm – Phase 2,3 & 4

- Phase 2: Condense into desirable range by building a smaller CF tree
- To meet the need of the input size range of Phase 3
- Phase 3: Global clustering
- To solve the problem 1
- Approach: re-cluster all subclusters by using the existing global or semi-global clustering algorithms
- Phase 4: Global clustering
- To solve the problem 2
- Approach: use the centroids of the cluster produced by phase 3 as seeds, and redistributes the data points to its closest seed to obtain a set of new clusters. This can use existing algorithm

BIRCH Clustering Phase1 – Threshold Value

Heuristic approach to increase the threshold

Try to choose new threshold value so that the number of data points that will be scanned under the new threshold value can double .

Approach 1: find the most crowded leaf node and the closest two entries on the leaf can be merged under new threshold.

Approach 2: Assuming that the volume occupied by the leaf clusters grows linearly with data points. a series of value pair: number of data point and volume ⇒ new volume (a new data point, using least squares linear regression) ⇒ new threshold

Using some heuristic methods to adjust the above two thresholds and choose one.

BIRCH Clustering Phase1 –Delay-Split option

When we run out of memory. There may be more data points that can fit in the current CF tree. We can continue to read data point and write those data points that require to split a node to disk until the disk space is run out. The advantage of this approach is that more data points can fit in the tree before we have to rebuild.

Performance Studies

- Complexity Analysis
- Experiment with Synthetic Datasets
- Performance Comparisons of BIRCH and CLARANS with Synthetic Datasets
- Experiment with Real Datasets

2)Association Algorithm:

A) MOST FREQUENT PATTERN MINING (MFPM) ALGORITHM

Association rule mining is one of the mostimportant and well defines technique for frequent pattern,associations in a dataset. Association rules [7] are widelyused in various areas such as market analysis, inventory[4] control, and stock data. Apriori algorithm for strongassociation among the patterns is highly recommended. Inthis work we proposed a new algorithm MFP thatefficiently generates frequent patterns and strongassociation between them Let we have set X of N items ina Dataset having set Y of attributes. This algorithm countsmaximum of each attribute values for each item in thedataset.

The algorithm is as follows
Input: Datasets (DS)
*Output:*Matrix Most Frequent Pattern (MFP):
MFP (DS)
Begin
For each item Xi in DS
a. for each attribute
i. count occurrences for Xi
C=Count (Xi)
ii. Find attribute name of C
having maximum count
Mi=Attribute (Ci)
Next [End of inner loop]
b. Find Most Frequent Pattern
i. MFP=Combine (Mi)
Next [End of outer loop

B) APRIORI ALGORITHM

The candidate-gen function takes *Fk*-1 and returns a superset (called the candidates) of the set of all frequent *k*-itemsets. It has two steps

- join step: Generate all possible candidate itemsets*Ck* of length *k*
- prune step: Remove those candidates in *Ck* that cannot be frequent.

Candidate –gen function

Function candidate-gen(*Fk*-1)
Ck¬Æ;
forall*f*1, *f*2 ÎFk-1
with*f*1 = {*i*1, … , *ik*-2, *ik*-1}
and*f*2 = {*i*1, … , *ik*-2, *i'k*-1}
and*ik*-1 <*i'k*-1 **do**
c¬ {*i*1, …, *ik*-1, *i'k*-1}; // join *f*1 and *f*2
Ck¬*Ck*È {*c*};
foreach (*k*-1)-subset *s* of *c***do**
if(*s*ÏFk-1) **then**
delete*c* from *Ck*; // prune
end
end
return*Ck*;

VI. EXPERIMENTS AND RESULTS

The data set we used includes data objects each one with seven attributes. Basing on one of the attributes the clustering technique is applied. The attribute for clustering [9] is decided accordingly. In our application of stock data [7], clustering is performed based on the attribute of quantity sold. The sample data is shown below.

TABLE I. DATA SET

Item_id	Color	Season	Size	Company	Gender	Qty_sold
2.0	green	s	5.0	jack	f	0.0
8.0	red	s	6.0	jack	m	0.0
9.0	green	s	7.0	wikkey	m	0.0
10.0	white	w	1.0	kips	m	0.0
19.0	black	w	6.0	jack	m	5.0
20.0	white	s	7.0	wikkey	m	5.0
27.0	white	s	2.0	wikkey	f	9.0
28.0	red	s	4.0	jack	f	9.0
2.0	green	s	5.0	jack	f	9.0
10.0	white	w	1.0	kips	m	12.0
20.0	white	s	7.0	wikkey	m	12.0
28.0	red	s	4.0	jack	f	12.0
3.0	white	w	3.0	wikkey	m	12.0
11.0	green	s	2.0	jack	m	18.0
16.0	green	s	9.0	wikkey	m	18.0
15.0	red	w	10.0	wikkey	m	18.0
22.0	red	s	6.0	kips	f	25.0
4.0	green	s	2.0	kips	f	25.0
12.0	black	s	4.0	imp	f	30.0
17.0	white	w	8.0	imp	m	30.0
23.0	green	w	7.0	jack	f	30.0
5.0	black	a	3.0	wikkey	m	30.0
18.0	green	s	7.0	kips	m	30.0

We apply the clustering technique for the initial grouping of the whole data it then give three clusters of DS, SM and FM stock. So the process has two phases as given below

The evaluation of K means, PAM and BIRCH is done and the execution time is tabulated

A. *Interaction Effects Study*

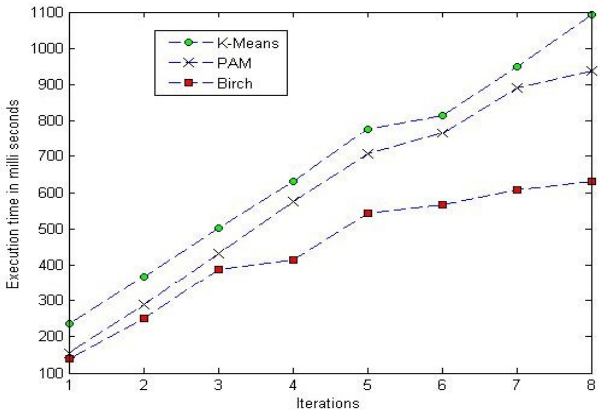
The interaction effects study gives us the interaction between Execution time and the iterations. We take the interaction plot between the execution time and iterations. The following can be observed.

• When K-means is considered there is a increase in the execution time with respect to the iterations
• When K-Medoids [15] is compared to k means there is decrease in the execution time which may improve the overall computational efficiency.
When Birch is compared to these algorithms there has been much improvement in the execution time.

Based on these effects the results are specified in a tabular form showing the iterations and the execution time in milliseconds for clustering algorithms k-means, PAM and BIRCH in combination with MFP.A graph is then plotted basing on these tabulated values.

TABLE II. RESULTS OF ALGORITHMS

Iterations	MFP with K-Means	MFP with PAM	MFP with BIRCH
1	235	155	140
2	368	288	250
3	503	431	388
4	631	576	413
5	777	707	543
6	816	764	567
7	950	890	607
8	1094	937	630



VII. CONCLUSION AND FUTURE WORK

By comparing K-means with MFP, PAM with MFP, Birch with MFP, K-means with Apriori, PAM with Apriori, Birch with Apriori where K-means, PAM and Birch are Clustering algorithms while Apriori and MFP are Association rule mining algorithms. We will achieve that which clustering algorithm is best and time efficient and which association algorithm is best and time efficient. Here the comparison parameter are - 1)Number Of Iteration 2)Size of dataset

REFERENCES

[1] Yan Meng ,Xiyu Liu . "Application of K-means Algorithm Based on Ant Clustering Algorithm in Macroscopic Planning of Highway Transportation Hub ",published in ©2007 IEEE 1-4244-1385-0/07.

[2] Chen-Chia Chuang ,Jin-TsongJeng ,Chih-Wen Li . "Fuzzy C-Means Clustering Algorithm with Unknown Number of Clusters for Symbolic Interval Data ",SICE Annual Conference 2008 .August 20-22, 2008, The University Electro-Communications, Japan .

[3] Sheng-Yi Jiang , Xia Li . "A Hybrid Clustering Algorithm ",2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

[4] Odukoya, O.H, Aderounmu G.A. And Adagunedo, E.R. "An Improved Data Clustering Algorithm for Mining Web Documents",published in ©2010 IEEE 978-1-4244-5392-4/10/

[5] FuhengQu ,Yating Hu ,Shuangzi Sun ,"A New Possibilistic Clustering Algorithm with Its Application to Fault Diagnosis ",published in ©2011 IEEE 978-1-4244-9857-4/11 /

[6] WangChun-hong,Nan Li-li,Ren Yao-Peng,"Research on th Text Clustering Algorithm based on Latent Semantic Analysis and optimization",published in ©2011 IEEE 978-1-4244-8728-8/11 /

[7] Huiying Wang ,Xiangwei Liu ,"Study on Frequent Term Set-based Hierarchical Clustering Algorithm ,2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) .

[8] Li Liu ,Yun Shao ,Fengli Zhang ,Xu Lu ,"The Discrepancies Caused By Different Cluster Merging Algorithms In Fully Polarimetric SAR Classification",published in ©2012 IEEE 978-1-4673-1159-5/12/

[9] Francesco Maiorana ,Angelo Mongioj ,"A data mining approach for bill of materials for motor revision ",published in ©2012 IEEE 978-83-60810-48-4/

[10] NeethiNarayanan ,J.E.Judith, Dr.J.Jayakumari ,"Enhanced Distributed Document Clustering Algorithm Using Different Similarity Measures ",Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[11] XuDegang, Zhao Panlei, GuiWeihua, Yang Chunhua, XieYongfang ,"Research on Spectral Clustering Algorithms Based on Building Different Affinity Matrix",published in ©2010 IEEE 978-1-4673-5534-6/13/

[12] Sachinkambrey, R. S. Thakur,ShaileshJalori,"Application of Data Mining Technique in Stock Market : An Analysis", International Journal of Computer & Communication Technology (IJCCT) ISSN (Online): 2231 - 0371 ISSN (Print): 0975 –7449 Vol-3, Iss-3, 2012

[13] M.Rajeswari ,Y.Ramu,"Frequent Patterns Mining Of Stock Data Using Hybrid Clustering Association ",International Journal of Systems,Algortihm and application Volume 1, Issue 1, December 2011 .

[14] Aurangzeb Khan, BaharumBaharudin, Khairullah Khan," Mining Customer Data For Decision Making Using New Hybrid Classification Algorithm,publication of Little LionScientificR&D,IslamabadPakistan,Journal of Theroteicaland applied Information Technology,15th May 2011.Vol.27 No.1.

[15] Sunil Joshi ,Dr. R. S . Jadon ,Dr. R. C. Jain ,"An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function ",International Journal of Computer Applications (0975 – 8887) Volume 9– No.9, November 2010

[16] M.SureshBabu ,Dr. N.Geethanjali, Prof B.Satyanarayana ,"Clustering Approach to Stock Market Prediction ",Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291 (2012) .

[17] D.V.S. Shalini ,M.Shashi, A.M.Sowjanya , "Mining Frequent Patterns of Stock Data UsingHybrid Clustering (2010) .