International Research Journal of Computer Science (IRJCS)
Volume 1 Issue 1 (September 2014)
ISSN: Applied
www.irjcs.com

Enhancing Pattern Recognition in Social Networking Dataset by Using Bisecting KMean with Accuracy Rates

Ms. Shilpa V. Gajbhiye
M.Tech., Dept. of Computer Engineering
Bapurao Deshmukh College of Engineering
Sevagram, Maharashtra, India

Prof. Sudhir W. Mohod
Dept. of Computer Engineering
Bapurao Deshmukh College of Engineering
Sevagram, Maharashtra, India

Abstract - Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The research uses social networking data set for pattern recognition, because it is one of the emerging application areas in data mining. We are using the Facebook 100 dataset and applying the Bisecting KMean algorithm on it, by which we would get better clustering results. Bisecting KMean first bisects the data into 2 parts and selects the part with greater number of elements, then applies clustering on it again. This goes until we have N Number of clusters. We would apply this to our dataset to get desired results. We have calculated the accuracy by using Rand Index for KMean and Bisecting KMean. With this we are going to compare Bisecting KMean algorithm with other data mining algorithm and we are going to find out different pattern from social networking dataset. The patterns are depending on Male Students, Female Students, Male Faculty, Female Faculty, Major and Minor. Graphical representation is also provided with accuracy rates of various algorithms.

***Index Terms*—Social Networking Site, Bisecting KMean, KMean, Cluster, Pattern,Rand Index.**

I. INTRODUCTION

Innovative organizations worldwide are already using data mining to locate and appeal to higher value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. The first and simplest analytical step in data mining is to describe the data summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). As emphasized in the section on the data mining process [1], collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. We have to build a predictive model based on patterns determined from known results, and then test that model on results outside the original sample. A good model should never be confused with reality (for example a road map isn't a perfect representation of the actual road), but it can be a useful guide to understanding our business. The final step is to empirically verify the model. But Data mining is a tool not a magic stick. It won't sit on database watching what happens and sends e-mail to get your attention when it sees an interesting pattern. It doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining assists business analysts with finding patterns and relationships in the data it does not tell you the value of the patterns to the organization. The data set on which the data mining is going to applied plays very important role. Social networks have become omnipresent in today's life [2].Many Social Network sites (SNS) are now available like Orkut, Face Book, and Twitter. Face book is a social networking service and website which was launched in February 2004. As of February 2012, Face book has more than 845 million active users. This Research uses Face book 100 university dataset which defines various attributes like ID, Student/Faculty flag, Gender, Major, Second Major, Dorm /house/ Year and High school of 100 Universities. Our work focuses on Mining Association Patterns in only a subset of 100 Universities by randomly choosing some universities out of 100 and projects the association between a different attributes. The work also concentrates to evaluate the performance of clustering algorithms on the universities based on the accuracy in grouping the data in specific way. This research work focuses on extracting patterns from the Face book 100 universities and projects some specific association rules when applied to the dataset.

To enhance the pattern discovery on the data set the clustering is very important task. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups(clusters) [3]. 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 informatics. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular

© 2014, IRJCS- All Rights Reserved

statistical distributions. Clustering can therefore be formulated as a multi objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

II. Related Work

Zhao Yongli (Zhao et.al, 2013) describes an improved feature selection algorithm to identify most appropriate subset of features for a certain attack in a network. The method proposed by them is based on MAHALANOBIS Distance feature ranking and an improved exhaustive search for choosing a better combination of features [4]. They evaluate the approach on the KDD CUP 1999 datasets using SVM classifier and KNN classifier. They proved that classification can be done with high classification rate and low misclassification rate with reduced feature subsets.

Letao Qi, Harris (Letao et.al, 2013) implemented two representative clustering algorithms using update queries against the SPARQL endpoint of the RDF store [5]. They compare the time complexity and the communication complexity of algorithms with of those that require direct centralized access to the data and hence have to retrieve the entire RDF dataset from the remote location. They used Flicker dataset for their work.

David Combe (David et.al 2012) presents different combined clustering methods and evaluates their performances. The dataset used by them contains a scientific social network in which textual data is associated to each vertex and the classes are known [6].They also showed that good clustering results can be obtained using simple methods, when having a scenario adapted to the data and having precise criteria characterizing a good cluster.

Zhiwen Hu(Zhiwen et.al,2012) proposed a new algorithm called Community Detection algorithm for mining interesting communities or groups in a Campus Mobile Social Network (CMSN) [7]. The algorithm composed of two main components, one for community partition and other for selecting small communities to combine into a big community. They show that performance of their algorithm is better than the state-of-the-art Newman Clustering algorithm for mining community in CMSN.

Jia-Yi Li (Jai et.al, 2012) applied non negative matrix factorization algorithm and visualization method to data collected from online and real-life social networks, and discover the link patterns of web based and non web based social networks among a certain group of students [8]. They compared the networking patterns, and proved the existence of behavior unconformity and show how behavior unconformity might strengthen the ties between the individuals.

Joseph J. Pfeiffer (Joseph et.al 2012) proposed an extension to the Chung Lu random graph model, the Transitive Chung Lu (TCL) model, which incorporates the notion transitive edges. They combined the standard Chung Lu model with edges that are formed through transitive closure (e.g., by connecting a 'friend of a friend'). They prove TCL's expected degree distribution is equal to the degree distribution of the original input graph, while still providing the ability to capture the clustering in the network [9]. They demonstrate the performance of TCL on four real world social networks, including an email dataset with hundreds of thousands of nodes and millions of edges, showing TCL generates graphs that match the degree distribution, clustering coefficients and hop plots of the original networks.

Cheng-T Li (Cheng et.al 2012) presents a novel framework for knowledge discovery in heterogeneous social networks. They proposed a tensor-based model with operations about relation sequences to catch the direct and indirect information for nodes. Based on the devised model, three brand new centrality measures for heterogeneous social networks are proposed. They also propose a role based clustering schema to group nodes based on their relational semantics [10]. Their outcomes on both real and artificial dataset not only demonstrate the usability of their framework but also show the tool can assist human analyst in making more accurate, efficient, and confident decisions.

Amanda L (Amanda et.al, 2011) studied the social structure of Facebook "friendship" networks at one hundred American colleges and universities at a single point in time, and examined the roles of user attributes gender, class year, major, high school, and residence at these institutions [11]. They investigate the influence of common attributes at the dyad level in terms of associativity coefficients and regression models. Then examine larger-scale groupings by detecting communities algorithmically and comparing them to network partitions based on the user characteristics. They compare the relative importance's of different characteristics at different institutions, finding for example that common high school is more important to the social organization of large institutions and that the importance of common major varies significantly between institutions.

Krzysztof Juszczyszyn(Krzysztof et.al,2011) presents a new approach to the description and quantifying evolutionary patterns of social networks illustrated with the data from the Enron email dataset. They have propose the discovery of local network connection patterns (in this case: triads of nodes), measuring their transitions during network evolution and present the preliminary results of this approach. The Triad Transition Matrix (TTM) containing the probabilities of transitions between triads, then the result show how it can help to discover the dynamic patterns of network evolution [12]. Also, they analyze the roles performed by different triads in the network evolution by the creation of triad transition graph built from the TTM, which allows them to characterize the tendencies of structural changes in the investigated network.

Zhu Wang (Zhu et.al, 2012) proposed work was based on the user-venue check-in relationship and user/venue attributes. They come out with a novel community profiling framework. Specifically, they first adopt edge-clustering to simultaneously group both users and venues into communities, and then based on the rich metadata of users and venues we put forward a quantitative community profiling mechanism to indicate the preferences, interests and habits of a community. The efficiency of their approach is validated by intensive empirical evaluations using the collected foursquare dataset of 266,838 users with 9,803,764 check-ins over 2,477,122 venues worldwide [13].

© 2014, IRJCS- All Rights Reserved

R.Chithra and S.Nickolas (2010) used a novel algorithm for generating hybrid dimensional association rules. By providing appropriate data structure, with four level linked structures is used for this algorithm. Many datasets consists of one or more multivalued attributes [14]. The strength of the algorithm is, to store the transaction numbers along with 1-itemset to avoid multiple scan of the dataset. This structure need not compare item sets straightway; instead it checks with attribute combination whether to proceed with inter dimensional join or intra dimensional join. They reduced the comparison time to find relevancy among different values of different attributes. They applied algorithm for different datasets, with multiple values, and performance is evaluated.

III. MOTIVATION

There are many issues which came across while survey. Studies were done on practical databases and as practical databases are omnipresent, it slows down the performance. For such type of practical databases KMean algorithm is generally used. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Kmeans algorithm is fully deterministic, once initial centroid is selected. In this initial centroid plays a very important role, bad choice of initial centroid leads to poor cluster which may lead to poor clustering output. When considering association rule generation frequent item set or pattern discovery and searching interest in that is very important. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal [18] introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. When social network is considered the shaping of a network is a complex process and there are many factors and reasons that lead to the formation and breaking up of connections. And identification of central node is also important. From these issues we motivate to use the large document collection, which may be used in many applications like digital libraries or web. There is additional interest in methods for more effective management of information, like Abstraction a process by which data and programs are defined with a representation similar in form to its meaning, while hiding away the implementation details, Browsing which supposed to identify something of relevance for the browsing organism, Classification which may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood, Retrieval an activity of obtaining information resources relevant to an information need from a collection of information resources. Again clustering is the means for achieving better organization of information. In this the data space is partitioned into groups of entities with similar content.

IV. PREVIOUS WORK DONE

In our previous work, We have downloaded the FACEBOOK 100 dataset. As the dataset was not readable, extraction of the data from the dataset is done. The fields of the dataset are Student/ Faculty, Gender, Major, Second Major, Dorm, Year, High School. And the data was displayed in excel format. After extraction, discretization of data was done.

```
       Student Female  00251   000000000243    0094    2007    00000003701
       Student Female  00251   000000000000    0000    2006    00000000000
       Student   Male  00242   000000000256    0000    2006    00000008823
Student/Faculty Gender  Major   Second Major    Dorm    Year    High School
       Student   Male  00000   000000000000    0088    2008    00000017537
       Student Female  00283   000000000273    0000    2006    00000002277
       Student Female  00292   000000000267    0000    0000    00000000000
       Student Female  00000   000000000000    0086    2009    00000018158
```

Then particular dataset was selected and on that dataset the KMean algorithm is implemented.
Algorithm:
Step 1] Place randomly initial group centroids into the 2d space.
Step 2] Assign each object to the group that has the closest centroid.
Step 3] Recalculate the positions of the centroids.
Step 4] If the positions of the centroid didn't change go to the next step,else go to step 2.
step 5] End.
Clusters are displayed with the time required for clustering. With this intra distance and inter distance between cluster s are also calculated.

V. IMPLEMENTED WORK

In this , We have implemented the bisecting KMean algorithm.
Bisecting KMean algorithm is given below:
Step 1] Pick a cluster to split.
Step 2] Find two sub clusters using the basic KMean algorithm (Bisecting Step).
Step 3] Repeat Step 2,the bisecting step for ITER times and take the split that produces the clustering with the highest overall similarity.

© 2014, IRJCS- All Rights Reserved

Step 4] Repeat steps 1, 2 and 3 until the desired number of clusters are reach.

A. Reading of FACEBOOK 100 dataset.
B. Calculate the size of FACEBOOK 100 dataset.
C. Initially all fields are kept to zero.
D. Records are written and displayed .
E. Enter the number of cluster needed.
F. Start the timer.
G. Bisect the cluster in two parts

```
Index =KMean(current_data,2)
```

CL1 and CL2 will become two clusters with index 1 and 2.

```
CL1=length (find (index==1)));
CL2=length (find (index==2)));
```

Check for CL1>CL2, If yes save CL2 and pass CL1 for current data to cluster,else check for CL2>CL1.
Out_cluster will contain the cluster.

```
Out_clusters (clusters)=mat2cell(current_data);
```

H. Stop the clock.
I. Initially inter distance and intra distance are kept zero.
J. Inter distance and intra distance is calculated,
K. Calculation of inter distance.

```
Data1 = data of first cluster.
Data2 = data of second cluster.
distance = (Data1 - Data2);
distance = distance .* distance;
distance = sqrt(distance);
distance = distance / length(Data1);
Inter_distance = inter_distance + distance;
```

L. Calculation of intra distance.

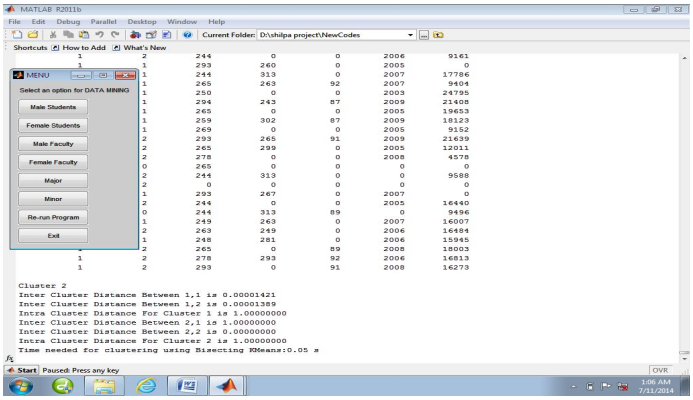
```
Calculate the absolute value of cluster1 data
dist2 = dist2 / size (Cluster1Data,1);
intra_distance(1) = dist2;
Calculate the absolute value of cluster 2 data
dist2 = dist2 / size (Cluster2Data,1);
intra_distance(2) = dist2;
```

M. Result

```
Inter Cluster Distance Between 1,1 is 0.00001421
Inter Cluster Distance Between 1,2 is 0.00001389
Intra Cluster Distance For Cluster 1 is 1.00000000
Inter Cluster Distance Between 2,1 is 1.00000000
Inter Cluster Distance Between 2,2 is 0.00000000
Intra Cluster Distance For Cluster 2 is 1.00000000
Time needed for clustering using Normal Kmeans : 0.68 s
Inter Cluster Distance Between 1,1 is 0.00001421
Inter Cluster Distance Between 1,2 is 0.00001389
Intra Cluster Distance For Cluster 1 is 1.00000000
Inter Cluster Distance Between 2,1 is 1.00000000
Inter Cluster Distance Between 2,2 is 0.00000000
Intra Cluster Distance For Cluster 2 is 1.00000000
Time needed for clustering using Bisecting Kmeans : 0.05 s
```

N. Apriori Algorithm is used for pattern recognition.
O. Menu for (“Male Student, Female Student, Male Faculty, Female Faculty, Major, Minor”)

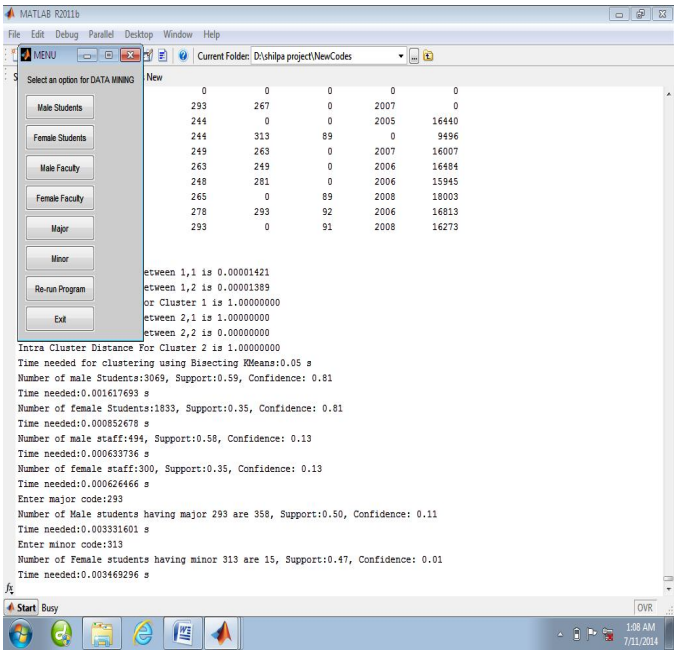
© 2014, IRJCS- All Rights Reserved



P. Support=0, Confidence=0
If first column match with 1 increment Confidence with1.
If second column match with 1 increment Support with 1.
Repeat these conditions with every menu provided.

Q. Support and Confidence is calculated as
Support=Support/Confidence
Confidence=Confidence/size (local_info,1).

R. Ttime is calculated.



VI. ACCURACY OF CLUSTERING ALGORITHMS

The accuracy of various algorithms is calculated using Rand Index. Rand Index is the measure of the similarity between clustering. A form of the Rand index may be defined that is adjusted for the chance grouping of elements; this is the adjusted Rand index. From a mathematical standpoint, Rand index is related to the accuracy, but is applicable even when class labels are not used.

RI (Accuracy) = (TP+TN)/(TP+FN+FP+TN)

Where,

True positive (TP): Number of positive sample correctly Predicted. If the outcome from a prediction is positive and the actual value is also positive, then it is called true positive.

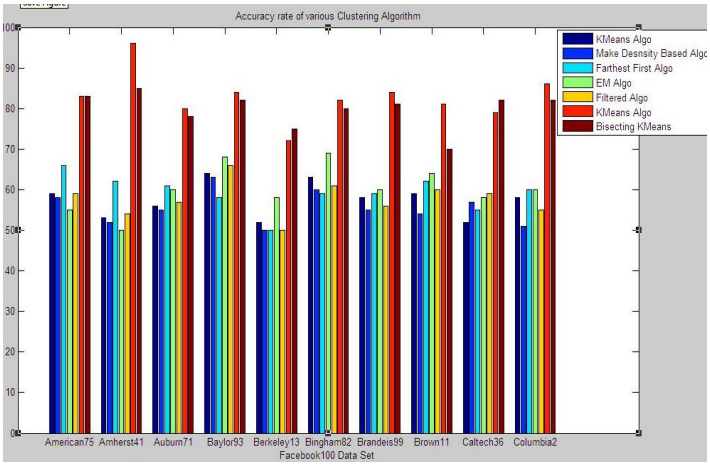
False positive (FP):- Number of negative sample incorrectly predicted as positive. If the actual value is negative then it is named false positive (FP).

False negative (FN):-Number of positive sample incorrectly predicted.

True negative (TN):-Number of negative samples correctly predicted.

Graphical Representation of accuracy rate for various clustering algorithms,

© 2014, IRJCS- All Rights Reserved



VII. CONCLUSION

This work will find the best possible clustering for the FB 100 dataset and in the most efficient manner, Again it highlights the formation of association rules between the attributes and explores the association rule between different parameter, and discovers the patterns.

REFERENCES

[1] Potomac, MD ,Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery,1999.

[2] PI. Nancy, G. Ramani,"Discovery of Patterns and evaluation of Clustering Algorithms in SocialNetwork Data (Face book 100Universities) through Data Mining Techniques and Methods" *International Journal of Data Mining & Knowledge Management Process" (IJDKP)* Vol.2, No.5, September 2012.

[3] Cluster analysis, en.wikipedia.org/wiki/Cluster_analysis.

[4] Z. Yongli ,Z. Yungui ,T. Weiming ,C. Hongzhi ,"An Improved Feature Selection Algorithm Based on MAHALANOBIS Distance for Network Intrusion Detection ",2013 *International Conference on Sensor Network Security Technology and Privacy Communication System* (SNS & PCS).

[5] L. Qi, H. Lin, V. Honavar ,"Clustering Remote RDF Data Using SPARQL Update Queries ",*ICDE Workshops 2013* @IEEE.

[6] D. Combe, C. Largeron, E. Zsigmond, M. Gery ,"Combining relations and text in scientific network clustering ",2012 IEEE/ACM *International Conference on Advances in Social Networks Analysis and Mining*.

[7] Z. Hu, X. Wang,K. Xu,"Mining Community in Social Network using Call Detail Records ", 2012 9th *International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2012).

[8] J. Yi Li, Z. Yuan Zhang, R. Yang Zhang, X.Mo,S. Wang,"An Empirical Study of Behavior Unconformity in Web-based and Non-web-based Social Networks ",2012 9th *International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2012).

[9] J. Pfeiffer, T. Fond , S. Moreno, J. Neville ,"Fast Generation of Large Scale Social Networks While Incorporating Transitive Closures ", 2012 *ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*.

[10] C. Li,"C entrality Analysis,Role-based clustering and ego centric abstraction for heterogeneous social networks",2012 ASE/IEEE *International Conference on Social Computing* and 2012 *ASE/IEEE International Conference on Privacy, Security, Risk and Trust*.

[11] A.L. T. Mucha and M. Porter,"Social Structure of Facebook Networks ",February 11, 2011.

[12] K. Juszczyszyn, M. Budka, K. Musiał,"The Dynamic Structural Patterns of Social Networks Based on Triad Transitions", 2011 *International Conference on Advances in Social Networks Analysis and Mining*.

[13] Z. Wang Daqing Zhang, D. Yang, Z. Yu ,Xingshe Zhou, Zhiwen Yu ,"Investigating City Characteristics based on Community Profiling in LBSNs",2012 *Second International Conference on Cloud and Green Computing* .

[14] R.Chithra ,S.Nickolas ,"A Novel Algorithm for Mining Hybrid-Dimensional Association Rules ",©2010 *International Journal of Computer Applications* (0975– 8887) Volume 1 – No. 16.

[15] C. Lin,Z. Huang ,F. Yang ,Q. Zou ,"Identify content quality in online social networks ",ET Commun., 2012, Vol. 6, Iss. 12, pp. 1618–1624, 2012.

[16] Q. Zhao and S. Bhowmick,"Association Rule Mining: A Survey ",Technical Report, *CAIS*, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[17] M. Goldberg, M. Hayvanovych, A. Hoonlor ,"Discovery, Analysis and Monitoring of Hidden Social Networks and Their Evolution ".

[18] S. Swami, "Mining association rules between sets of items in large databases". Proceedings of the 1993 *ACM SIGMOD international conference on Management of data - SIGMOD* '93. p. 207.

© 2014, IRJCS- All Rights Reserved