International Journal of Research In Science & Engineering
Volume: 3 Issue: 6 Nov-Dec 2017
e-ISSN: 2394-8299
p-ISSN: 2394-8280

IMPLEMENTATION OF BISECTING KMEAN ALGORITHM ON SOCIAL NETWORKING DATASET FOR ENHANCING PATTERN RECOGNITION

Prof. Ashish P. Nanotkar[1], Prof. Aniruddha A. Kolpyakwar[2], Prof.Pallavi A. Chaudhari[3]

[1]Bapurao Deshmukh College of Engineering (Sevagram),Wardha, India., ashishnanotkar@gmail.com
[2]Jagdamba College of Engineering & Technology, Yavatmal, MS, India. anirudha.kolpyakwar@yahoo.com
[3]pallavi.chaudhari17@gmail.com

Abstract**—*****Databases today can range in size more than terabytes. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do we draw meaningful conclusions about the forest? The newest answer is data mining, which is being used both to increase revenues and to reduce costs. 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 KMeans algorithm on it, by which we would get better clustering results. Bisecting KMeans first bisects the data into 2 parts and selects the part with greater number of elements, then applies clustering on it again. This goes on till we have N Number of clusters. We would apply this to our dataset to get desired results. With this we are going to compare Bisecting K Mean algorithm with other data mining algorithm. And finally we are going to find out different pattern from social networking dataset.***

Keywords— SNS, Bisecting KMean, KMean, Cluster, Pattern

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. Furthermore, the patterns uncovered by data mining must be verified in the real world. 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]. It gives way to share information between people anywhere and at anytime. 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 bioinformatics. 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 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. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification primarily their discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal.

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 assortativity 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].

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.

Lin (C.Lin at.el, 2012) analyzes the effects of distinguishing features on UGC (user generated contents) quality in different types of OSNs (online social network). Extensive statistical analysis leads to the discovery of existence of diverse patterns of human information sharing activity in dissimilar OSNs [15]. This discovery is employed as prior knowledge in the classification framework, which decompose the original highly imbalanced problem into several balanced sub-problems. Ensemble classifiers are adopted in samples from clusters generated by incompact features. Experiments show the proposed framework is both effective and efficient for several OSNs. Contributions of this study are twofold: (i) model posting activity in different types of OSNs; (ii) propose novel classification framework to identify UGC quality.

Qiankun Zhao and Sourav S. Bhowmick(2003) surveyed the list of existing association rule mining techniques [16] like AIS Algorithm which focus on improving the quality of databases together with necessary functionality to process decision support queries. AIS is just a straightforward approach that requires many passes over the database, generating many candidate item sets and storing counters of each candidate while most of them turn out to be not frequent. Apriori is more efficient during the candidate generation process for two reasons; Apriori employs a different candidate's generation method and a new pruning technique. FP-Tree(Frequent Pattern Tree) Algorithm break the two bottlenecks of Apriori series algorithms, some works of association rule mining using tree structure have been designed. FP-Tree [Han et al. 2000], frequent pattern mining, is another milestone in the development of association rule mining, which breaks the two bottlenecks of the Apriori. The frequent item sets are generated with only two passes over the database and without any candidate generation process. FP-Tree was introduced by Han et al in [Han et al. 2000]. By avoiding the candidate generation process and less passes over the database, FP-Tree is an order of magnitude faster than the Apriori algorithm. The frequent patterns generation process includes two sub processes: constructing the FT-Tree and generating frequent patterns from the FP-Tree. RARM (Rapid Association Rule Mining) [Das et al. 2001] is another association rule mining method that uses the tree structure to represent the original database and avoids candidate generation process. RARM is claimed to be much faster than FP-Tree algorithm.

Mark Goldberg (Mark et.al) presents a set of tools for discovery, analysis and monitoring evolution of hidden social groups on the internet and in cyberspace. One is based on statistical analysis of communication network without considering communication content. The other was focused on communication content and analyzes recursive patterns arising in it [17]. They present a software system SIGHTS (Statistical Identification of Groups Hidden in Time and Space), designed for the discovery, analysis, and knowledge visualization of social coalition in communication networks by analyzing communication patterns. The algorithms extract groups and track their evolution in Enron-email dataset and in Blog data. The goal of SIGHTS is to assist an analyst in identifying relevant information.

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. PROPOSED WORK AND OBJECTIVE

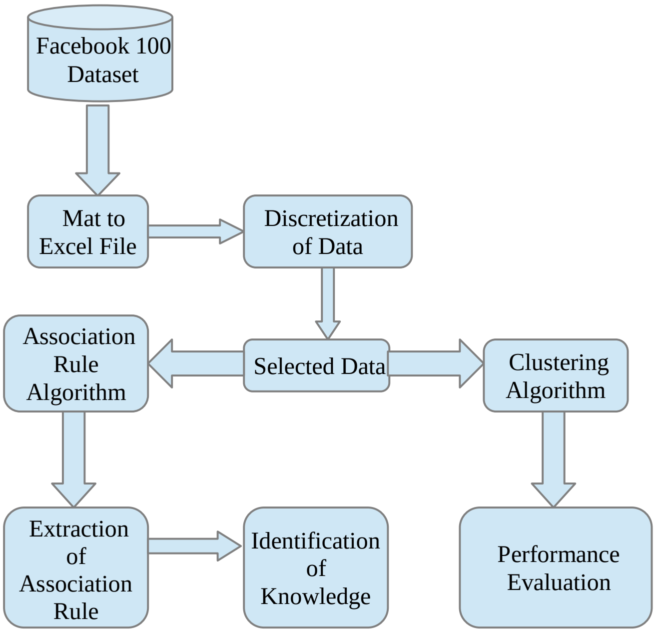
We are using the Facebook 100 dataset and applying the Bisecting KMeans algorithm on it, by which we would get better clustering results. Bisecting KMeans first bisects the data into 2 parts and selects the part with greater number of elements, then applies clustering on it again. This goes on till we have N Number of clusters. We would apply this to our dataset to get desired results.

A. Basic Bisecting K-means Algorithm for finding K clusters

This algorithm tries to improve quality over K Means. It starts with one large cluster of all the data points and divides the whole dataset Into two clusters. K Means algorithm is run multiple times to find a split that produce maximum intra cluster similarity. Then the cluster with largest size is picked to split further. This cluster can be chosen based upon minimum intra cluster similarity also. This algorithm is run k – 1 time to get k clusters. This algorithm performs better than regular K Means because bisecting K Means produces almost uniform sized clusters. While in regular K Means there can be notable difference between sizes of the clusters. As small cluster tends to have high intra cluster similarity, large clusters have very low intra cluster similarity and overall intra cluster similarity decreases. The algorithm is as described below:

- Pick a cluster to split.
- Find 2 sub-clusters using the basic k-Means algorithm (Bisecting step)
- Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
- Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

B. DESIGN VIEW



C. OBJECTIVES

- Social networking dataset can be clustered in many ways but to find the best possible clustering from the dataset is the main task
- The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There of course is a common denominator: a group of data objects. However, different cluster models can be used, and for each of these cluster models again different algorithms can be given.
- Understanding these "cluster models" is key to understanding the differences between the various algorithms. The centroid models for example the k-means algorithm represents each cluster by a single mean vector. But as it is having drawbacks Bisecting Kmean algorithm is used.
- By using algorithm, the dataset is clustered in the most efficient manner with least waste of time and effort.
- After clustering the formation and explore of association rule between different parameter and so the useful patterns are discovered.

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 100 Universities) through Data Mining Techniques and Methods "*International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.2, No.5, September 2012.

[3] cluster analysis http://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.