



A Survey on Online Stock forum using Subspace Clustering

Pooranam.N

PG Scholar,

Dept of CSE, Sri Shakthi Institute of
Engineering and Technology, India

Shyamala.G

Asst. Prof,

Dept of CSE, Sri Shakthi Institute of
Engineering and Technology, India

Abstract- *Financial stock Data Analysis and future prediction in terms of Sentiments is great challenge in the big data research. Among the unlabelled opinion, opinion classification in terms of unsupervised learning algorithm will lead to classification error as data is sparse and high dimensional. To overcome this problem, the sentiment analysis to extract the opinion of each word in the stock data has been proposed. Moreover the data size is large, hence the singular value decomposition to resolve the inconsistent constraints correlating to the large dimensions, and dimensionally reduced feature set is been used. The dimensionally reduced feature set is classified into clusters through employment of Principle component analysis with utilization of the domain knowledge. Cluster data which further inconsistent with the outlier probability can further reduced through subspace clustering. Experimental results prove that the proposed framework outperforms the state of art approaches in terms of precision, recall and Fmeasure.*

Keywords: *Opinion mining, sentiment analysis, Singular Value Decomposition, Principle Component Analysis, Subspace Clustering.*

I. INTRODUCTION

Big data is a structured, unstructured, semi structured data that have three different characteristics obtained they are volume, velocity, variety (V3). The key features of the big data mainly focus on 1.) Enhance the storage capacity 2.) High processing power 3.) Data accessibility. Some of available data types are social networks and mobile devices, internet of things etc.,. The tools used in big data are defined as follows Database – NoSQL, HBase, map reduce- hadoop, hive, pig, storage- s3, Hadoop Distributed File System (HDFS) [9]. The challenges facing in analytics are fault tolerance , scalability, quality of data. Applications are Telecommunication Network Monitoring, social network, instant-messenger network, health care etc, [9] The main issues in big data are associated in adapting to the technology and its tool. Storage and transport of data, managing data is a big issue in analytics [1].

II. LITERATURE REVIEW

A. “David R. Hardoon” [10]

Have illustrated about solving the clustering problems and discover biological significant. Centroid-based clustering clusters are represented by a central vector also known as centroid which is not necessarily a member of the data set. They anticipated a new framework called CATSeeker which integrate domain knowledge. The two main problems in subspace clustering are 1.) Usefulness 2.) Usability this problem can be solved by mounting the domain knowledge. CATS is a three dimensional data they are 1.) Trimming the search space 2.) Finding the objects centroid which has high utilities. 3.) Mining the subspace. This CATSeeker three module a.) Calculating the value of the data set using SVD b.) Estimating the value of methods c.) Mining CATs. Comparisons are made between different algorithms to reduce the redundancy. CATS denote utility function of the objects and which concurrently handles the multifacts and effectiveness of the data.

B. “A. Zimek” [3]

Have made a study on subspace clustering and relative problems. Many comparisons have been made to solve the problem with different solutions but they did not clarify about exact problem definition. They compared experimentally whether the two different solutions track same problem and the assumptions are made on outcome of an algorithm. They try to clarify 1.) General difficulty in defining the subspace clustering, 2.) Difficulties on the field of research 3.) Assumptions of different approaches and finally how solutions deal with different problems. The main objective of cluster analysis gives a superior kind of structured records. The cluster analysis is used in indexing and data compression. Applications of clustering high dimensional data where studied and described a.) Gene Expression analysis b.) Clustering the rows c.) Clustering the columns d.) Co clustering Rows and Columns e.) Metabolic Screening f.) Customer Recommendation System g.) Text Documents. The problem of high – dimensional data can be overcome feature selection methods like Principal component Analysis (PCA) which is used in the direction of mapping the original data space to dimensional data.

C. “Guimei Liu”, [2]

Have described about the efficiency in mining maximal nCluster and accuracy of MaxnCluster. They have anticipated a new model known as nCluster model which uses sliding window approach to protect major clusters and generate more bins. They develop an algorithm called Maxn-cluster uses different methods to speed the mining plus reduce the result size. This is more efficient and accurate. Non-maximal nclusters are trimmed. The problem defines about subspace δ -neighbors and subspace δ -nCluster. The Maxn Cluster algorithm is customized as of model development algorithms which define about support and frequency. This uses a dense tree formation called FPO- tree to store the support count. This tree uses two techniques 1.) Pruning derived from support node and child node. 2.) Pruning by the closure items. The performance study is made on the quality and the efficiency. PROCLUS, ORCLUS and FINDIT are the distance and screened based subspace clustering algorithm to find the non-overlapping clusters by assigning weights for each cluster.

D. “Lizhuang Zhao”, [4]

Have introduced an algorithm called TRICLUSTER which is the 3D microarray subspace clustering method. TRICLUSTER identify some cluster having large overlap. They present a set of metrics to calculate the value of cluster. This algorithm define some challenges a.) The maximal Tricluster satisfy certain criteria. b) The clusters arbitrary overlapping regions c.) Tricluster have identical value for all subsets of the dimensionality d.) This has an inherent property for effective mining. TRICLUSTER construct multigraph which yield the set of bicluster. The bicluster is a recursive algorithm and apply a Depth First Search (DFS) on multigraph. Tricluster can find significant cluster in actual data. They have planned to enlarge new method for reducing the space.

E. “Haoliang Jiang”, [8]

Have anticipated a clustering algorithm known as gTRICLUSTER which is to find significant cluster in microarray GST data. This is based on common 3D cluster model find in more biological gene cluster than TRICLUSTER. This TRICLUSTER is robustness to noise. They discuss about the new model that keep away from the proportional property of TRICLUSTER with genetic surveillance. The new algorithm expands the real world data for effectiveness. This deal with the Tricluster mining which uses a SRC (Spearman rank correlation) to estimate the relationship of arbitrary profile. GST data clustering has examine the subset of the sample during the time series. The data has qualitative behavior rather than actual value. The two algorithms are compared and the performance is evaluated the results are plotted in the form of graph. The results are obtained to find the characteristics of different gene. If the two similar patterns appear in same time region then the time latency is calculated for the effectiveness of the data.

F. “Amir Adler”, [5]

Have discussed about a new approach called probabilistic subspace clustering capable of clustering large signal collections. The mixture model is derived from co-occurrences matrix which consists of both atom and signals. The component of mixture model is obtained from non-negative matrix factorization (NNMF) and subspace of maximum likelihood (ML). There are four approaches discuss in subspace clustering algorithm a.) A review approach b.) State-of-the-art approach c.) Low-Rank Representation (LRR). d.) Closed Form Solution of LRR (LRR-CFS). They define some problems addressed on the collection of data at large amount of signals. The reason may be due to the tasks required to handle image and video stream and require processing of large data sets. They formulate the clustering problems based on subspace separation, signal quality, model accuracy. They proposed a model called aspect model which overcome the probabilistic subspace clustering problem formulations. The performance evaluation is made on different data collection where the computation time depends only on linear data.

G. “Casey Whitelaw”, [6]

Have described about the non-topical text analysis which is characterized as opinions, feelings and attribute expressions. The problem have been defined in this district is sentiment classification which has been labeled as positive and negative objects. Some applications include web mining, market research etc.,. This analysis has two main approaches the first is document classifier based frequency and second is based on two classes. The main goal of this study on a method called Appraisal Group comprises of head adjective with attitude type. The appraisal group can significantly improve both lexicon and sentimental classification. The group contains four types they are Attitude, Orientation, Graduation and Polarity. The automated technique lexicon has been constructed to find the attribute value and is stored in appraisal adjective. The method is defined for factor sets like words by attitude, system by attitude, systems by attitude and orientation, bag of words. The challenges of classification are mainly focus on accurate identification without filtering.

H. “Abdullah Embong”, [11]

Have described about two main approaches called subspace clustering and projected clustering in high dimensional spaces. They have studied and compared the results of three algorithms namely PROCLUS, P3C, STATPC. And analyze the properties of different cluster method. The PROCLUS is better in performance while calculating the least number of un-clustered data. STATPC perform better accuracy in both cluster points and relevant attributes when compare to PROCLUS and P3C. PROCLUS is focus on cluster method to find small projection subspace of high dimensional data. P3C is an effective algorithm to discover data while minimizing the number of required parameters. They made a study

on experimental results how the three algorithm works and what parameter is been evolved and the performance accuracy has been analyzed.

I. “Eugene Agichtein”, [7]

Have described about ordering of web search improvement. They incorporate the comment into ranking process and discover the web search feature. The main goal of this process to know how implicit comment can be used in operational location. This is focused mainly on analysis of alternatives behavior into web search ranking and a model has been derived for mining and feedback. There are two approaches to rank the comment 1.) Implicit comment as independent ranking result. 2.) Implicit comment directly features into ranking algorithm. They explored effectiveness of noisy implicit comment to get better web search ranking. The experiments are made on significant process that do not consider implicit comment feature. Feedback is valuable for a particular query that will reduce ranking results.

J. “Carlos Hurtado”, [12]

Have proposed a method to present the query into search engine that advice a directory of related queries. This method is based on a query clustering process and discover related queries also ranks according to two particular criteria. They are a.) The similarity of the input queries b.) The measure of support for recommended query. The group of related query is found by clustering process. This process is based in terms of weight vector of the query. The effectiveness of the method is experimentally shown over query log. The rank score of the query is measured using two notations a.) Similarity and b.) Support. The experiment is made on different queries and the result is drawn through graph.

Table I. Comparison Table

S.No.	Methods	Usage	Advantage	Disadvantage
1	Centroid-based clustering using CATSeeker framework and SVD	To encore domain knowledge of cluster data	The usefulness and usability is improved and reduce redundancy	Calculating the value of data is hard and slow.
2	Principal component Analysis (PCA)	To plot the original data	More efficient in solving the problems	This method is difficult to construct.
3	ncluster model and max cluster algorithm	Store large amount of data	ncluster model increases mining and reduce the size	Produce more bins space are dissipated.
4	3D microarray subspace clustering	Find significance of actual data	Efficient in mining the data	Creates more overlapping region.
5	TRICLUSTER, g TRICLUSTER	Similarity of arbitrary profile and create gene cluster which estimate the relationship of the profile	Effectiveness of the data and more robustness to noise	The time sequence is relatively high.
6	Aspect model	The signals are collected from each cluster	Overcome the probabilistic clustering problems and signals are more accurate	Linear data require more time to generate signals
7	Lexicon	To collect opinions, feelings and attribute expressions	An automated technique focuses on accurate identification of expressions and improves classification.	Performance is imperfect due to group.
8	Subspace clustering and Projected clustering	Analyze the performance of clustering	Performance is high in different dimensions of data.	Calculating the least number of un cluster is complex.
9	Web search ranking	To discover the web search characteristics	Improves ranking process in implicit comment.	Some comments produce unfortunate creativity of ranking results.
10	Query clustering process	To group relative query and find weight vector.	The rank score is accurate and easy to measure. Effectiveness is based on query logs	The prediction on each query log is decreased.

III. CONCLUSION

This paper describes about various algorithms, methods and models involved in cluster analysis. This illustrates many techniques for classifying and analyzing data. This model uses different approaches for significant clusters and effectiveness of noise. This survey focus on different models like PCA, ncluster, aspect, which is mainly involved in high dimensional clustering data. The comparison is made between different models and methods for the efficiency of high dimensional data. This paper offers field information on clustering Analysis.

REFERENCES

- [1] Katal, A . ; Dept. of CSE, Graphic Era Univ., Dehradun, India ; Wazid, M. ; Goudar, R.H.”Big data: Issues, challenges, tools and Good practices”2013.
- [2] G. Liu, K. Sim, J. Li, and L. Wong, “Efficient Mining of Distance-Based Subspace Clusters,” *Statistical Analysis Data Mining*, vol. 2, nos. 5/6, pp. 427-444, 2009.
- [3] H.-P. Kriegel, P. Kroger, and A. Zimek, “Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering,” *ACM Trans.Knowledge Discovery from Data*, vol. 3, no. 1, pp. 1-58, 2009.
- [4] L. Zhao and M.J. Zaki, “TRICLUSTER: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data,” *Proc. ACM SIGMOD Int’l Conf. Management of Data*, pp. 694-705. 2005.
- [5] Amir Adler, Michael Elad, Fellow, IEEE, and Yacov Hel-Or “Probabilistic Subspace Clustering Via Sparse Representations“ *IEEE signal processing letters*, vol.20, no. 1, january 2013.
- [6] Casey Whitelaw, Navendu Garg and Shlomo Argamon “Using Appraisal Taxonomies for Sentiment Analysis”2005.
- [7] Eugene Agichtein Eric Brill Susan Dumais”Improving Web Search Ranking by Incorporating User Behavior Information “ *SIGIR’06*, August 6–11, 2006.
- [8] Haoliang Jiang¹ , Shuigeng Zhou^{1,2} , Jihong Guan³ , and Ying Zheng¹”gTRICLUSTER: A More General and Effective 3D Clustering Algorithm for Gene-Sample-Time Microarray Data” Li et al. (Eds.): *BioDM 2006*, LNBI 3916, pp. 48–59, 2006.
- [9] http://www.planetdata.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf
- [10] Kelvin Sim, Ghim-Eng Yap, David R. Hardoon, Vivekanand Gopalkrishnan, Gao Cong, and Suryani Lukman“Centroid-Based Actionable 3D Subspace Clustering” *IEEE transactions on knowledge and data engineering*, vol. 25,no. 6,june 2013.
- [11] Rahmat Widia Sembiring¹, Jasni Mohamad Zain², Abdullah Embong³”clustering high dimensional data using subspace and projected clustering algorithms”*International journal of computer science & information Technology (IJCSIT)* Vol.2, No.4, August 2010.
- [12] Ricardo Baeza-Yates¹ , Carlos Hurtado¹, and Marcelo Mendoza²”Query Recommendation using Query Logs in Search Engines”2004.