DATA MINING OF SOCIAL NETWORKS USING CLUSTERING BASED-SVM

Pooja Sikka
Assistant Professor, AIMT/ CSE Deptt., Indri-Karnal, Haryana, India
Email: pooja.nov22@gmail.com

Abstract— Data mining refers to extracting or mining of useful information from large amounts of records or data. Data mining includes the task of data clustering, association analysis and evolution analysis. SVM (Support vector machines) had been the most developed method for classification and regression technique due to its favourable features such as margin maximization and systematic nonlinear classification. SVM have not been favoured for large data sets for mining so to overcome this k-means micro-clustering technique will be implied with SVM.

Keywords- HRV (Heart rate variability) and MER (Microelectrode recordings).

1. INTRODUCTION

Data mining is a technique to analyze data from different perspectives and to summarize it into useful information which can be further used for increasing revenues, cost cutting etc. This technique provides tools to extract required information. Applications of data mining are wireless sensor networks, pattern recognition, data compression, machine learning etc. [12].

Data mining can be viewed as a result of the natural evolution of information technology. The efficient mining of the information will be done by retrieving information from large scale data helps to discover valuable knowledge of paramount importance in different fields like marketing, banking, government and defense. Data mining include various sub tasks:

- Association Analysis - This task will show relationships between attribute values sets.
- Classification and prediction - This task will discover a model that describes and were presented.
- Cluster analysis - This task will group similar data objects based on some similarity measures. Each group of similar data can be viewed as a class of objects.

1.1 Introduction to Data Mining

Data mining can be viewed as a result of the natural evolution of information technology. Data mining in social networks is the process of extracting application-oriented models and patterns with acceptable accuracy from a continuous, rapid, and possibly non ended flow of data streams from networks [7]. In this case whole data cannot be stored and must be processed immediately.

Data mining algorithm has to be sufficiently fast to process high-speed arriving data. The conventional data mining algorithms are meant to handle the static data and use the multistep techniques and multiscan mining algorithms for analysing static data-sets.

Data mining techniques has shown to be capable of mining large data generated on social networking sites [12]. This is made possible by way of extracting information from large data sets generated on social networks and transforming them into understandable structure for further use. These days many people are becoming interested in social networks for information, breaking news and other diverse subject matters. Users find out what other people’s views are about certain product/service, film, school, or even more major issues like seeking other people’s opinion on political candidates in national election poll.

The data mining techniques that have been developed recently for relational data include probabilistic relational models (PRMs), Bayesian logic programs (BLPs), first-order Bayesian classifiers, and relational probability trees (RPTs). In each case both the structure and the parameters of a statistical model can be learned directly from data, easing the job of data analysts and greatly improving the fidelity of the resulting model.
1.2 Emotional Analysis on Social Networks

Information about social relationships can be used to improve user-level emotional analysis in social networks. Emotional analysis could be referred to as discovery and recognition of positive or negative expression of opinion by people on diverse subject matters of interest [11]. A social network is the collection of social organizations or individuals and set of interconnections between them. The most well-known social networking sites are Facebook, Twitter, and Myspace. The social networks are complex networks which are the representation of a complex system from real life in terms of nodes and edges where a node is an individual member in the system and an edge is a link between nodes according to a relation in the system. The main motive regarding this technique is that users that are connected may be more likely to hold similar opinions; therefore, relationship information can complement what we can extract about a user's viewpoints from their utterances [8].

1.2.1 Challenges in Emotional Analysis

A number of challenges facing the realisation of utilising data mining techniques in sentiment analysis could be identified as follows [11]:

- Ambiguity in the emotions portrayed by users:- The expressions expressed by users sometimes may be direct or indirect. The condition of ambiguity occurs when more than one user express similar sentiments at same time.
- Noisy texts such as grammatical mistakes, missing punctuation and slang are still a big challenge to most emotional analysis systems.
- Need for modelling compositional sentiments:- Accurate calculations and manipulation of the overall sentence sentiment of the sentiment-bearing words, the sentiment shifters, and the sentence structure.

1.3 Introduction to Mining Techniques

- Naïve Bayes method- In this algorithm data related to conditional probabilities which are collected by counting the occurrence of values and combinations of values from historical data. The other name for this technique is conditional probability. This technique is depended on hypothetical data collection and manipulation. It is used for large data set. This technique is most efficient for weather forecast mining.
- Support vector machine- This technique is most preferred for emotional analysis of social network. It is based on kernel-based supervised learning technique. Due to some of its favourable properties this method is not suitable for large scale data mining as for pattern recognition or machine learning because of training complexity because it is highly dependent on size of data set [5]. In this the boundary function is described by the support vectors in which the data is located closet to boundary.
- Decision Tree- A decision tree is a classifier described as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree which is a directed tree with a node called root that has no incoming edges. The technique is used with review sets of positive and negative as leaf nodes. The most significant review formed the root of the tree. Decision tree is used as trade-off for highly predictive techniques.
- K- Nearest neighbor- This technique is versatile and simple to understand. It is a non-parametric classification method. This technique is much less preferred than SVM and Naïve bayes method. This does
not include the training data points to do any generalization. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

- **K-Means Clustering:** This method classifies the dataset through a certain number of K clusters. Then it defines K centres for K clusters which are placed as far as possible from each other. Then each point in dataset is associated to the nearest data centre. The following formula finds the new cluster centre [11].

\[ V_1 = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i \]

Where \( c_i \) represents the number of data points in \( i_{th} \) cluster and \( x_i \) is the data point. The above process is repeated till no new centre is reassigned.

### 2. Proposed Model

In new proposed method SVM technique and K-Means clustering technique are used for data mining in social networks. This method will increase the training efficiency by mutating it. The new technique is named as K-Means Clustering Based SVM (KMCS). The complexity of SVM depends on no. of input variables and support vectors. The proposed method will reduce the number of support vectors only the necessary ones. When the size of large data sets were used earlier, SVM tends to perform worse with training from the entire data than training the dataset with a fine quality of data samples. The SVM can build the model including about 6000 support vectors while the proposed method will build model with only 100 support vectors. So it will be 60 times faster than SVM technique which is highly required for large datasets like social networks. The proposed model will apply a clustering algorithm that scans entire data set only to provide the high quality samples that will carry statistical data. This will provide finer description closer to boundary and farther to boundary. KMCS would be used for classifying large data sets of relatively low dimensions in large warehouses. The performance difference between SVM and KMCS increase from earlier methods as the data sets increases.

### 3. Literature Survey

Inoshika Dilrukshi et al. [1] proposed a research to classify news into different groups so that the user could identify the most popular news group in a given country for a given time. To extract the short messages several active news groups were chosen. Each short message was classified manually into 12 groups. These classified data were used to train the machine learning techniques. This data was trained using SVM (Support Vector Machine) machine learning techniques. SVM supports high dimensional data that is why SVM is used for the current research. The performance of the system will be the effectiveness of the system. The results show that the system provides high performance for most groups.

Chendi Wang et al. [2] proposed a method to evaluate human’s emotion and stress based on heart rate variability (HRV). Initially 4 kind of emotions were induced and the corresponding electrocardiogram (ECG) changes were measured in a laboratory setting; Secondly a fast and improved denoising method based on wavelet transform threshold denoising to process the noisy ECG signal. Finally a wide range of physiological features from various analysis domains including time, frequency and nonlinear analysis were proposed in order to find the best emotion-relevant features and to correlate them with emotional states.

Jerzy Surma et al. [3] discussed that the classification and regression trees (C&RT) model for identifying users of on-line social network was aimed at using the advanced data mining methods to enable business usage of social networks. The research presented in this paper confirmed the usage of data mining techniques in marketing campaign optimization. This was justified by significant improvement in response rate.

P. Guilleen et al. [4] proposed an efficient methodology for the characterization of Microelectrode Recordings (MER) which was obtained during deep brain stimulation surgery for Parkinson’s disease using Support Vector Machines. The methodology was based in two algorithms: (1) an algorithm which extracts multiple computational features from the microelectrode neurophysiology and (2) integrates them in the support vector machines algorithm for classification. The SVM algorithm performed quite well achieving 99.4% correct classification. In conclusion, the use of a computer-based system was intended to avoid human subjectivity in the localization of the subcortical structures and mainly the subthalamic nucleus (STN) for neurostimulation. Jun Wang et al. [5] proposed an improved method of extracting emotional attribute of relations from Chinese
events based on syntactical analysis. In proposed work the author improved the defect of previous method that cannot deal with multiple entities in one sentence. In addition a large-scale Chinese emotional dictionary not only emotional verbs was used in the extraction of emotional attribute.

Piotr Bródka et. al. [6] proposed a new method for the group evolution discovery called GED in this paper. The results of the first experiments were presented on the email based social network together with comparison with two other methods of group evolution discovery.

Kazi Masudul Alam et. al. [7] proposed an emotion tagger that takes content from diversified Internet-based services and maps the analysed content to an emotional value which could be used to augment semantic values of Internet-based service. The author add the proposed emotion tagger with earlier developed context-aware social network framework called SenseFace to show a proof of concept working environment and investigate the efficiency of the tagger.

Michael Steve Stanley Laine et. al. [8] described that the nature of the interaction between content and connections is fundamentally important not just from a social science perspective. The author suggested that YouTube recently added the ability for users to form groups in which explicit category affiliation is noted too. The author also investigates the role of users in groups, how groups evolve and the structure of these groups organized under a category change over time. Finally the author find what form of linkage motivates new members to join these groups.

Michiel van Meeteren et. al. [9] described a multi-method approach to delineate a real world community of practice from a large N dataset derived from the social networking site Twitter. The author proposed multi-method approach allows micro level inferences from a macro dataset of which the individual Twitter user might be completely unaware. The accuracy of this approach give rise to the discussion of the value for businesses and market research since it offers opportunities to identify and monitor target audiences at a finely grained level. However the author should be wary of the serious consequences with regard to privacy and ethics. The results could have consequences for the anonymity of key persons behind the scenes of social and political movements or any other communities whose members are active on Twitter or other social networks.

Yadong Zhou et. al. [10] analysed the dynamics of online groups discussing incidental popular topics and presenting a new model for predicting the dynamic sizes of incidental topic groups. The author discovered that the dynamic sizes of incidental topic groups follow a heavy-tailed distribution. The author developed an adaptive parametric method for predicting the dynamics of this type of group based on the data collected. The model presented in the article is validated using actual data from Live Journal and Sohu blogs. The empirical results show that the model can effectively predict the dynamic characteristics of incidental topic groups over both short and long timescales and outperforms the SIR model. The author conclude by offering two strategies for promoting the group popularity of incidental topics.

Nan Li et. al. [11] created an algorithm to automatically analyse the emotional polarity of a text and to obtain a value for each piece of text. This algorithm was combined with K-means clustering and support vector machine (SVM) to develop unsupervised text mining approach. The author used the proposed text mining approach to group the forums into various clusters with the centre of each representing a hotspot forum within the current time span.

Dr. Jaideep Srivastava et. al. [12] Proposed that social network analysis has gained prominence due to its use in different applications from product marketing to search engines and organizational dynamics. The author provide an introduction to social network analysis and then briefly survey the research in this field. The author presented the work in two areas: (i) data mining for socio-cognitive analysis of email networks and (ii) data mining on logs from massively multi-player online (MMO) games to understand social and group dynamics amongst players.

4. REFERENCES


[2] Chendi Wang, Feng Wang “An Emotional Analysis Method Based on Heart Rate Variability” International Conference on Biomedical and Health Informatics, Hong Kong and Shenzhen, China, 2-7 Jan, 2012


[10] Yadong Zhou, Xi’an Juatong “Group Dynamics in Discussing Incidental Topics over Online Social Networks” IEEE, 2010


