ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 7, July 2015

Prediction of Collective Behavior Using Edge-Centric Clustering Technique

Pranjali K. Deshmukh, Prof.R.M.Gawande

Abstract— This study is to predict behavior of people, which are connected through social media. The social media provides platform for people to express their thoughts, opinions, voice etc. Thousands of people like to co1nnect though social media which gives opportunity to learn pattern of behavior. Lot of data are generated by social media like Facebook, Twitter, Flickr, and YouTube, etc. However there are many challenges present in order to study collective behavior of the people. It is also interesting to extract behavioural features of user social activities and include them with social networking. These features are used for improving prediction performance of multimode network. This approach can handle millions of individuals to demonstrate a comparable prediction performance to other nonscalable methods. For scalability issues, apply an edge-centric clustering scheme for extracting sparse-social dimensions. As there are heterogeneous connections present in social media, a social-dimension-based approach has been shown beneficial for community detection. By using edge-centric clustering algorithm the category to which the users belong is predicted.

***Index Terms*— collective behavior, multimode, edge-centric clustering, social dimensions, Communities.**

I. INTRODUCTION

THE proliferation in the usage of Social networking sites has conferred people of various demographics and professions with innovative ways of associations, interactions and sharing of knowledge and information in (or of) groups. Enormous number of online users voluntarily writes encyclopedia article of extensive scale and scope. Endorsements by online bazaars of various merchandises are done by studying the customer's interests and behaviors. Even the experience of influence on social and governmental or political actions has been noticed largely. Social media provides opportunities to study human interactions and collective behavior on large scale.

In this work, prediction of behavior by giving some human behavior information and individual preference, try to predict behavior of other individual with same network. It helps better understand behavioral patterns of users in social media for applications like social advertising and recommendation. Collective Behavior: Collective behavior includes lot of activities which performed with group such as joining any group, sharing some post, clicking some adds, etc. then their behaviors are not independent. That is, their behaviors can be influenced by the behaviors of their friends. This naturally leads to behavior correlation between connected users. Collective Behavior prediction is not simply aggregation of individual behavior. Many social user are influenced by the behavior of friends Consider as an example, mostly we like to buy those things which our friends buy, without more investigation of those things because of homophily.

Homophily is a term coined in the 1950s to explain our tendency to link with one another in ways that confirm, rather than test, our core beliefs [9]. According to Homophily we link to connect with those peoples who share some similarity with us this is not only with physical world, but also with online systems. Simply we can say friends in a social network behave similarly according to homophily.

Recently various social sites encourage people to connect with each other. Different age group peoples like to communicate with each other. The social behavior includes many activities such as join groups, like photos, share some videos, tags etc. In this work consider the influence behavior rather than the correlation of users.

Motivation behind this study is to predict people social behaviors and personal choices in social networking media. The conventional relational classification model focuses on the single-label classification problem. But the real-world relational datasets contain instances associated with multiple labels. Connections between instances in multi-label networks are driven by various casual reasons. This paper is to predict the behavior of individuals by studying behavior of some other individuals in the same social network [3] which will help to know behavioral patterns of individuals in social networking environment for applications like social marketing and endorsements. Heterogeneous users are connected with each other. The users may be classmates, colleagues, and family members etc. The heterogeneity with network connections, limits the effectiveness of a commonly used technique – collective inference for network classification [2].

The objective of this paper is to find affiliation between individuals by applying the edge-centric view to find the sparse social dimensions. The edge-centric clustering algorithm is used to predict communities of users with similar behavior. The edge-clustering algorithm is variant of k-means clustering algorithm [1]. The scalability is the main issue that occurs with previous methods. For the scalability issue the edge-clustering framework is used to find user community. The extracted dimensions show the features of node. On the basis of connections present in network generate an instance based matrix. Link between two individuals is denoted as edge *m*, and the individual which denoted as node *n*. The edge-centric clustering which is variant of k-means algorithm generate clusters i.e. community. For its regularization Linear Support Vector Machine (SVM) can be used. The linear SVM work apply on clusters to finish work with linear time. Sometimes one node belongs to more than one community i.e. multimode; hence regularization shows effective work for classification.

***Manuscript received July, 2015*.**

Pranjali Kalyanrao Deshmukh, *PG Computer Science Student*, Shavitribai Phule Pune University. *India.*

Prof. R.M.Gawande, *Project Guide, Computer engineering Department* MCOERC Nashik, Shavitribai Phule Pune University. *India.*

II. Literature Survey

Data classification with network instance is known as within network classification [7]. In the conventional data mining methods the data instances are not identically distributed. A Markov dependency assumption is applied on data to find label of every node it depends on attribute of its neighbor node. This work like lazy learner, the node label is dependent upon the neighbor node label.

Relational classifiers are constructed on the base of relational features of labeled data. Update the class membership of data for every node while the label of neighboring node is fixed. Iteratively repeat the same process while inconsistency between two neighboring node is less. The drawback with Markov assumption is that it's only applicable on local dependency of network.

Instead of this method a simple weighted vote relational neighborhood classifier (wvRN) [11] works well and set a baseline for comparison. The wvRN gives weight to connection and calculate relation between two nodes.
The network clustering is based on weight assignment that is not sufficient for very large data base.

However L. Tang and H. Liu's work on soft clustering scheme [2] consider the different heterogeneous relations represents potential affiliation between actors. The soft clustering method is used to extract futures, and support vector machine can be used for classification. The soft clustering method solved the heterogeneous relation problem, but to deal with dense social dimensions are difficult to handle. Practically to handle the million node data is difficult because extra resources are required to store data. The S. Fortunato [8] gives a comparative survey of matrix factorization, spectral clustering, modularity maximization [4], Probabilistic methods for a comprehensive survey shows the challenges need to consider while performing soft clustering methods. Another method to find overlapping community by Palla et al. [6] is a clique percolation method to find overlapping communities. In this method first find all cliques of size k in a graph. k-1 nodes are shared if they are connected with two k-cliques. With respect to k-cliques connected component we can divide them into two different communities. For dividing nodes in corresponding community a clique method allow to represent the node in both communities. One node can be involved with two or more communities mean overlapping node, this issue solved by clique method.

Newman-Girvan [12] method find the overlaping community by recursively removing edges between the graph until it divides into different communities. This method only removes those edges which create and bridge with communities. But it gives output as only nonoverlapping communities. To overcome the problem S. Gregory [9] also handles overlapping communities with node (instead of edge). The algorithm recursively splits nodes those are likely to reside in two communities or removes edges that seem to bridge two different communities. Repeat the process until the network is disconnected into the desired number of communities. These overlapping methods require more computational cost for large-scale networks.
T. Evans al. [10] found a simple method construction of line graph. When we consider large network then formation of cycles is increased. Also the line graph requires all nodes of same degree. Practically it's not possible that every node have same affiliation with other node. Studying all this methods and considering scalability issue we formulate the overlapping community detection problem. The edge-clustering algorithm which is variant of k-means algorithm can handle the scalability issue [1]. Less amount of memory is required where the previous methods are failed due to dynamic nature of data.

III. Proposed Method

This paper is based on social networking where social media generate big data. It gives more challenges to handle that data. social information is available with multimode (tags, links, comments etc.) form. This paper is work on within network multimode data. For processing consider network is represented in the form of graph where user consider as node and two node connection i.e. link. This link is edge G(V,E) which connect two nodes. For processing considering whole profile of user is difficult, it requires more cost, by considering all attribute of node. So network handling issues will arose, that is explained with previous techniques [2][10]. By considering only link (edge) between two nodes is better solution for handling large network.

The following figure shows of Framework of Collective Behavioral Learning

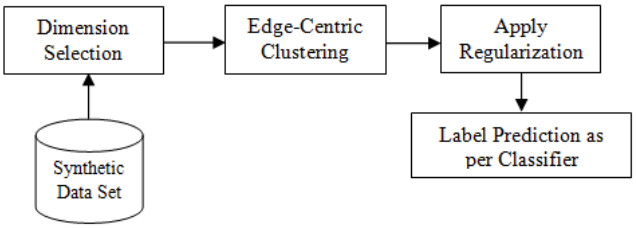

Fig1: Framework of Collective Behavioral Learning

For this framework we generate the synthetic dataset for better work. To consider all attributes for processing not possible so select the sparse dimension. That is possible by dividing network into disjoint form:

A. *Create edge centric view of network*

To solve this problem, network node is divided into disjoint sets. Instead of considering all attributes of node, consider only the edge between them for processing. Every edge having two end points, so one node may belong with many affiliations. Overlapping of community was the issue that is solved by this method. A network may be sparse, but the extracted social dimensions may not be sparse. Let's consider a toy network example [1].

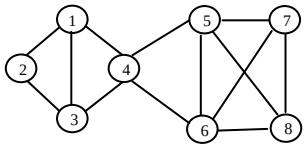

Fig.2. A Toy example

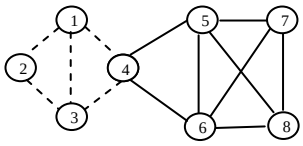

Fig.3. Edge clusters

After selection of proper dimensions, we get affiliation of all the nodes. On the basis of affiliation construct an instance based matrix. The edge between two nodes is treated as feature of that node. Table 1 shows the edge centric view of network data. The edge shows an extracted feature of toy network. Fig. 2 shows a Toy network; divide the network into disjoint sets. In Fig 3 the dashed edges represent one affiliation with one community, and the remaining edges denote the second affiliation as second community.

TABLE I
Edge Instances of Toy Network

Edge	Features								
	1	2	3	4	5	6	7	8	9
e(1,4)	1	0	0	1	0	0	0	0	0
e(1,3)	1	0	1	0	0	0	0	0	0
e(2,3)	0	1	1	0	0	0	0	0	0
	.....								

The table shows an instance of edge centric view of toy network. This scheme can extract sparse social dimensions. With such a scheme, we can also update the social dimensions efficiently when new nodes or new edges arrive. Then a typical clustering algorithm like k-means clustering can be applied to find disjoint partitions.

B. *K-means algorithm for clustering*

Input: Data instances{xi|1≤i≤m}, Number of clusters *k*
Output: Number of clusters $\{idx_i\}$
1. Construct a mapping from feature to instances.
2. Initialize the centroid for new cluster $\{C_j \mid 1 \le j \le k\}$
3. repeat
4. Reset $\{MaxSim_i\}$, $\{idx_i\}$
5. for j=1: k
6. identify relevant instances S_j to centroid C_j
7. for *i* in S_j
8. Compute $sim(i,C_j)$ of instance *i* and C_j
9. if $sim(i,C_j) > MaxSim_i$
10. $\text{MaxSim}_i = sim(i,C_j)$
11. $idx_i = j$
12. for i=1: m
13. update centroid $Cidx_i$
14. until change of objective value < ∈

The k-means algorithm which maximizes within cluster similarity Shown in

$$\arg\max_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \frac{x_j \cdot \mu_i}{\|x_j\| \|\mu_i\|},$$

Where k is the number of clusters, $S = \{S_1, S_2, \ldots, S_k\}$ is the Set of clusters, and μ_i is the centroid of cluster S_i. In algorithm value of *MaxSim* represents the maximum similarity between one data instance and a centroid.

For cluster every iteration first identify relevant instance to centriod. This avoids the iteration for each instance and each centriod. This clustering algorithm is able to partition its edges into disjoint sets. This k-means is applicable for large dynamic nature network.

C. *Regularization on Communities*

This instance matrix is given as input to k-means. The k-means clustering algorithm generates cluster similarity. Then apply the regularization method i.e. support vector machine (SVM) to classify network. The connection between larger communities is weaker. So we can build an SVM relying more on communities of smaller sizes by modifying SVM objective function.

$$\min \lambda \sum_{i=1}^{n} |1 - y_i(\mathbf{x}_i^T \mathbf{w} + b)|_+ + \frac{1}{2}\mathbf{w}^T \Sigma \mathbf{w},$$

This study shows that the algorithm can be interpreted as an iterative latent semantic analysis process, which allows for extensions to handle networks with actor attributes and within-mode interactions. Experiments on both synthetic data and real world networks demonstrate the efficacy of our approach and suggest its generality in capturing evolving groups in networks with heterogeneous entities and complex relationships.

IV. Experiments and Results

Experiment is performed with the synthetic dataset, where we can change the number of node dynamically. The following table shows statistics on sample node.

TABLE II
Statistics of Synthetic DataSet

Data	**Dataset with 500 Nodes**	**Dataset with 300 Nodes**
Nodes(n)	500	300
Edges(m)	127268	45755
Number of Cluster (K)	50	50
Time (ms)	138	86

On the basis of experiment we can say if total number of clusters change then it does not effect on execution time. But if the node size is increased then it requires more time to execute the program.

V. Conclusions

The Prediction of Collective Behavior (PCB) System predicts collective behavior of people using the multimode data. By applying edge-centric clustering algorithm, very efficiently we can predict the label of unobserved individual. This algorithm also processed the multiple social dimensions and improved the performance of classification. This approach solves the scalability issue of social network. Also gives feasible solution for prediction of collective behavior of unobserved user.

ACKNOWLEDGMENT

P. K. Deshmukh thanks Prof.R.M.Gawande for his valuable suggestions and comments to improve the quality of this paper. I am also very much thankful to all those who indirectly helped in preparing this paper successful.

REFERENCES

S[1] Lei Tang al., Scalable Learning of Collective Behavior, IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 6, June 2012.

[2] L. Tang and H. Liu, Relational Learning via Latent Social Dimensions, KDD 09: Proc. 15th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining, pp. 817-826, 2009.

[3] L. Tang and H. Liu, Toward Predicting Collective Behavior via Social Dimension Extraction, IEEE Intelligent Systems, vol. 25, no. 4, pp. 19-25, July/Aug. 2010.

[4] M. Newman, Finding Community Structure in Networks Using the Eigenvectors of Matrices, Physical Rev. E (Statistical, Non- linear, and Soft Matter Physics), vol. 74, no. 3, p. 036104, http:// dx.doi.org/10.1103/PhysRevE.74.036104, 2006.

[5] M. McPherson, L. Smith-Lovin, and J.M. Cook, Birds of a Feather: Homophily in Social Networks, Ann. Rev. of Sociology,vol. 27, pp. 415-444, 2001.

[6] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature, vol. 435, pp. 814-818, 2005.

[7] S.A. Macskassy and F. Provost, Classification in Networked Data: A Toolkit and a Univariate Case Study, J.Machine Learning Research, vol. 8, pp. 935-983, 2007.

[8] S. Fortunato, Community Detection in Graphs, Physics Reports, vol. 486, nos. 3-5, pp. 75-174, 2010.

[9] S. Gregory, An Algorithm to Find Overlapping Community Structure in Networks, Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 91-102, 2007.

[10] T. Evans and R. Lambiotte, Line Graphs, Link Partitions, and Overlapping Communities, Physical Rev. E, vol. 80, no. 1, p.16105,2009.

[11] S.A. Macskassy and F. Provost, A Simple Relational Classifier, Proc. Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2003.

[12] M. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks, Physical Rev. E, vol. 69, p. 026113, http://www.citebase.org/abstract?id=oai:arXiv.org:cond-at/0308217,2004