ISSN:0975-9646

A Review on Prediction of Collective Behavior

Pranjali Deshmukh[#1], R.M.Gawande[#2]

[1,2]*MCOERC, Nashik, Maharashtra, India.*
Department of Computer Engineering,
Affiliated to Savitribai Phule Pune University.

***Abstract*—The social media provide platform to connect peoples with each other and they can share their views, idea, thoughts. Now a day lot of data are generated by social media. The heterogeneous peoples are like to connect with each other through the social sites such as Facebook, Twitter, Flickr, and YouTube etc, to classify this large data is also a challenging task. This paper is to study of predict collective behavior of individuals on the social network. For classification of large network into specific communities by consider pattern of social behaviour. The k-means variant edge-centric clustering algorithm is used for finding the communities of people. To classify the large network we are going to use the edge-centric clustering technique to extract the sparse social dimensions from network. The social dimension based approach has been shows beneficial for the study. With this approach we can handle millions of user data to demonstrate behavioural approach of user in social sites for marketing purpose.**

***Keywords*— Collective behavior, multimode, edge-centric, social dimensions, scalable learning**

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.

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. The problems with social sites for predicting behaviors are the users are not homogenous; the 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].

Collective behavior of users according to homophily [5] is, we are more likely to connect to others who share certain similarities with us. Social media provides facility to connect with each other, according to homophily we can say that through networking sites related friends behave similarly. Consider as an example, mostly we like to buy those things which our friends buy, without more investigation of those things.

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. The edge-centric view generates the large instances of network. For its regularization Linear Support Vector Machine (SVM) can be used. SVM work linearly, it can apply on large instances 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.

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 distributed data. To find label of every node it depends on attribute of its neighbor node. 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 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 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 the dense social dimensions are difficult to handle. Practically to handle the million node data is difficult. 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 finding 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 share 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 overlapping community by recursively removing edges between the graph until it divides into different communities. This method only removed that edges which creates and bridge with communities. But it gives output as only non-overlapping communities. To overcome the problem S. Gregory [9] also handles overlapping communities with node (instead of edge). The algorithm recursively splits nodes that 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 in 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. In this system, by considering multi-attribute edge-centric clustering issue, viewing actor attributes such as tags, comments, inlink, outlinks etc. as features for clustering.

III. METHODS

For classification of large network various methods are used that methods are discus in this section.

A. SocioDim with modularity maximization

The Social dimensions extracted according to soft clustering, such as modularity maximization and probabilistic methods, are thick [4]. The drawback with Modularity maximization requires us to compute the top eigenvectors of a modularity matrix, which is the same size as a given network. The Social networks grow, with new members joining and new connections occurring between existing members each day. This dynamic nature of networks entails an efficient update of the model for collective behavior prediction. Efficient online updates of eigenvectors with expanding matrices remain a challenge.

B. Edge Partition via Line Graph Partition

A simple method construction of line graph it represent the whole network with line graph. When we consider large network then formation of cycles in increased. Also the line graph requires all nodes of same degree. But practically it's not possible that every node have same affiliation with other node.

C. Edge-centric clustering technique

The multimedia network generates large data. For processing network is represent in the form of graph where user consider as node and connection link between them consider as edge G(V,E). By considering all attribute of node data is generated with large size, so network handling is issues for previous techniques [2][10]. By considering only link (edge) between two nodes is better for handling large network. Generate an instance based matrix of network.

For this study, consider user as a node to 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].

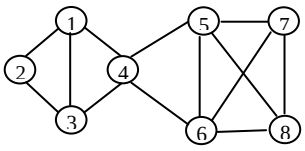

Fig.1. A Toy example

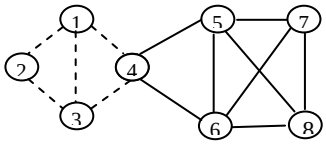

Fig.2. Edge clusters

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. 1 shows a Toy network; divide the network into disjoint sets. The dashed edges represent one affiliation of 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.

IV. PROPOSED SYSTEM

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

The proposed system can also handle the users' other social behavioral activities like comment, inlink, outlink and length of the blog. To find sparse social dimensions, considering all these social activities and apply the edge-centric clustering algorithm on the blogger dataset to find the most influenced blog. Also find the influenced blog on the basis of number of comments, number of likes, inlink, outlink of the blog etc. and predict the behavior of unobserved user. Advantage of this system is, a scalable approach having capacity to handle scalable network, where earlier models failed to handle millions of user network data. The following diagram shows a framework of scalable learning Fig. 3.

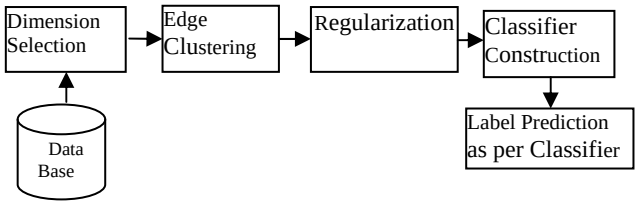

Fig 3. The formwork of Prediction of Collective Behavior

To perform operation by considering all attribute of node is not possible, so extract the meaningful dimensions. The work flow is first select the required dimensions of user from data base. Then generate an edge centric view of data (refer section C). After creation of edge-centric view of given network, the sparse instance based matrix is generated. 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 find sparse communities. 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.

In this work, try to address the scalability problem by employing the other activities like inlink, outlink,comments etc. information to analyze a multi-mode network. A framework and its convergence property are carefully studied. To show 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 realworld networks demonstrate the efficacy of our approach and suggest its generality in capturing evolving groups in networks with heterogeneous entities and complex relationships.

V. CONCLUSIONS

The Scalable Learning System predicts collective behaviour of people using the multimode data. By applying edge-centric clustering algorithm, handle multimode data. This algorithm also processed the multiple social dimensions like tags, comments, inlink, outlink and length of blog and improved the performance of classification. This approch solves the scalability issue of social network. This method presents a feasible solution for prediction of collective behavior of unobserved user. Future direction is needed to find dynamic attributes of user for prediction of user label.

ACKNOWLEDGMENT

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

REFERENCES

[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-mat/0308217,2004.