RESEARCH ARTICLE OPEN ACCESS

An Approach to Extract Social Dimensions in Social Media Data Based On Modularity

T.P.R.S.Lahari [1], Somesh Katta [2]
Department of CSE [1], Assistant Professor [2]
VITAM COE, Visakhapatnam
India

ABSTRACT

Social media such as Instagram, Flickr, Facebook, Twitter and blogs etc., presents data in a linked format rather than independent and identical distributed (i.i.d) fashion. To deal with the interdependency among data instances, relational learning has been proposed, and collective inference based on network connectivity is adopted for prediction. However, the connections or relations in social media are often multi-dimensional. A user can connect to another user through various factors, e.g., belongs to same family, classmates, living in the same country or town or sharing similar interest, etc. Many collective inference methods normally do not differentiate these links or relations. In this paper, a relational learning approach is proposed to extract social dimensions based on network linked information first, and then utilize them as features for discriminative learning. These social dimensions describe different affiliations or collaborations of social network users hidden in the network, and the subsequent discriminative learning can automatically determine which affiliations are better aligned with the class labels. Such a scheme is preferred when multiple diverse relations are associated with the same network. The proposed method is tested with real world social media data and evaluated its performance. The results show that the dimensions generated by the proposed method are promising to predict the behaviour of a new user.

Keywords:- Social Dimensions, Affiliations, Modularity, Discriminative Learning, Precision and recall.

I. INTRODUCTION

Large complex graphs representing relationships among sets of nodes are importantly a common focus of network analysis. Examples include social networks, blogs, Web graphs, telecommunication networks, semantic networks etc. A real fundamental problem related to such networks is the discovery of affiliations or communities as shown in figure 1. Intuitively, an affiliation or community refers to a collection of individuals with dense connections patterns internally and sparse connections externally.

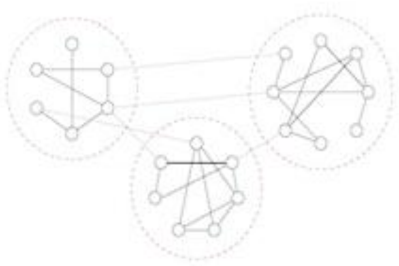


Figure 1: A small network with community structure

The study of community structure in networks has its roots from graph partitioning. It is closely related to the ideas of graph partitioning in graph theory and computer science, and hierarchical clustering in sociology.

Now-a-days a vital aspect of Social Network Analysis is to predict the behaviour of individual users based on collective inference obtained from affiliations. Once the community structure of a group of labeled or unlabeled networked nodes is obtained, based on collective inference of the community structure unlabeled nodes can be labeled. It appears like a simple conventional data problem but the challenge here is to deal with network of related or connected instances rather than independently identically distributed (IID) instances in conventional data mining. So the key aspect is to leverage the social network data for accurate classification when limited labeled instances are available.

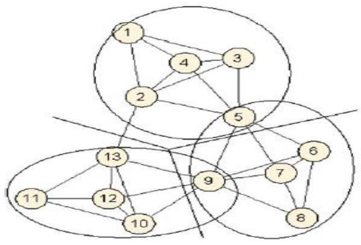


Figure 2: Different affiliations with Node 5 and Node 9

With limited information and the network connectivity, categorizing connections into different affiliations is not an easy task as the same user is involved in multiple affiliations as shown in figure 2. Moreover, the same connection can be associated with multiple affiliations. Instead of identifying affiliations among actors via categorizing connections directly, a collective inference mechanism which considers all plausible affiliations among the users in community detection process is needed.

In this work, a relational learning framework based on social dimensions is proposed. Each social dimension can be considered as the measure of a likely affiliation between users. The extracted social dimensions can used for discriminative learning such as SVM or logistic regression. A discriminative learning approach may be employed to select the relevant social dimensions for classification automatically. In the prediction phase, different from existing relational learning methods, collective inference becomes unnecessary as the selected social dimensions have already included the relevant network connectivity information. This proposed framework is flexible and it is different from existing relational learning works, which mainly concentrate on entity resolution, web page or publication classification, we specifically focus on classification associated with social media where the network is noisy and typically has a composite of multiple relations among actors.

II. LITERATURE REVIEW

In social network analysis, relational learning refers to the classification when users affiliated to multiple affiliations. In this paper, classification in network data is studied. The data instances in a social network are not independently identically distributed (IID) as in traditional data mining scenarios. In order to get the autocorrelation between labels of neighbouring users, a Markov dependency assumption is applied. That is, in a network, labels of one node depend on the labels (or attributes) of its neighbouring nodes. D. Jensen et al proposed a collective inference approach for prediction. In traditional data mining approach, a classifier is constructed based on the relational features of labelled data, and then an iterative process is required to determine the class labels for the unlabelled data. Q. Lu et al showed that a simple weighted vote relational neighbourhood classifier works well on some real world benchmark social media data. X. Zhu and Z. Gahramani et al. proved that this method is related to Gaussian field for semi-supervised learning on graphs.

Majority of the relational classifiers are designed based on Markov assumption and can capture only the local dependencies. To capture the long-distance autocorrelation, new models are proposed based on user's cluster membership. But the model needs high computational cost for inference, which hinders their direct application to large networks. So Neville and Jensen et al. proposed a hard clustering algorithm to find the hard cluster membership of each user, and then fix the latent group variables for later inference. Since the social media networks are not homogeneous, some nodes do not show a strong affiliation membership and hard clustering might assign them randomly. The affiliation structure may change drastically even with the removal of one single edge or node in the network. Lei Tang et al. proposed a relational learning approach based on social dimensions where social dimensions are represented as continuous values and allow each node to involve at different dimensions in a flexible degree in conjunction with the discriminative classifiers.

Now-a-days, community detection has been an important field in social network analysis and various methods have been proposed including stochastic block model, latent space model, spectral clustering, and hierarchical clustering based on various measures such as shortest-path betweenness or modularity. In this paper, modularity procedure is used for soft community detection to extract social dimensions.

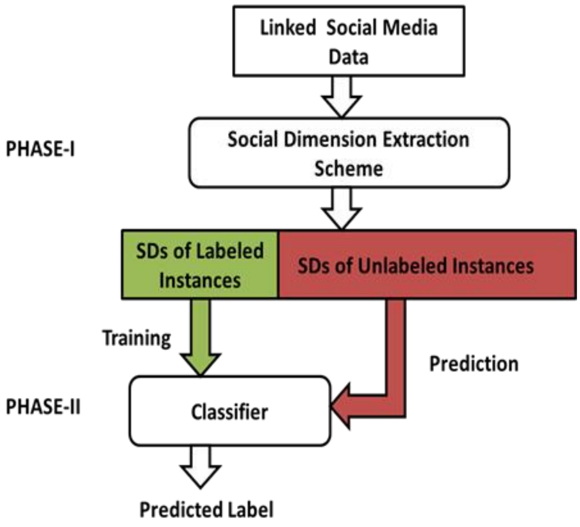
III. METHODOLOGY

In social media, individuals are highly idiosyncratic. Each user's interest cannot be captured by a single class label. A user may participate in several affiliations. In this paper, rather than looking at univariate cases for classification in social network data (a node has only one class label), much attention is paid on more challenging tasks that each node in a network may have multiple labels. This problem can be formulated formally as

Suppose there are K class labels $y = \{c_1, c_2, ..., c_k\}$.Given network $G = (V, E, Y)$ where V is the vertex set, E is the edge set and $Y_i \subseteq y$ are the class labels of a vertex $v_i \in V$, and given known values of Y_i for some subsets of vertices V^L,how to infer the values of Y_i (or a probability estimation score over each label) for the remaining vertices $V^U = V - V^L$?

Social media connections are not homogeneous. Users can connect to their family, colleagues, college class mates, or some other users met online. Some of these relations are helpful to determine the targeted behaviour (labels) but not necessarily always so true.

Figure 3: Proposed Methodology



In a social network, people are involved in different affiliations and connections are emergent results of those affiliations. These affiliations have to be differentiated for behaviour prediction. However, the affiliation information is not readily available in social media. Direct application of collective inference or label propagation treats the connections in a social network homogeneously. This is especially problematic when the connections in the network are heterogeneous. To address the connection heterogeneity, several methods based on social dimensions are devised for collective behaviour learning.

The proposed framework (figure 3) involves two steps:

Phase **I**: Social dimension extraction, and
Phase **II**: Discriminative learning.

PHASE-I

Social Dimension Extraction

In the first step, social dimensions are extracted based on network structure to capture the potential affiliations of users. These extracted social dimensions represent how each actor is involved in diverse affiliations. In this frame work a new community measure known as modularity is used to extract social dimensions. Modularity directly takes the degree distribution into consideration and stands as an effective community measure in many complex network structures

Modularity measures how far the interaction deviates from a uniform random graph with same degree distribution. It is defined as

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{ij}-\frac{d_i d_j}{2m}\right]\delta(s_i,s_j).$$

where A_{ij}, is represents Adjacency matrix of the social network nodes, d_i and d_j represent degree of two nodes(v_i and v_j), m represents total number of edges of the network and s_i and s_j represent community membership of vertices v_i and v_j respectively, and $\delta(s_i,s_j)=1$ if s_i and s_j belong to same affiliation.

A higher value of modularity indicates higher degree of interaction with in a community. To find the communities of higher degree of interaction one has to maximize Q. But maximizing the modularity over hard community detection is a NP-hard problem. With some relaxation on the problem it can be solved efficiently. i.e. problem in discrete domain is changed to continuous domain. Then the modularity is reformulated as

$$Q=\frac{1}{2m}Tr\left(S^T BS\right)$$

where B represents the modularity matrix defined as

$$B=A-\frac{d.d^T}{2m}$$

d = degree matrix of nodes

$S\in\{0,1\}^{n\times k}$ is a community indicator matrix defined as

$$S_{ij}=\begin{cases}1 & if\ vertex\quad i\quad belongs\ to\ community\\ 0 & otherwise\end{cases}$$

Since S is discrete, by relaxing S to be continuous, optimal S is the $top-k$ Eigen vector of the modularity matrix.

PHASE-II

Construct Discriminative Classifier

The above social dimensions can be treated as features of users for the subsequent discriminative learning. Since the entire network is converted into features, any supervised classifier that suits the requirements such as support vector machine, decision tree or logistic regression can be employed.

This step is critical as the classifier will determine which dimensions are relevant to a class label. More powerful methods like structural SVM, decision tree can also be employed. Once the classifier is ready, prediction can be done easily, since the latent social dimensions have been calculated for unlabeled data in Phase-I. Note that collective inferences not required for prediction.

IV. EXPERIMENTAL RESULTS

In this section, the performance of proposed learning method is examined. The proposed frame work is implemented using MATLAB.

A. Datasets used

BlogCatalog: Table I lists some statistics of the network data. As seen in the table, the connections among the social actors are extremely sparse. The degree distribution is highly imbalanced, a typical phenomenon in scale-free networks. Both data sets are available from the first author's homepage.

TABLE I: STATISTICS OF SOCIAL NETWORK DATA

Parameter	BlogCatalog
Categories(k)	30
Users(n)	10,312
Links(m)	333,989
Density	$6.3\text{x}10^{-3}$
Maximum Degree	3,992
Average Degree	65
Average Labels	1.4

B. *Performance Metrics used*

A user may belong to more than one social dimension. Therefore a thresholding process is required to generate ranking of labels. It was shown that various thresholding methods lead to quite different performance. To overcome the concern, it is assumed that the number of labels on the test date already known and it is to check how the top predictions match with the true labels. The quality of prediction is evaluated by two commonly used Multiclass Metalabeler metrics, Micro F_1 and Macro F_1

Macro-F1 is the F1 averaged over categories.

$$Macro-F1=\frac{1}{K}\sum_{k=1}^{K}F_1^k$$

For a category C_k the precision (P_k) and the recall (R_k) are calculated as,

$$P^k=\frac{\sum_{i=1}^{N}y_i^k\hat{y}_i^k}{\sum_{i=1}^{N}\hat{y}_i^k},$$

$$R^k=\frac{\sum_{i=1}^{N}y_i^k\hat{y}_i^k}{\sum_{i=1}^{N}y_i^k}.$$

Then *F1* measure, defined as the harmonic mean of precision and recall is computed as follows:

$$F_1^k=\frac{2P^kR^k}{P^k+R^k}=\frac{2\sum_{i=1}^{N}y_i^k\hat{y}_i^k}{\sum_{i=1}^{N}y_i^k+\sum_{i=1}^{N}\hat{y}_i^k}$$

Micro *-F1* is computed using the equation of F_1^k and considering the predictions as a whole. More specifically, it is defined as,

$$Micro-F1=\frac{2\sum_{k=1}^{K}\sum_{i=1}^{N}y_i^k\hat{y}_i^k}{\sum_{k=1}^{K}\sum_{i=1}^{N}y_i^k+\sum_{k=1}^{K}\sum_{i=1}^{N}\hat{y}_i^k}.$$

According to the definition, macro-F1 is more sensitive to the performance of rare categories while micro-F1 is affected more by the major categories. During performance evaluation, both the measures are examined carefully.

C. *Performance Evaluation*

Table II presents the performance of various approaches for the BlogCatalog data. We gradually increase the number of labelled nodes from 10% to 90%. For each setting, we randomly sample a portion of nodes as labelled. This process is repeated 10 times and the average results are reported.

TABLE II: PERFORMANCE ON BLOGCATALOG DATA

Training Ratio (%)	Micro-F1 (%)	Macro-F1 (%)
10	27.35	17.36
20	30.74	20.00
30	31.77	20.80
40	32.97	21.85
50	34.09	22.65
60	36.13	23.41
70	36.08	23.89
80	37.23	24.20
90	38.18	24.97

V. CONCLUSION AND FUTURE SCOPE

Social media provides a virtual social networking environment. The classical IID assumption of data instances is not applicable for effective social media analysis. A framework relational learning framework based on social dimensions is proposed. Based on the extracted social features using modularity, a discriminative classifier is employed to determine which dimensions are informative for classification. Experimental results on real world social media data demonstrated that the proposed social dimension approach performs well, especially when the labelled data are few. Further investigation is needed to address the some concerns related to modularity maximization such as Eigen-value problem as well as concerns related relational learning when social network is very dynamic in nature.

REFERENCES

[1] J.Tang and H.Liu. Unsupervised Feature Selection for Linked Social Media Data. *KDD'12* August 12-16, 2012.

[2] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.

[3] L. Tang and H. Liu. Scalable Learning of Collective Behavior Based on Sparse Social Dimensions. *CIKM'09*, November 2-6-2009, Hong Kong.

[4] L. Tang and H. Liu. Relational Learning via Latent Social Dimensions. *KDD'09,*June 28-July 1,2009.

[5] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physics Revies E(Statistical, Nonlinear, and Soft Matter Physics)*, 74(3),2006.

[6] M. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8582, 2006.

[7] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[8] Y.Zhang. Community Detection Methods Using Eigenvectors of Matrices.

[9] Rong-En Fan and Chih-Jen Lin. A Study on Threshold Selection for Multi-Label Classification.

[10] H.Liu and L.Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering.

[11] L.Tang, S.Rajan and V.K.Narayanan. Large Scale Multi-Label Classification via MultiLabeler. In the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use and personal use by others. *WWW 2009*, April 20–24, 2009.

[12] L.Tang and H.Liu. Levaraging social media networks for classification. 10 August 2009 / Accepted: 15 December 2010.

[13] Xufei Wang, Lei Tang and Huiji Gao and Huan Liu , BlogCatlog Dataset "http://dmml.asu.edu/users/xufei/datasets.html"

[14] Vincent Van Asch. Macro- and Micro- averaged evaluation measures [[BASIC DRAFT]]. September 9, 2013.