基于流形学习的社会化媒体网络数据分类*

史仍浩[a], 陈秀真[b], 李生红[a]
(上海交通大学 a. 电子工程系; b. 信息安全工程学院, 上海 200240)

摘 要: 社会化媒体提供了海量的、大尺度的异质网络数据,如何对网络数据进行分类是一个亟待解决的新问题。基于潜在社会维模型,提出利用流形学习中的拉普拉斯特征映射算法进行社会维抽取。实验表明,在精确匹配率、微平均、宏平均等性能指标上,均优于基于模块度最大化的原有社会维模型。该算法能更好地获取用户的隐性联系,从而更好地分析网络用户行为。

关键词: 流形学习; 拉普拉斯特征映射; 社会化媒体; 网络数据分类; 多标签
中图分类号: TP181; TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)03-0692-03
doi:10.3969/j.issn.1001-3695.2013.03.012

Networked data classification in social media based on manifold learning

SHI Reng-hao[a], CHEN Xiu-zhen[b], LI Sheng-hong[a]
(*a. Dept. of Electronic Engineering, b. School of Information Security Engineering, Shanghai Jiaotong University, Shanghai* 200240, *China*)

Abstract: Social media provided massive, large-scale heterogeneous networked data. Classification in networked data is a new problem that needed to be solved. Based on latent social dimension model, this paper proposed using Laplacian eigenmaps from manifold learning to extract social dimensions . Experiments show that it is superior to original modularity maximization social dimension model in performance metrics like exact match ratio, micro average and macro average. The algorithm can capture implicit user relations better and analysis Web user behavior better.

Key words: manifold learning; Laplacian eigenmaps; social media; networked data classification; multi-label

0 引言

随着互联网的飞速发展和 Web 2.0 时代的来临,在线论坛、博客、社交网络、内容分享社区、微博客、维基等各种类型的社会化媒体(social media)[1]不断涌现,网络社会化的趋势不断加强。传统 Web 网站的传播方式是单向的,大部分用户被动地接受少数权威或专家提供的内容。而在社会化媒体环境下,普通用户既是内容的消费者,又是内容的制造者,而且丰富的用户交互形成了一个社会网络。

数据分类是数据挖掘的一个基本问题。传统的算法假设数据样本是独立同分布的,只根据数据自身的属性进行分类[2]。而社会化媒体产生的海量、大尺度的数据是可以用图表示的网络数据[3]。网络数据分类不仅要考虑各数据样本自身的属性,还要考虑它们之间的相互关系。目前,网络数据分类的主要研究进展都是基于统计关系学习或概率图模型进行协作分类(collective classification)[4]。这些算法的缺点是认为节点之间关系是单一的、同质的,而实际上网络中用户的关系是多关系、异质的。以 Facebook 中的某一用户来说,他(她)的好友中可能有研究生同学、大学同学、中小学同学等。在大多数情况下,难以获取社会化媒体网络中用户的隐性关系。社会维模型[5,6]很好地解决了上述问题,该模型主要分为两步:先用软聚类方法进行社会维抽取,再把社会维作为潜在特征进行分类。在社会网络中,具有相似社会关系的用户会形成一个社区,社会维的抽取可看做是一个软聚类社区发现问题。此外,网络中用户的行为和兴趣也是多样的,可用多个标签来表示,因此是个多标签分类(multi-label classification)问题[7]。为此,本文基于潜在社会维网络数据多标签分类模型,提出一种利用拉普拉斯特征映射进行社会维抽取的改进算法。实验表明该方法的效果优于采用模块度最大化的基本社会维模型。

1 基于流形学习的网络数据分类

1.1 网络数据多标签分类的形式化描述

给定:

a) C 个类别 $C=\{C_1,C_2,\cdots,C_c\}$;

b) 一个体现了节点交互关系的网络 $V=G(V,E,C)$,其中 V 是顶点集合,E 是边集合,每个顶点 V_i 都对应一个标签集合 C;

c) 全体顶点集合 V 的一个子集 V^L 的标签已知,其中 $V^L\in V$,$C_{ij}\in\{+,-\}$ 代表节点 V_i 在类别 C_j 上的标签信息;

求:

d) 剩下的顶点集合 $V^U=V-V^L$ 对应的未知的标签 C^U。

1.2 基于流形学习的社会维模型

1.2.1 算法设计

正如前面所述,真实的社会化媒体网络都是多关系、异质的,不同的社会关系体现了用户潜在的社会维度,可以用代表用

收稿日期: 2012-07-18; **修回日期**: 2012-08-23 **基金项目**: 国家"973"计划资助项目(2010CB731403,2010CB731406);国家自然科学基金资助项目(61071152,61271316);国家"十二五"科技支撑计划重点项目(2012BAH38B04)

作者简介: 史仍浩(1988-),男,硕士研究生,主要研究方向为在线社会网络分析(srh08@126.com);陈秀真(1977-),女,副教授,博士,主要研究方向为网络安全、在线社交网络分析;李生红(1971-),男,教授,博士,主要研究方向为网络内容安全、网络集群行为分析.

户在每种社会关系上的参与情况的向量来表示社会维。假设网络中共有 k 种社会关系,那么顶点 V_i(或说用户 U_i)的社会维可用一个 k 维的向量表示,即 $S_i = \{S_{i1}, S_{i2}, \cdots, S_{ij}, S_{ik}\}, 1 \leqslant j \leqslant k$。整个网络的社会维可以用矩阵 $S = \{S_{ij}\} \in \mathrm{R}^{n \times k}$ 来表示。

基于流形学习的社会维模型的整体框图如图 1 所示。

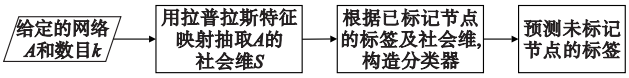

图1　基于流形学习的社会维模型框图

模型的第一步是利用软聚类方法抽取社会网络的潜在社会维,是算法的关键部分。流形是微分几何、拓扑学中的一个基本概念,可以简单地理解为由 $f(x) = 0, x \in \mathrm{R}^d$ 确定的 R^N ($d << N$)空间中的 d 维曲面。流形学习的主要目标[8]是根据有限的离散样本数据学习和发现嵌入在高维空间中的低维光滑流形。Belkin 等人[9]提出的拉普拉斯特征映射是流形学习的一种典型方法,其基本思想是在高维空间中离得很近的点投影到低维空间中的像也应该离得很近。通过保持数据的局部信息,该算法隐式地强调了数据的聚类特性,可以解释为一种软聚类算法。

与此同时,该算法避免了模块度自身定义带来的分辨率极限的缺点[10],模块度最大化方法难以发现小规模的聚类,严重影响算法的实际应用。实际网络中的各个社区往往是大小不一的,有些小规模的社区可能没有被检测出来。模块度对网络中个别的几个连接也十分敏感,这可能会导致错误的社区划分结果。因此,本文提出利用流形学习中的拉普拉斯特征映射进行社会维的抽取。

模型的后两步是用抽取出的社会维作为特征,然后构造分类器对未标记的数据进行分类。与文献[5,6]一样,分类器采用的是一对多(one-against-all)线性 SVM(http://www.csie.ntu.edu.tw/~cjlin/liblinear/)。

1.2.2　拉普拉斯特征映射

算法的目标是找到一个在平均意义上保留数据点局部特性的映射,即最小化目标函数 $L(f) = \frac{1}{2}\sum_{i,j}(f_i - f_j)^2 W_{ij}$,其中 L 表示拉普拉斯—贝尔特拉米算子。根据谱图理论[11],如果数据均匀取样于高维空间中的低维流形,那么流形上的拉普拉斯—贝尔特拉米算子可以由图的拉普拉斯算子逼近。

由于网络数据本身就有图结构,因此不再需要原算法中构造邻接图的一步。这样算法的具体步骤变为:

a)建立加权邻接矩阵 W。有热核法和简单法两种,定义如下:

$$W_{ij}(\text{热核法}) = \begin{cases} \mathrm{e}^{-|x_i - x_j|^2/t} & t \in \mathrm{R}, \text{如果节点 } i \text{ 和 } j \text{ 有连接} \\ 0 & \text{否则} \end{cases}$$

$$W_{ij}(\text{简单法}) = \begin{cases} 1 & \text{如果节点 } i \text{ 和 } j \text{ 有连接} \\ 0 & \text{否则} \end{cases}$$

实际的社会化媒体网络数据可用 n 阶方阵 $A \in \{0,1\}^{n \times n}$ 来表示,对应这一步中的简单法,即 $W = A$。

b)特征映射。即计算方程 $L_v = \lambda D_v$ 的广义特征值和特征向量,其中 D 为对角矩阵,$D_{ii} = \sum_j W_{ji}$,L 为图的拉普拉斯矩阵,$L = D - W$。设求得的 $d+1$ 个最小的特征值对应的特征向量为 $v_0, v_1, \cdots, v_d$,则映射后的低维空间可以用 $v_1, v_2, \cdots, v_d$ 来表示。这个问题最终可以归结为计算归一化图拉普拉斯矩阵 $L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2}$ 的 $d+1$ 个最小特征值[9,12]。

可以用一个如图 2 所示的简单示例($d = 1$)来说明如何利用拉普拉斯特征映射进行社会维抽取。

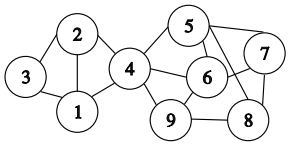

图2　示例

其社会维矩阵($d+1$ 个特征向量)为

$$S = \begin{bmatrix} -0.3162 & -0.4552 \\ -0.3162 & -0.4552 \\ -0.2582 & -0.4643 \\ -0.4082 & -0.0893 \\ -0.3651 & 0.2748 \\ -0.3651 & 0.2415 \\ -0.3162 & -0.3144 \\ -0.3162 & 0.3077 \\ -0.3162 & 0.1864 \end{bmatrix}$$

矩阵的第一列没有社区划分信息,它对应的是最小的特征值 0。而从第二列可以发现两个社区或者两种社会关系:节点{1,2,3}对应的数值是十分接近的负数,为一个社区;节点{5,6,7,8,9}对应的数值则是十分接近正数,为另一个社区;而节点 4 的数值在上述两者之间(接近于 0),所以该节点可以划分到两个社区或者说拥有两种社会维度。这一结果与谱图理论相一致:图的拉普拉斯矩阵的第二小的特征值对应的特征向量能把图 2 分成两个子图。从这个例子中可以看出,本文提出的模型能够在一定程度上获取用户的多种社会关系。

2　实验

2.1　实验数据

本文的实验数据采用 Tang 等人[5,6]所在的研究组提供的数据集(http://www.public.asu.edu/~ltang9/social_dimension.html)。该数据集包括博客网站 BlogCatalog 和照片内容分享网站 Flickr 两个真实的社会化媒体网络。BlogCatalog 网络中的用户在发表博文时一般会加上一个或多个的类别信息,而 Flickr 网络中的用户常常会加入一些不同的兴趣小组。BlogCatalog 数据包含 39 个博文类别,Flickr 数据则有 195 个兴趣小组。本文把这两种信息分别作为多标签分类问题中的类别信息。两个网络的一些统计特性如表 1 所示。

表 1　数据集

属性	BlogCatalog	Flickr
类别数 C	39	195
用户数 n	10 312	80 513
链接数 m	333 983	5 899 882
网络密度	6.3×10^{-3}	1.8×10^{-3}
最大度	3 992	5 706
平均度	65	146
平均标签	1.4	1.3

2.2　评测标准

精确匹配率(exact match ratio,EMR)是评价多标签分类性能的一种指标,是对二分类中常用的正确率的扩展。对给定的 N 个测试数据,设 $y_i^k, \hat{y}_i^k \in \{0,1\}^K$ 分别表示实体 x_i 实际的标签和预测的标签。精确匹配率的计算方法为

$$\mathrm{EMR} = \frac{1}{N}\sum_{i=1}^{N} I[y_i^k = \hat{y}_i^k], \text{其中 } I[s] = \begin{cases} 1 & s \text{ 为真} \\ 0 & s \text{ 为假} \end{cases}$$

精确匹配率没有考虑多标签分类中的部分匹配问题,即一个数据的部分标签的分类结果是正确的,它把这种情况当成是错误的。考虑了这一问题的指标是二分类中的 F_1 值在多标签分类中的扩展,即微平均 F_1 值和宏平均 F_1 值。

宏平均 F_1 值是每一个类别的性能指标 F_1 值的算术平均值

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

其中:类别 C_k 的准确率 P^k、召回率 R^k、F_1 值 F_1^k 分别为

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

而微平均 F_1 值是每一个实体的性能指标 F_1 值的算术平均值:

$$\text{Micro-}F_1 = \frac{2\sum_{k=1}^{K}\sum_{i=1}^{N} y_i^k \hat{y}_i^k}{\sum_{k=1}^{K}\sum_{i=1}^{N} y_i^k + \sum_{k=1}^{K}\sum_{i=1}^{N} \hat{y}_i^k}$$

2.3 实验结果

Tang 等人建议该数据集的潜在社会维度 k 取值在 400 ~ 600 之间比较理想[5],本实验中把它设定为 500。将本文提出的基于拉普拉斯特征映射这一流形学习的方法与基于模块度最大化的基本社会维模型进行对比。对于中小规模 BlogCatalog 数据集(10 000 个节点左右),实验时把训练节点的比例(TR)从 10% 逐步增加到 90%;而对于较大规模的 Flickr 数据集(80 000 个节点左右),标签信息是十分有限的,所以实验中训练节点的范围是 1% ~ 9%。选取训练集的方法是随机抽样。

在 BlogCatalog 和 Flickr 数据集上的实验结果分别如表 2、3 所示。可以发现,在精确匹配率、微平均、宏平均等多标签分类性能指标上,使用拉普拉斯特征映射(LE)进行社会维抽取的效果均优于模块度最大化(MM)方法。与模块度最大化相比,拉普拉斯特征映射法能够更精确地捕获到网络数据中潜在的社会关系。由此也可以看出,社会维的抽取方法是整个社会维模型的关键。另外还可以发现,对于三个多标签分类指标,不管是模块度最大化还是拉普拉斯特征映射,微平均 F_1 值最大,而宏平均 F_1 值比没有考虑部分匹配的精确匹配率还小。这是因为计算宏平均 F_1 值时认为每一个类别拥有相同的权重,即假设类别是平衡的,而实际的数据如本文的两个数据集,是不平衡数据。

表 2 在 BlogCatalog 数据集上的性能

TR/%	EMR		Micro-F_1		Macro-F_1	
	MM	LE	MM	LE	MM	LE
10	0.212 5	0.229 0	0.269 8	0.287 1	0.169 8	0.204 4
20	0.241 8	0.269 6	0.302 7	0.333 8	0.200 9	0.237 2
30	0.254 8	0.298 4	0.313 1	0.360 4	0.204 5	0.252 3
40	0.258 4	0.318 7	0.325 5	0.385 6	0.213 5	0.271 7
50	0.271 1	0.334 0	0.340 6	0.403 3	0.223 3	0.283 1
60	0.281 0	0.344 1	0.353 6	0.413 5	0.229 5	0.287 9
70	0.301 6	0.359 2	0.374 5	0.431 3	0.240 6	0.302 3
80	0.300 7	0.373 4	0.382 1	0.450 1	0.252 3	0.307 7
90	0.291 9	0.374 4	0.370 3	0.445 3	0.257 1	0.312 6

表 3 在 Flickr 数据集上的性能

TR/%	EMR		Micro-F_1		Macro-F_1	
	MM	LE	MM	LE	MM	LE
1	0.184 0	0.190 5	0.239 9	0.294 4	0.109 9	0.154 9
2	0.196 2	0.205 9	0.252 4	0.313 0	0.129 2	0.187 0
3	0.208 8	0.215 6	0.270 5	0.325 8	0.141 4	0.193 7
4	0.214 1	0.223 4	0.279 1	0.337 5	0.150 2	0.201 3
5	0.214 8	0.236 1	0.282 0	0.3420	0.158 0	0.213 1
6	0.212 0	0.250 8	0.279 5	0.3531	0.162 1	0.217 5
7	0.215 0	0.264 9	0.284 5	0.3611	0.163 3	0.223 0
8	0.214 8	0.276 6	0.286 3	0.3686	0.166 8	0.228 4
9	0.218 2	0.293 0	0.290 9	0.3750	0.166 5	0.235 1

3 结束语

本文提出的一种利用流形学习中的拉普拉斯特征映射算法进行社会维抽取的改进模型,充分利用拉普拉斯特征映射的软聚类特性,很好地解决了社会化媒体网络的异质问题。实验表明该算法优于基于模块度最大化的基本社会维模型。

在下一步的工作中,可以把网络结构特征与节点特征(如用户个人信息、用户生成内容等)结合起来作为特征,再使用合适的不平衡数据分类算法对网络数据进行分类,以提高多标签分类的性能。

参考文献:

[1] KAPLAN A M, HAENLEIN M. Users of the world, unite! The challenges and opportunities of social media[J]. Business Horizons, 2010, 53(1): 59-68.

[2] GETOOR L, DIEHL C. Link mining: a survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.

[3] MACSKASSY S A, PROVOST F. Classification in networked data: a toolkit and a univariate case study[J]. Journal of Machine Learning Research, 2007, 8(5): 935-983.

[4] SEN P, NAMATA G, BILGIC M, *et al*. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93-106.

[5] TANG Lei, LIU Huan. Relational learning via latent social dimensions [C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 817-826.

[6] TANG Lei, LIU Huan. Toward predicting collective behavior via social dimension extraction[J]. IEEE Intelligent Systems, 2010, 25(4): 19-25.

[7] TSOUMAKAS G, KATAKIS I. Multi label classification: an overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.

[8] 罗四维,赵连伟. 基于谱图理论的流形学习算法[J]. 计算机研究与发展, 2006, 43(7): 1173-1179.

[9] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.

[10] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174.

[11] CHUNG F R. Spectral graph theory[M]. Providence: American Mathematical Society, 1997.

[12] 西奥多里蒂斯. 模式识别[M]. 4 版. 李晶皎, 王爱侠, 王骄, 译. 北京: 电子工业出版社, 2010.