基于种子节点选择的网络环境下多标签分类算法研究

吴信东[1,2],赵银凤[1],李 磊[1]
(1. 合肥工业大学计算机与信息学院,安徽合肥 230009; 2. 佛蒙特大学计算机科学系,美国伯灵顿 VT05405)

摘 要: 多标签分类在基因分类,药物发现和文本分类等实际问题中有着广泛的应用.已存在的多标签分类算法,通常都是从网络中随机的选取节点作为训练集.然而,在分类算法执行的过程中,网络中不同节点所起的作用不同.在给定训练集数目的情况下,选择的训练集不同,分类精度也会不同.所以我们引入了种子节点的概念,标签分类从种子节点开始,经过不断推理,得到网络中其他所有节点的标签.本文提出了SHDA(Nodes Selection of High Degree from Each Affiliation)算法,即从网络的每个社团中,按比例的选取度数较大的节点,然后将其合并,处理后得到种子节点.真实数据集上的实验表明,将种子节点用作训练集进行多标签分类,能够提升网络环境下多标签分类的准确率.

关键词: 多标签分类; 网络; 种子节点; 推理; 社团
中图分类号: TP181; TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)09-2074-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.09.008

Multi-label Classification in Network Environments via Seed Node Selection

WU Xin-dong[1,2], ZHAO Yin-feng[1], LI Lei[1]
(1. *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui* 230009, *China*;
2. *Department of Computer Science, University of Vermont, Burlington VT* 05405, *USA*)

Abstract: Multi-label classification is widely used in genetic classification, drug discovery and text classification. The existing multi-label classification algorithms usually select nodes randomly from the network as their training set. However, during multi-label classification, different nodes have different effects. Given the number of nodes in the training set, a different training sub-set can lead to different classification accuracy. Hence, we introduce the concept of seed nodes, the classification procedure starts from the seed nodes, and after continuous reasoning, the labels of other nodes are inferred in the network. We propose an SHDA algorithm (Nodes Selection of High Degree from Each Affiliation) in which the nodes of high degrees from each affiliation belonging to the network are selected and merged, and after processing, the seed nodes are obtained. Experiments on several real-world datasets demonstrate that taking seed nodes as the training set to classify multi-labeled data can improve the classification performance.

Key words: multi-label classification; network; seed nodes

1 引言

目前,多标签分类问题已经取得了广泛关注,并且在实际问题中有很多应用,比如:基因分类,药物发现和文本分类[1].已存在的多标签分类算法,通常都是随机的选取节点作为训练集.然而,在分类算法执行的过程中,网络中不同节点所起的作用不同.在给定训练集数目的情况下,选择的训练集不同,分类精度也会不同.所以随机方法不能有效的利用网络的拓扑结构,导致节点的标签分类结果不稳定.

本文引入了种子节点的概念,分类从种子节点开始,通过不断推理,得到网络中其他节点的标签.应该如何选择种子节点,从而在给定的多标签分类算法下获得较高的分类精度,是本文所要解决的问题,例如:一个大学的所有学生组成了一个网络,学生的标签代表他们的兴趣爱好,如果用一部分学生的标签来预测其他学生的标签,

收稿日期:2015-01-30;修回日期:2015-05-18;责任编辑:覃怀银
基金项目:国家重点基础研究发展规划(973计划)项目(No. 2013CB329604);教育部创新团队(No. IRT13059);国家自然科学基金项目(No. 61229301, No. 61503114)

那么我们要解决的问题是,在网络中应该选择什么样的学生作为种子节点,从而使预测结果最好呢?

我们提出了 SHDA 算法来选择种子节点. 首先使用 EdgeCluster 算法来获取网络中的节点之间潜在的社团[2],接着根据每个社团中包含的节点数目大小,按比例的选取每个社团中度数较高的一些节点,然后将其合并,处理后得到种子节点. 本文将种子节点用作训练集进行多标签分类来验证 SHDA 算法的有效性.

2　相关工作

网络环境下的多标签分类数据一般具有同质性,即服从一阶马尔科夫假设,节点的标签倾向于与其直接邻居的标签相同. 关系学习利用对象之间的联系来进行多标签分类,可以取得较好的分类结果[3]. 集体推理可以进一步改善分类的性能,减少分类错误[3~5]. wvRN(weighted-vote Relational Neighbor classifier)算法计算节点 v_i 属于类 c 的概率,$P(y_i=c|v_i)$,是其邻居中属于类 c 的概率的带权平均值:

$$P(y_i=c|v_i)=\frac{1}{Z}\sum_{v_j\in N_i} w_{ij}\times P(y_j=c|N_j) \qquad (1)$$

其中 $Z=\sum_{v_j\in N_i} w_{ij}$,$w_{ij}$是节点 v_i 和节点 v_j 之间的权值[3].

还有一种方法是利用标签间的相关性,郑等提出了一种基于随机游走模型的多标签分类算法 MLRW(Multi-Label Random Walk algorithm)[7].

然而传统方法无法区分异质网络中对象和边的类型的差异. Ji 等[8] 提出了基于排序的迭代分类模型(Rankclass),提高了异质网络中的分类性能. Kong 等[9] 通过挖掘异质网络中对象间和标签间的关系进行多标签分类. Tang 等[2] 提出了 EdgeCluster 算法,其以连接边的节点作为特征,使用 Scalable K-means Variant 方法将网络的边聚类成多个互不相交的集合,根据聚类结果,构建网络的社会维度,将其作为特征,使用 SVM(Support Vector Machines)算法进行多标签分类. SCRN(multi-label iterative Relation Neighbor Classifier that employs Social features)算法[10]拓展了关系邻居分类器,其计算节点 v_i 属于类 c 的概率为:

$$P(l_i^c|N_i,F(v_i))=$$
$$\frac{1}{Z}\sum_{v_j\in N_i} P_{\mathrm{CP}}(l_i^c|F(v_i))\times w(v_i,v_j)\times P(l_j^c|N_j) \qquad (2)$$

其中 Z 是归一化因子,$P_{\mathrm{CP}}(l_i^c|F(v_i))$是节点 v_i 的对类 c 的传播概率. Zhou 等[11] 提出了以活动边为中心的多标签分类框架来挖掘异质网络,使用迭代学习算法动态的改善分类性能.

在种子节点选择方面,Qian 等通过在隐式社交网络中选择种子节点解决了网络的影响力最大化问题[12]. Jankowski 等通过选择种子节点进行社交网络中的信息传播[13].

3　SHDA 算法

我们首先介绍一下种子节点的定义.

定义 1(种子节点)　种子节点是网络环境下多标签分类的起点,通过整个网络对标签的传播和扩散,可预测出网络中其他节点的标签.

下面介绍本文提出的 SHDA 算法的步骤:

第 1 步:计算网络中各节点的度数.

第 2 步:使用 EdgeCluster 算法[2] 提取网络的社会维度. 社会维度是节点对各个社团从属程度的描述,一个节点可以从属于多个社团[14]. 我们去除每个社团中属于测试集的节点,计算 $\eta=\{\eta_1,\eta_2,\cdots,\eta_N\}$,即每个社团中包含的节点数目,以节点度数由高到底的顺序对每个社团中的节点进行排序.

第 3 步:计算需要从每个社团中选取的节点数目 m_i:

$$m_i=\lceil n_{tr}\times\eta_i/\sum_{i=1}^{N}\eta_i\rceil \qquad (3)$$

其中 n_{tr}代表训练集的大小,向上取整函数保证了 m_i 是一个整数. 从每个社团中选取前 m_i 个节点,这些节点即是度数排在前 m_i 个的节点.

第 4 步:合并上述从每个社团中所选的节点作为种子节点集合,由于一个节点可以从属于多个社团,导致节点的重复选取,同时向上取整函数也会导致节点选择的增多,使得种子节点集合的数目与训练集大小不相等,所以要去除重复节点,然后调整种子节点集合大小.

去除重复节点后的种子节点集合大小记为 n_{seed},若 $n_{\mathrm{seed}}>n_{\mathrm{tr}}$,我们就从种子节点集合中随机选取 n_{tr}个节点作为种子节点,若 $n_{\mathrm{seed}}<n_{\mathrm{tr}}$,则再从除去测试集和种子节点集合的剩余节点集合 G'中,随机选取 $n_{\mathrm{tr}}-n_{\mathrm{seed}}$个节点,与种子节点集合合并作为种子节点.

算法流程总结起来就是:无向图 G 被划分成多个社团,在每个社团中去除属于测试集的节点,利用式(3)选取前 m_i 个度数较大的节点,然后合并为种子节点集合,最后去除重复节点,调整种子节点集合大小后,得到种子节点.

种子节点的标签是已知的,并且标签分类从这些节点开始,经过不断推理,得到网络中其他节点的标签. 选择的种子节点度数都比较高,与其他节点的联系比较紧密,同时种子节点在每个社团上分布的比较均匀,有利于标签传播过程中,更准确的把标签扩散至整个网络,减少分类错误,改善多标签分类的性能.

4　实验设计

我们选择了 3 个真实数据集,DBLP、BlogCatalog 和

YouTube 进行实验[2,10]. 关于数据集的详细信息见表 1.

我们使用 SHDA 算法选择的种子节点作为训练集,并分别使用 SCRN、wvRN 和 Edge(EdgeCluster)算法进行网络环境下的多标签分类实验,记为 SHDA + SCRN、SHDA + wvRN 和 SHDA + Edge. 在对比算法中,我们随机的从网络中选择节点作为训练集,分类算法也采用 SCRN、wvRN 和 Edge 算法,记为 randomly + SCRN、randomly + wvRN 和 randomly + Edge. 通过比较 SHDA + SCRN/wvRN/Edge 和 randomly + SCRN/wvRN/Edge 方法的分类评估结果,来验证 SHDA 算法的有效性.

我们将 DBLP、BlogCatalog 和 YouTube 数据集的社会特征的维数分别设置为 1000、5000 和 1000[2,10],训练集选择的比例分别设置为 5% 到 30%,10% 到 60% 和 1% 到 10%. 我们采用网络交叉验证 NCV(Network Cross-Validation)方法[15]来减少测试样本的重叠选取. 评估指标选用宏观 F1 值,微观 F1 值和汉明损失(Hamming Loss)[10]. Micro-F1 值和 Macro-F1 值越大,分类性能越好,Hamming Loss 值越小,分类性能越好. 实验结果取 10 次实验的平均值.

表 1 数据集描述

Data	YouTube	BlogCatalog	DBLP
Nodes	100,000	10,312	8,865
Links	648,092	333,983	12,989
Categories	47	39	15
Network Density	3.2×10^{-3}	6.3×10^{-3}	3.3×10^{-4}
Maximum Degree	16,007	3,992	86
Average Degree	64	65	3

5 实验结果与分析

5.1 实验结果

实验结果记录在表 2 ~ 4 中,代表 randomly + SCRN/wvRN/Edge 方法和 SHDA + SCRN/wvRN/Edge 方法在各个数据集上的多标签分类结果.

YouTube 数据集上的实验结果记录在表 2 中,分析数据可知:在 Micro-F1 指标下,SHDA + SCRN、SHDA + Edge 和 SHDA + wvRN 方法分别比 randomly + SCRN、randomly + Edge 和 randomly + wvRN 方法提高了 15.63%、3.16% 和 -6.45%. 在 Macro-F1 指标下,提高了 46.78%、2.53% 和 4.42%. 在 Hamming Loss 指标下,降低了 4.26%、1.15% 和 -4.47%.

BlogCatalog 数据集上的实验结果记录在表 3 中,分析结果可得:SHDA 算法应用于 SCRN、EdgeCluster 和 wvRN 算法分别比随机方法应用于这些算法,提高了 9.08%、-0.68% 和 4.34% 的 Micro-F1 值,提高了 14.82%、-0.64% 和 5.32% 的 Macro-F1 值,降低了 2.81%、-0.30% 和 1.54% 的 Hamming Loss 值.

DBLP 数据集上的实验结果记录在表 4 中,由结果可得:SHDA + SCRN、SHDA + Edge 和 SHDA + wvRN 方法分别比 randomly + SCRN、randomly + Edge 和 randomly + wvRN 方法,提高了 3.18%、5.89% 和 1.15% 的 Micro-F1 值,提高了 3.47%、7.92% 和 0.94%. 的 Macro-F1 值,降低了 3.69%、5.33% 和 1.18% 的 Hamming Loss 值.

综合所述,大部分情况下,SHDA 算法在多标签分类方法上的性能要比随机方法好. 某些情况下,SHDA

表 2 YouTube 数据集上实验结果

训练集比例		1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Micro-F1 (%)	randomly + SCRN	30.42	31.29	32.87	34.43	35.20	35.87	36.75	37.59	38.13	38.37
	SHDA + SCRN	35.19	35.99	36.03	36.61	36.88	37.24	37.49	37.68	37.99	38.04
	randomly + Edge	23.14	25.99	26.78	28.45	31.51	32.00	33.60	33.09	34.91	34.97
	SHDA + Edge	25.04	27.57	28.04	29.43	30.58	32.89	33.40	34.67	35.44	36.23
	randomly + wvRN	34.83	35.04	36.35	38.93	38.47	38.33	40.04	40.33	40.27	40.51
	SHDA + wvRN	34.42	34.95	34.91	35.43	35.66	35.99	36.25	36.50	36.85	36.78
Micro-F1 (%)	randomly + SCRN	12.81	13.81	15.32	16.03	17.62	18.59	19.51	20.06	20.95	21.68
	SHDA + SCRN	20.94	22.91	24.14	25.02	25.20	25.99	26.74	27.27	28.70	28.60
	randomly + Edge	19.05	21.52	22.05	22.33	24.44	24.60	25.02	24.56	25.43	26.15
	SHDA + Edge	19.10	21.75	22.43	22.69	23.77	25.38	25.42	26.35	26.97	27.52
	randomly + wvRN	20.84	22.00	24.24	25.06	25.31	26.25	27.52	27.53	27.84	28.43
	SHDA + wvRN	22.21	24.12	25.58	26.03	26.26	27.00	27.80	28.19	29.31	29.35
Hamming Loss (%)	randomly + SCRN	0.59	0.58	0.57	0.56	0.55	0.54	0.54	0.53	0.52	0.52
	SHDA + SCRN	0.54	0.54	0.53	0.53	0.53	0.52	0.52	0.52	0.52	0.51
	randomly + Edge	0.65	0.63	0.62	0.61	0.58	0.58	0.56	0.57	0.55	0.55
	SHDA + Edge	0.64	0.61	0.61	0.60	0.59	0.57	0.57	0.55	0.55	0.54
	randomly + wvRN	0.55	0.55	0.54	0.52	0.52	0.52	0.51	0.51	0.51	0.50
	SHDA + wvRN	0.56	0.55	0.55	0.55	0.55	0.54	0.54	0.54	0.54	0.54

算法表现的不如随机方法,可能是因为 SHDA 算法在最后一步时,从集合 G' 中随机选取 $n_{tr}-n_{seed}$ 个节点并入训练集,导致 SHDA 算法在多标签分类实验中表现的稍微不稳定. 但是从整体上来说,SHDA 算法由于利用了网络的拓扑结构,有助于提高网络环境下多标签分类的性能.

表 3　BlogCatalog 数据集上实验结果

训练集比例		10%	20%	30%	40%	50%	60%
Micro-F1 (%)	randomly + SCRN	20.45	24.77	27.32	29.21	31.11	31.86
	SHDA + SCRN	26.04	27.90	29.43	30.40	31.34	32.49
	randomly + Edge	28.69	30.65	32.52	33.63	34.02	34.69
	SHDA + Edge	27.62	31.32	32.21	33.43	33.89	34.49
	randomly + wvRN	23.49	25.85	27.38	29.03	30.91	31.47
	SHDA + wvRN	25.72	27.55	28.89	29.91	30.96	31.86
Micro-F1 (%)	randomly + SCRN	7.40	10.97	13.34	14.86	16.08	17.27
	SHDA + SCRN	11.26	13.13	14.36	15.53	16.55	17.61
	randomly + Edge	16.37	18.84	21.04	21.63	22.32	23.11
	SHDA + Edge	15.73	19.04	20.40	21.95	22.51	23.04
	randomly + wvRN	10.27	12.21	13.40	14.48	15.72	16.76
	SHDA + wvRN	11.11	12.96	14.02	15.21	16.48	17.28
Hamming Loss (%)	randomly + SCRN	5.73	5.41	5.23	5.09	4.96	4.90
	SHDA + SCRN	5.32	5.19	5.08	5.01	4.94	4.86
	randomly + Edge	5.13	4.99	4.86	4.78	4.75	4.70
	SHDA + Edge	5.21	4.94	4.88	4.79	4.76	4.72
	randomly + wvRN	5.51	5.34	5.23	5.11	4.97	4.93
	SHDA + wvRN	5.35	5.21	5.12	5.05	4.97	4.90

表 4　DBLP 数据集上实验结果

训练集比例		5%	10%	15%	20%	25%	30%
Micro-F1 (%)	randomly + SCRN	51.54	56.77	60.34	62.43	65.43	66.35
	SHDA + SCRN	53.58	59.46	62.30	64.56	66.41	67.82
	randomly + Edge	41.64	46.93	50.71	52.97	55.32	56.12
	SHDA + Edge	46.79	51.76	53.03	54.94	56.45	57.44
	randomly + wvRN	47.98	53.09	57.06	59.19	62.05	63.09
	SHDA + wvRN	49.20	54.67	57.53	59.84	61.70	63.12
Micro-F1 (%)	randomly + SCRN	44.55	49.25	53.38	55.73	58.51	59.41
	SHDA + SCRN	46.77	52.42	54.99	57.29	59.40	60.62
	randomly + Edge	34.51	39.74	43.80	46.26	47.89	49.42
	SHDA + Edge	39.88	45.09	46.03	48.82	50.04	51.09
	randomly + wvRN	41.96	46.38	51.05	53.37	55.93	56.53
	SHDA + wvRN	43.10	48.24	50.88	53.28	55.44	56.69
Hamming Loss (%)	randomly + SCRN	12.90	11.76	10.98	10.51	9.80	9.61
	SHDA + SCRN	12.52	11.21	10.59	10.04	9.56	9.22
	randomly + Edge	17.68	16.08	14.93	14.25	13.53	13.29
	SHDA + Edge	16.13	14.62	14.24	13.66	13.20	12.90
	randomly + wvRN	15.76	14.20	13.01	12.36	11.49	11.18
	SHDA + wvRN	15.40	13.74	12.87	12.17	11.61	11.18

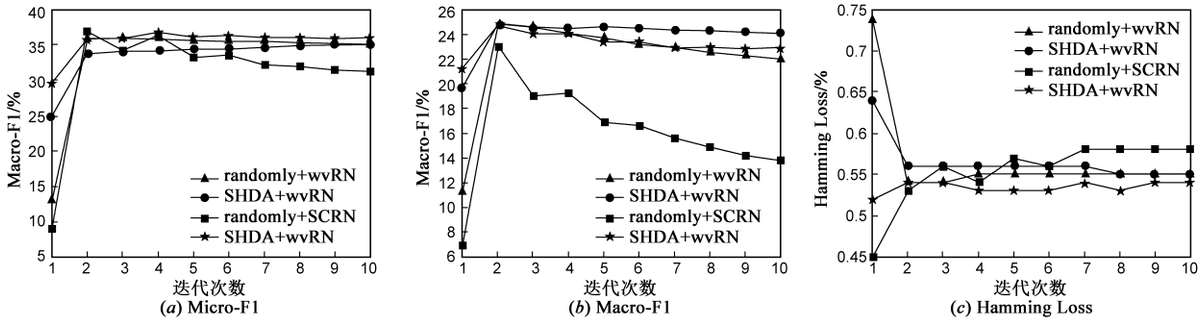
5.2　算法迭代次数分析

SCRN 和 wvRN 算法使用标签松弛法来进行集体推理. 在每一次迭代过程中,算法根据上一次的预测结果,逐次更新节点属于各个类的概率,更新类的参考特征,更新类的传播概率,根据更新后的结果预测标签,直至预测出的节点标签趋于稳定或者迭代次数到达最大值,算法终止.

我们在 YouTube、BlogCatalog 和 DBLP 三个数据集上分别做了实验来讨论 SHDA + SCRN/wvRN 和 randomly + SCRN/wvRN 方法的实验结果随着迭代次数的变化情况. EdgeCluster 算法利用社会特征使用 SVM 算法进行分类,不在本部分讨论的范围之内. 图 1 ~ 3 记录了 YouTube、BlogCatalog 和 DBLP 数据集上,训练集分别选择 2% 、5% 和 20% 时,实验结果随着迭代次数的变化情况.

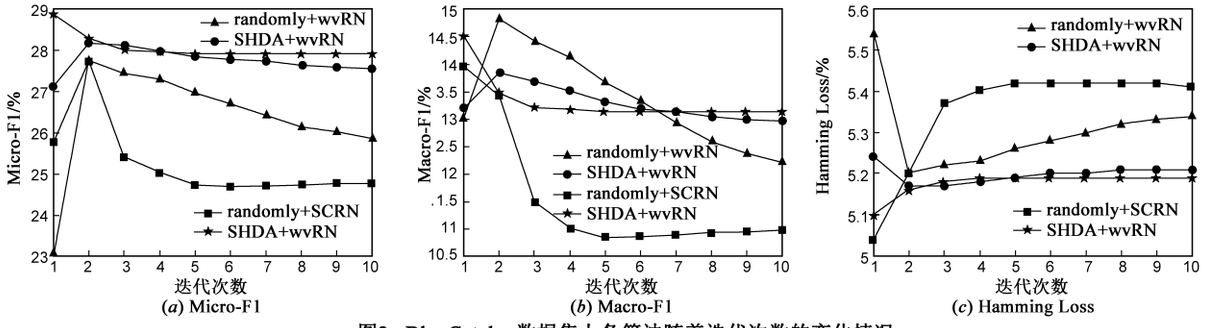
分析结果发现,SHDA + SCRN/wvRN 方法的实验

结果在迭代次数大于 5 的时候基本趋于稳定. 而 randomly + SCRN/wvRN 方法可能导致实验结果在迭代次数等于 2 的时候达到达最大值,然后随着迭代次数的增加而下降. 所以 SHDA 算法应用于多标签分类时有助于保持算法的稳定性.



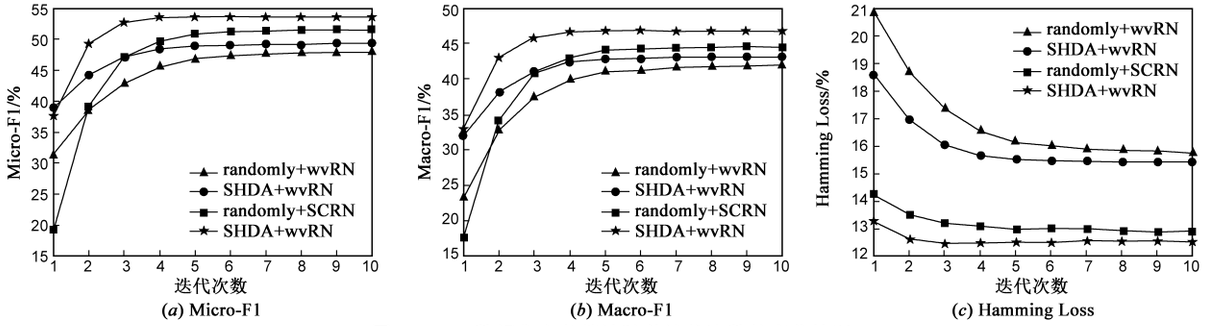
(*a*) Micro-F1 (*b*) Macro-F1 (*c*) Hamming Loss

图1 YouTube数据集上各算法随着迭代次数的变化情况



(*a*) Micro-F1 (*b*) Macro-F1 (*c*) Hamming Loss

图2 BlogCatalog数据集上各算法随着迭代次数的变化情况



(*a*) Micro-F1 (*b*) Macro-F1 (*c*) Hamming Loss

图3 DBLP数据集上各算法随着迭代次数的变化情况

5.3 算法运行效率分析

分析 SHDA 算法可知,该算法的执行时间与社会特征的维数,网络规模以及种子节点数目有关. 其中社会特征的维数对 SHDA 算法的影响最大. 而随机方法只与网络的节点数和训练集大小相关. 我们在 Intel(R) 双核 2.60GHz,32GB 内存的 PC 机上分别计算了 SHDA 算法和随机方法应用于多标签分类时,各算法的运行时间,结果记录在表 5 ~ 8 中,单位时间为 s.

分析结果可得:SHDA 算法在 BlogCatalog 数据集上耗时最长,其次是 YouTube 数据集,再次是 DBLP 数据集. SHDA 算法由于利用了网络的拓扑结构,执行的时间没有随机方法快,但是时间是在可以接受的范围内(10s 以内).

与随机方法相比,SHDA 算法应用于 SCRN 和 wvRN 算法时,降低了 SCRN 和 wvRN 算法的执行时间. 因为种子节点与网络中其他节点的关联性较强,能够加快标签传播和扩散的速度. EdgeCluster 算法在使用种子节点后,算法的执行时间增大,因为 EdgeCluster 算法利用节点的社会特征使用 SVM 分类器进行分类,种子节点带有的特征比一般节点多,使得算法需要花费更多的时间来分类.

SHDA + EdgeCluster 方法比 randomly + EdgeCluster

方法执行的时间长. SHDA + SCRN/wvRN 方法在 BlogCatalog 和 You Tube 数据集上比 randomly + SCRN/wvRN 方法执行的时间短,在 DBLP 数据集上两者执行的时间大致相当. 所以 SHDA 算法能够提升网络环境下多标签分类的性能,还能提升部分多标签分类算法的运行速度.

表 5　YouTube 数据集上各算法的运行时间

训练集比例	1%	2%	3%	4%	5%
randomly + SCRN	0.0716 +20220.209	0.0721 +21110.6866	0.0675 +20463.7392	0.0736 +16309.23	0.0756 +14418.0039
SHDA + SCRN	1.8924 +19120.8712	1.4899 +18846.2674	1.4006 +18287.7288	1.5488 +15999.571	1.7594 +12873.8918
randomly + Edge	0.0716 +1.3356	0.0721 +4.3049	0.0675 +7.3865	0.0736 +11.2170	0.0756 +16.4022
SHDA + Edge	1.8924 +4.9108	1.4899 +11.7078	1.4006 +19.1307	1.5488 +23.1049	1.7594 +26.9862
randomly + wvRN	0.0716 +11914.2346	0.0721 +12245.4936	0.0675 +11957.5732	0.0736 +11753.4907	0.0756 +11087.6778
SHDA + wvRN	1.8924 +11854.2659	1.4899 +12104.3575	1.4006 +11787.3501	1.5488 +11427.9402	1.7594 +11042.2835

表 6　YouTube 数据集上各算法的运行时间

训练集比例	6%	7%	8%	9%	10%
randomly + SCRN	0.0773 +14093.2313	0.0782 +14122.5027	0.0782 +13870.4589	0.0789 +13554.5594	0.0720 +12046.1614
SHDA + SCRN	1.3839 +12321.9526	1.5104 +12279.0958	1.4096 +12084.7095	1.5564 +11767.0459	1.5570 +11644.8778
randomly + Edge	0.0773 +19.7820	0.0782 +22.1468	0.0782 +27.4766	0.0789 +30.7452	0.0720 +32.6464
SHDA + Edge	1.3839 +28.8799	1.5104 +34.5318	1.4096 +36.4880	1.5564 +38.3377	1.5570 +39.0773
randomly + wvRN	0.0773 +10970.7951	0.0782 +11021.306	0.0782 +10960.4526	0.0789 +10562.2632	0.0720 +10507.8926
SHDA + wvRN	1.3839 +10780.9173	1.5104 +1069.1761	1.4096 +10403.0394	1.5564 +10340.6883	1.5570 +10334.9079

表 7　BlogCatalog 数据集上各算法的运行时间

训练集比例	10%	20%	30%	40%	50%	60%
randomly + SCRN	0.0100 +657.3252	0.0105 +576.988	0.0106 +489.379	0.0109 +419.7945	0.0112 +337.8141	0.0113 +266.8242
SHDA + SCRN	7.2481 +635.6806	7.1653 +534.9737	7.1919 +449.883	7.1589 +416.4779	7.1618 +329.2965	7.1833 +261.3027
randomly + Edge	0.0100 +3.0268	0.0105 +6.3613	0.0106 +10.4745	0.0109 +14.9554	0.0112 +19.4940	0.0113 +24.7098
SHDA + Edge	7.2481 +3.0087	7.1653 +9.2308	7.1919 +14.9193	7.1589 +18.9712	7.1618 +25.6156	7.1833 +29.3793
randomly + wvRN	0.0100 +270.0023	0.0105 +238.4652	0.0106 +207.3798	0.0109 +178.3946	0.0112 +144.8433	0.0113 +113.9107
SHDA + wvRN	7.2481 +212.0875	7.1653 +182.961	7.1919 +159.7029	7.1589 +133.749	7.1618 +111.648	7.1833 +89.4766

表 8　DBLP 数据集上各算法的运行时间

训练集比例	5%	10%	15%	20%	25%	30%
randomly + SCRN	0.0083 +59.0717	0.0085 +54.2758	0.0087 +51.4626	0.0088 +47.2634	0.0088 +42.6631	0.0089 +42.8946
SHDA + SCRN	1.2938 +58.4726	1.3624 +53.6815	1.4416 +49.6739	1.3356 +45.3047	1.3202 +40.8669	1.3653 +41.4465
randomly + Edge	0.0083 +0.0541	0.0085 +0.0992	0.0087 +0.1220	0.0088 +0.1494	0.0088 +0.1725	0.0089 +0.1864
SHDA + Edge	1.2938 +0.0584	1.3624 +0.0983	1.4416 +0.1376	1.3356 +0.1653	1.3202 +0.1867	1.3653 +0.2135
randomly + wvRN	0.0083 +26.8711	0.0085 +24.4639	0.0087 +22.8725	0.0088 +21.4823	0.0088 +20.0958	0.0089 +18.4917
SHDA + wvRN	1.2938 +26.5227	1.3624 +24.1004	1.4416 +22.8897	1.3356 +21.3200	1.3202 +19.5536	1.3653 +18.3373

6　总结

在网络环境下的多标签分类中,给定训练集的节点数目,如果选择的训练集不同,分类精度也会不同. 我们提出了 SHDA 算法,利用网络的拓扑结构来选择种子节点,首先根据网络中每个社团的大小,按比例的选取每个社团中度数较高的节点,然后将这些节点合并,处理后得到种子节点,将种子节点用作训练集进行多标签分类. 实验证明,相比于从网络中随机选取节点作为训练集,这种方法不仅有助于提高多标签分类的性能,还能改善部分多标签分类算法的运行速率.

SHDA 算法是从种子节点选择的方面来提高多标签分类的性能,我们今后的研究方向将从改善多标签分类算法这方面,研究提高多标签分类性能的更有效的方法.

参考文献

[1] Yang S, Jiang Y. Zhou Z. Multi-label learning with weak label[A]. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence[C]. Atlanta: AAAI, 2010. 593 – 598.

[2] Tang L, Liu H. Scalable learning of collective behavior based on sparse social dimensions[A]. Proceedings of the 18th ACM Conference on Information and Knowledge Management[C]. Hong Kong: ACM, 2009. 1107 – 1116.

[3] Macskassy S A, Provost F. A simple relational classifier

[A]. Proceedings of the Multi-relational Data Mining Workshop at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Washington:ACM,2003. 64 – 76.

[4] Lu Q,Getoor L. Link-based classification[A]. Proceedings of the Twentieth International Conference on Machine Learning[C]. Washington:JMLR,2003. 496 – 503.

[5] Neville J,Jensen D. Relational dependency networks[J]. Journal of Machine Learning Research,2007,8:653 – 692.

[6] Jensen D,Neville J,Gallagher B. Why collective inference improves relational classification[A]. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Seattle:ACM,2004. 593 – 598.

[7] 郑伟,王朝坤,刘璋,王建民. 一种基于随机游走模型的多标签分类算法[J]. 计算机学报,2010,33(8):1418 – 1426.
Zheng Wei,Wang Chao-kun,Liu Zhang,Wang Jian-min. A multi-label classification algorithm based on random walk model[J]. Chinese Journal of Computers,2010,33(8):1418 – 1426. (in Chinese)

[8] Ji M,Han J,Danilevsky M. Ranking-based classification of heterogeneous information networks[A]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Diego:ACM,2011. 1298 – 1306.

[9] Kong X,Cao B,Yu P S. Multi-label classification by mining label and instances correlations from heterogeneous information networks[A]. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Chicago:ACM,2013. 614 – 622.

[10] Wang X, Sukthankar G. Multi-label relational neighbor classification using social context features[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Chicago:ACM,2013. 464 – 472.

[11] Zhou Y,Liu L. Activity-edge centric multi-label classification for mining heterogeneous information networks[A]. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York:ACM,2014. 1276 – 1285.

[12] Qian T,Liu J. Influence maximization through identifying seed nodes from implicit social networks[A]. Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication[C]. Suwon:ACM,2010. 193 – 196.

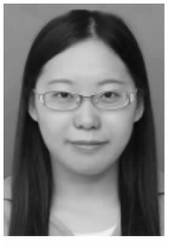
[13] Jankowski J,Michalski R,Kazienko P. Compensatory seeding in networks with varying availability of nodes[A]. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining[C]. Niagara:ACM,2013. 42 – 1249.

[14] Tang L,Liu H. Relational learning via latent social dimensions[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Paris:ACM,2009. 817 – 825.

[15] Neville J,Gallagher B,et al. Correcting evaluation bias of relational classifiers with network cross validation[J]. Knowledge and Information Systems,2012,30(1):32 – 52.

作者简介

吴信东 男,1963 年出生于安徽枞阳,教授,博士生导师,IEEE Fellow,AAAS Fellow,主要研究方向为数据挖掘,大数据分析,基于知识的系统和万维网信息探索.
E-mail:xwu@hfut.edu.cn



赵银凤 女,1990 年出生于安徽淮北,硕士研究生,主要研究方向为数据挖掘,社交网络.
E-mail:zhyinfeng@163.com

李 磊 男,1981 年出生于辽宁鞍山,副教授,主要研究社会计算,数据挖掘和信任计算.
E-mail:lilei@hfut.edu.cn