基于 NSGA2 的网络环境下多标签种子节点选择

李 磊*① 楚喻棋① 汪 萌① 韩 莉② 吴信东①③
①(合肥工业大学计算机与信息学院 合肥 230009)
②(科学技术部基础研究管理中心 北京 100862)
③(路易斯安那州立大学计算机与信息学院 拉斐特 70503 美国)

摘 要：随着社交网络规模的不断扩大，网络节点的标签分类也不再单一，变得丰富多样，这些促使了社交网络中的多标签分类问题成为一个重要的研究领域。以前的研究重点主要集中在提高预测网络节点标签的精度上，而忽略了得到节点信息所产生的包含时间消耗和计算资源等在内的系统开销问题。可现如今随着网络规模不断扩大且复杂性不断增强，之前所忽略的系统开销问题变得越来越严重，增加了预测标签的成本，加重了预测网络节点标签的难度。该文针对这一问题提出了基于 NSGA2 算法的网络环境下多标签种子节点选择算法(NAMESEA 算法)，目的是在能大大降低预测节点标签所消耗的系统开销的前提下一定程度上提高预测标签的精度。该文将 NAMESEA 算法与其他多标签预测算法在多个真实数据集上进行实验对比，结果证明 NAMESEA 算法大大降低了预测节点标签的系统开销并且提高了预测精度。

关键词：社交网络；多标签分类；NSGA2；系统开销
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2017)09-2040-08
DOI: 10.11999/JEIT161266

NSGA2-based Multi-label Seed Node Selection in Network Environments

LI Lei① CHU Yuqi① WANG Meng① HAN Li② WU Xindong①③
①(*School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei* 230009, *China*)
②(*Basic Research Management Center*, *Ministry of Science and Technology*, *Beijing* 100862, *China*)
③(*School of Computing and Informatics*, *University of Louisiana at Lafayette*, *Lafayette* 70503, *USA*)

Abstract: With the expanding scale of social networks, the label classification of nodes in the network is no longer single but various, which prompts the multi-label classification in social networks to become an important research area. The previous research focuses on how to improve the precision of the predicted labels, while ignoring the system overhead caused by obtaining the node information, such as time consumption and computing memory occupancy. Now, as both expansion and complexity of the networks are increasing, the problem of previously neglected system overhead is becoming the more and the more serious. It increases not only the cost but also the difficulty of predicting labels. In this paper, an NSGA2-based multi-label seed selection algorithm in network environments (NAMESEA) is proposed to improve the accuracy of label prediction on the condition that reducing both the time consume and the memory occupancy. Compared with other multi-label prediction algorithms on multiple real datasets, NAMESEA algorithm not only greatly reduces the system overhead but also improves the prediction accuracy.

Key words: Social networks; Multi-label classification; NSGA2; System overhead

1 引言

近年来随着社交网络应用的发展普及，社交网络吸引了越来越多学者的研究目光[1–5]，其中一个重要的研究方向就是社交网络中的多标签预测问题[1]。利用标签预测我们可以通过网络中已知用户标签预测得到未知用户标签，从而对用户进行分类和划分社区，进而有针对性地进行信息推荐。虽然已经有一些算法[1,6–8]来解决网络中的多标签预测问题，但随着社交网络规模的不断扩大，以及数据结构的复杂性不断增强，为了得到节点信息所引起的系统开销特别是花费的时间和消耗的系统内存不断

收稿日期：2016-11-24；改回日期：2017-04-11；网络出版：2017-05-11
*通信作者：李磊 lilei@hfut.edu.cn
基金项目：国家 973 规划项目(2013CB329604)，国家重点研发计划项目(2016YFB1000901)，国家自然科学基金项目(61503114)
Foundation Items: The National 973 Program of China (2013CB329604), The National Key Research and Development Program of China (2016YFB1000901), The National Natural Science Foundation of China (61503114)

增长。这无形中增加了预测网络中未知节点标签的成本。同时，这些额外增加的成本无疑给社交网络中多标签预测又增添了难度。为了节约系统开销，更好地对社交网络进行多标签分类，本文提出了基于 NSGA2 算法的多标签预测算法。

本文的创新点在于引入了遗传算法中的多目标优化算法 NSGA2 来优化选择社交网络中的节点，赋予种子节点包含提高预测精度、减少时间开销、降低内存占有等在内的功能特性。具体来说，本文根据设定目标模型综合考察每个节点在消耗时间，消耗内存和网络传播重要性等方面的作用，经过遗传算法中的交叉、变异和种群迭代等过程决策选择出符合目标模型要求的节点，将其作为种子节点，生成预测网络中未知节点标签的种子集。在社交网络多标签分类过程中，该种子集中的节点在能保证降低所花费的系统开销的前提下，较为精准地得到未知的节点多标签分类结果。

2　相关工作

2.1 多标签分类方法

传统解决多标签分类问题通常是将多标签问题分解成多个单标签分类问题[9−13]，而单标签分类问题利用成熟的关系模型可以很好地解决。Macskassy 等人[8]发现 RN(Relational Neighbor)分类模型预测标签虽然仅仅是依靠关联邻居节点的标签，但是其效果却好过关系概率模型、关系概率树模型等其他复杂模型。在此基础上 Wang 等人[1]扩展了 RN 分类器，利用网络拓扑结构，从中抽取每个节点相对应的社会维度，借此来判定节点归属标签的概率，从而获取更侧重节点内在属性的标签。

除了上述 RN 及其相关算法，现在还有很多其他的多标签分类方法。例如，申超波等人[14]利用改进的平衡 k-means 方法将训练集改进为由潜在的重要标签组成，因此得到了更好的分类结果。郑伟等人[15]引入随机游走模型，将未分类数据在多标签转化的随机游走模型中得到的顶点概率作为标签概率进行分类，较好地解决了标签排序问题。张振海等人[16]通过计算特征与标签之间的信息增益来判定标签的倾向性，提出了一种基于信息熵的多标签特征选择算法。

2.2 NSGA2遗传算法

NSGA2 算法是 Kalyanmoy 等人[17]对 NSGA 算法的改进，主要思路是在找到符合设定目标最优解的前提下大大降低时间复杂度和最大程度地保留下种群中的精英解。

NSGA2 针对 NSGA 的时间复杂度过大和缺少精英策略以导致种群中最优个体流失这两个问题在算法内部进行了大量的改动，因此两者本质并不相同[17]。NSGA2 算法主要改进了 3 点：一是在保证快速找到 pareto 前沿最优解的同时改掉了 NSGA 的非劣分层思想，引入快速非支配排序方法，统计个体在可行解空间中所支配个体的数量，然后按照支配个体数量进行分层排序，层级越低越好，这一新策略很好地解决了 NSGA 算法非劣分层重复次数过多，时间复杂度较大的问题；二是加入精英策略，将种群的父代和子代合并到一起进行分层排序，较好地保留住种群中的精英个体；三是加入拥挤距离这一思想，计算第 i 个体与它相邻的第 $i+1$, $i-1$ 个体在目标函数值上差的和，该值越大说明它们之间的差异越大，如式(1)，然后选择两个相距远的个体能较好地保证种群的多样性[13]。拥挤距离的计算公式为

$$d_{I_j^m}=d_{I_j^m}+\frac{f_m^{(I_{j+1}^m)}-f_m^{(I_{j-1}^m)}}{f_m^{\max}-f_m^{\min}} \tag{1}$$

NSGA2 算法的核心是多目标优化问题，目前被广泛地运用在各行各业以寻找最优方案。例如，刘晓娟等人[18]运用 NSGA2 算法解决了多目标并行调度问题。孙建龙等人[19]选择出投资成本少、电压稳定和功率损耗少的电源配置。张利[20]找到能使控制器进行低频阻尼振荡的参数。正是因为 NSGA2 算法在寻找多目标最优解上的出色表现，本文在筛选符合社交网络目标要求的种子节点上运用 NSGA2 算法。

3　基于 NSGA2 的多标签种子节点选择算法

3.1 NSGA2算法建模

在运用 NSGA2 算法选择出所需的种子节点集之前最重要的一步就是根据社交网络的需求设定目标，构建种子节点决策模型。设推算社交网络 W 中未知节点的标签，$W=\{1,2,\cdots,w,\cdots,|w|\}$ 为网络中所有节点集合，选定种子节点集合为 W 的子集，设为 S。

在选种子集 S 中节点时，本文设定其选择决策受消耗时间、占用系统内存、社会维度这 3 方面因素的影响。其中为了减小上述预测社交网络中未知节点标签所产生的系统开销且保证预测精度，我们设定模型目标为以下 3 个，其中 W_i 代表网络中的节点，t_i 代表该节点消耗的时间，c_i 代表该节点所占用的内存，SF_i 代表该节点所对应的社会维度特征值。

目标 1：时间最小，Min $f_1=\sum_{w=1}^{|W|}w_it_i$；目标 2：占用内存最小，Min $f_2=\sum_{w=1}^{|W|}w_ic_i$；目标 3：

社会维度特征值最大，Max $f_3=\sum_{w=1}^{|W|}w_i\mathrm{SF}_i$。其中，$w_i=\begin{cases}1, & w_i\in S\\ 0, & \text{其它}\end{cases}$。

3.2 NAMESEA算法步骤

本节将介绍基于 NSGA2 的多标签种子节点选择算法(NAMESEA 算法)具体步骤，伪代码详见表 1。

步骤 1 设定判断种子节点的目标函数及约束条件(算法 1,1 行)。

步骤 2 确定得到社交网络中节点信息所花费的时间 t_i，所占用的内存 c_i，对网络中 n 个节点的时间和内存值进行归一化处理；同时采用 Scalable K-means edge clustering 方法[13]计算得到该社交网络中第 i 个节点对应的社会维度向量，并对该网络的社会维度特征集合进行统计分析，根据整体情况选定一个阈值 β 值，统计节点 i 中大于 β 值的数量作为社会维度特征值 SF_i (算法 1 的 2, 3 行)。

步骤 3 根据该社交网络中的节点数量确定所需种子节点个数 ns，确定种群规模 J，代数 gn，对种群进行随机初始化，得到种群的初始父代 G_0，$G_0=\{g_1,g_2,\cdots,g_e,\cdots,g_j\}$；其中 g_e 代表种群中的一个个体，即为一条染色体，根据目标函数分别计算第 g_e 条染色体在 M 个目标函数上的值 Mg_e (算法 1 的 4~7 行)。

步骤 4 在初始父代 G_0 的基础上利用二进制锦标赛法排序，然后进行选择、交叉和变异操作，产生子代 Q_0，将 g_0' 和 Q_0 合并进 Rt 中，其中 Rt 的初始状态为 0，对 Rt 进行快速非支配解排序，构造等级 F，$F=\{F1,F2,F3,\cdots\}$；然后根据染色体 g_e 所处等级 $F1,F2,F3,\cdots$ 进行排序(算法 1 的 8~15 行)。

步骤 5 根据每条染色体的等级 F 将所属同一等级的所有个体放入集合 C 中计算染色体 g_e 的拥挤距离 De，对种群中的所有染色体进行排序，De 越大排名越靠前，取排名前 J 的染色体组为新一代种群 G_1 (算法 1 的 16~19 行)。

步骤 6 对新子代 G_1 重复步骤 4 和步骤 5 的操作进行 gn 代得到的种群 G_{gn} 即为最优种群，其中每条染色体为一个解，包含一组最优种子节点的选择，本文得到 J 组不同的选择，从中任选 N 组代入到社交网络中完成标签的传播(算法 1 的 20~23 行)。

完成上述步骤本文就能得到 N 组解集，每一组解集代表了一种种子节点选择策略，将解集中选定的节点提取出来作为训练集代入社交网络中利用 wvRN 和 SCRN 算法进行多标签分类，以验证算法的节约性和准确性。

表 1 NAMESEA 算法流程

算法 1 NAMESEA 算法

输入：社交网络 W，时间消耗，内存消耗，种群 J，代数 N

输出：种子节点

1 设定目标函数，约束条件，种子节点个数
2 用边聚类的方法计算每个节点的社会维度
3 对社会维度，时间消耗，内存消耗进行归一化处理
4 For 代数 1 to N do as follows
5 根据种子节点个数初始化染色体
6 计算每条染色体在目标函数上的值
7 构建 J 条染色体作为第 1 代
8 用二进制锦标赛法对染色体进行排序
9 对染色体进行选择、交叉和变异操作
10 将父代与子代合并到一起用快速非劣解排序法进行排序
11 if p 支配 q
12 把 q 加到 p 支配的解集中，p 支配的解集数加 1, p 属于第 1 级, q 属于下一级
13 Else
14 Q 属于第 1 级，q 属于下一级
15 End if
16 计算每条染色体的拥挤距离
17 if $(i_{\text{rank}}<j_{\text{rank}})\ i\succ j$，选择第 i 个染色体
18 if $(i_{\text{rank}}=j_{\text{rank}})$ and $(i_{\text{distance}}>j_{\text{distance}})\ i\succ j$，选取第 i 条染色体
19 End if
20 选取前 J 条染色体
21 End for
22 获得 pareto 最优解，抽取种子节点
23 Return

4 实验结果

4.1 数据集

本文采用 DBLP[1]数据集和 BlogCatalog[13]数据集，数据集统计信息如表 2 所示。

4.2 度量标准

在精度方面，本文使用 3 个经典的标签分类度量标准 Macro-F1 值、Micro-F1 值、汉明损失值[1,2,21]来衡量多标签分类的精度，其中汉明损失是用来衡量标签预测的失真性。

在降低时间消耗方面，本文用时间降低百分比 t_s 来衡量：

$$t_s=\left(t_a-t_N\right)\Big/t_a \tag{2}$$

表 2 数据集信息

数据集	DBLP	BlogCatalog
节点数目	8865	10312
链接数	12989	333983
标签数	12	39
网络密度	3.3×10^{-4}	6.3×10^{-3}
节点最大度数	86	3992
节点平均度数	3	65

在减少占用内存方面用 c_s 来衡量：

$$c_s = (c_a - c_N)/c_a \tag{3}$$

其中，t_a, c_a 为比较算法所花费的时间，内存；t_N, c_N 为本文 NAMESEA 算法所花费的时间，内存。

4.3 实验结果分析

4.3.1 NSGA2 中的实验参数设计　在本文 NSGA2 算法中，种群成员之间交叉概率为 0.8，变异概率为 0.1。本文对社交网络环境下的种群代数和规模进行了测试，以寻找使 pareto 最优解前沿分布较好的结果。本文分别对种群规模为 50, 70, 90, 100，代数分别为 100, 150, 200 进行了实验，结果发现当种群规模为 90，代数为 150 时在社交网络环境下节点进化得到的结果最好。图 1，图 2，图 3 为种群规模 N 分别为 50, 70, 90，代数 gn 分别为 100, 150, 200 的实验结果 3 维图，从图 3 可以看出 pareto 解分布均匀，效果较好。其中坐标参量 t 为所选染色体在目标 1 上的代表值；c 为所选染色体在目标 2 上的代表值；SF 为所选染色体在目标 3 上的代表值。

4.3.2 NAMESEA 算法实验结果　根据上述实验本文选择 J 为 90, gn 为 150，在得到 90 种不同种子节点选择策略后分别从中随机抽取 5, 10, 15 个方案进行实验对比，求取均值以验证 NAMESEA 算法在时间消耗和内存占用减少上的提高，具体分析如下：

表 3 和表 4 记录了在时间消耗和内存占有上的具体值。我们可以看到在两个消耗中本文的 NAMESEA 算法相比随机抽取(random)和 SHDA 算法均要少 100~200，该数值是经过归一化处理后的统计值，若与无归一化处理的值相比，随着网络规模的扩大与复杂性增长，该值将呈线性增长。

表 5 和表 6 记录了 NAMESEA/random, NAMESEA/SHDA, SHDA/random 在降低时间上的百分比，从表中分析我们可以得到 NAMESEA/random 降低了 10%~15%, NAMESEA/SHDA 降低了 13%~16%, SHDA/random 上升了 1%~3%；就内存占有而言，NAMESEA/random 降低了 13%~

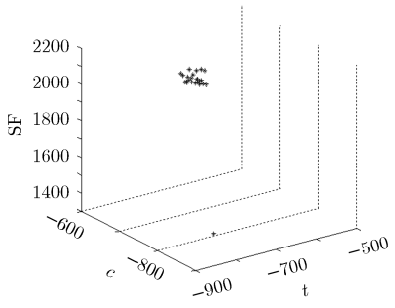

图 1　N=50, gn=100 时 pareto 最优解的前沿分布情况

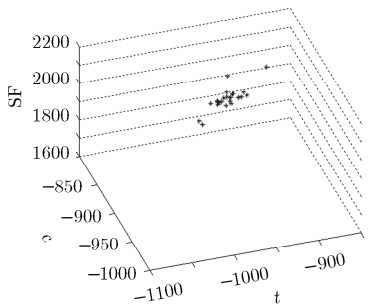

图 2　N =70, gn=200 时 pareto 最优解的前沿分布情况

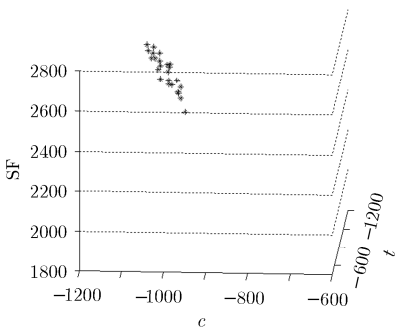

图 3　N =90, gn=150 时 pareto 最优解的前沿分布情况

表 3　DBLP 数据集上时间和内存具体消耗

		训练集占数据集比(%)								
		10			20			30		
选择策略个数		5	10	15	5	10	15	5	10	15
消耗时间(归一化值)	NAMESEA	**385.65**	**382.70**	**378.08**	**810.31**	**805.62**	**806.40**	**1130.66**	**1126.36**	**1118.77**
	random	433.78	433.78	433.78	931.24	931.24	931.24	1327.27	1327.27	1327.27
	SHDA	445.62	445.62	445.62	945.93	945.93	945.93	1331.52	1331.52	1331.52
占用内存(归一化值)	NAMESEA	**386.19**	**382.65**	**383.28**	**815.38**	**813.35**	**809.88**	**1164.01**	**1147.95**	**1121.35**
	random	439.11	439.11	439.11	941.12	941.12	941.12	1326.35	1326.35	1326.35
	SHDA	457.43	457.43	457.43	960.90	960.90	960.90	1351.01	1351.01	1351.01

表 4 BlogCatalog 数据集上时间和内存的具体消耗

		训练集占数据集比(%)								
		15			25			35		
选择策略个数		5	10	15	5	10	15	5	10	15
消耗时间(归一化值)	NAMESEA	688.56	682.56	675.20	1039.97	1032.93	1039.83	1458.35	1451.56	1454.36
	random	773.10	773.10	773.10	1174.36	1174.36	1174.36	1679.45	1679.45	1679.45
	SHDA	795.74	795.74	795.74	1195.43	1195.43	1195.43	1703.50	1703.50	1703.50
占用内存(归一化值)	NAMESEA	701.65	696.70	682.98	1041.97	1044.96	1045.82	1472.66	1466.53	1464.19
	random	797.12	797.12	797.12	1192.05	1192.05	1192.05	1695.00	1695.00	1695.00
	SHDA	812.45	812.45	812.45	1210.32	1210.32	1210.32	1736.87	1736.87	1736.87

表 5 DBLP 数据集上时间和内存降低比例

		训练集占数据集比(%)								
		10			20			30		
选择策略个数		5	10	15	5	10	15	5	10	15
消耗时间(%)	NAMESEA/random	11.10	11.78	12.84	12.99	13.49	13.41	14.81	15.41	15.70
	NAMESEA/SHDA	**14.13**	**14.12**	**13.14**	**14.34**	**14.83**	**14.73**	**15.09**	**15.41**	**15.98**
	SHDA/random	−2.73	−2.73	−2.73	−1.58	−1.58	−1.58	−0.32	−0.32	−0.32
占用内存(%)	NAMESEA/random	12.05	12.17	12.71	13.36	13.58	13.95	12.24	13.45	15.46
	NAMESEA/SHDA	**15.57**	**16.35**	**16.21**	**15.14**	**15.36**	**15.72**	**13.84**	**15.03**	**16.99**
	SHDA/random	−4.17	−4.17	−4.17	−2.10	−2.10	−2.10	−1.86	−1.86	−1.86

表 6 Blogcatalog 数据集上时间和内存的降低比例

		训练集占数据集比(%)								
		15			25			35		
选择策略个数		5	10	15	5	10	15	5	10	15
消耗时间(%)	NAMESEA/random	10.94	11.71	12.66	11.44	12.04	11.46	13.17	13.57	13.40
	NAMESEA/SHDA	**13.47**	**14.22**	**15.15**	**13.00**	**13.59**	**13.02**	**14.39**	**14.78**	**14.63**
	SHDA/random	−2.93	−2.93	−2.93	−1.79	−1.79	−1.79	−1.43	−1.43	−1.43
占用内存(%)	NAMESEA/random	11.98	12.60	14.32	12.59	12.34	12.27	13.12	13.48	13.62
	NAMESEA/SHDA	**13.63**	**14.25**	**15.94**	**13.91**	**13.66**	**13.59**	**15.21**	**15.56**	**15.70**
	SHDA/random	−1.92	−1.92	−1.92	−1.53	−1.53	−1.53	−2.47	−2.47	−2.47

17%, NAMESEA/SHDA 降低了 14%~17%, SHDA /random 上升了 1%~4%。

综上分析，我们可以得到 NAMESEA 算法选择的种子集在降低时间消耗，内存占用上相比随机选择节点，SHDA 算法要好很多，经过 NSGA2 算法建立目标模型决策后的种子集达到目标要求。

表 7 和表 8 可以看出在 Micro-F1, Macro-F1 评估指标下，NAMESEA+wvRN 算法相比 wvRN, SCRN 均提高了 1%左右，这是因为目标 3 会优先考虑社会维度特征值高的节点，保证所选种子节点在社交网络中的重要地位。但相比 SHDA+wvRN 和 SHDA+SCRN 要低 1%~3%。这是因为 SHDA 算法在设计时优先考虑的是精度的提高，而 NAMESEA 在设计时优先考虑的是系统开销小，对于那些需要花费大量时间和内存才得到信息的节点进行了一定程度上的摒弃。所以精度相比 SHDA 难免会有所下降，但在表 5 和表 6 中 NAMESEA 算法相比 SHDA 算法在节约系统开销方面提高了 14%~17%。本文的主要目标是在减小时间消耗和内存占有的前提下一定程度上提高多标签预测的精度，NAMESEA 算

表 7　DBLP 数据集上的实验结果

		训练集占数据集比(%)								
		10			20			30		
选择策略个数		5	10	15	5	10	15	5	10	15
Micro-F1(%)	NAMESEA+wvRN	39.36	39.61	39.77	43.23	43.86	43.48	47.20	47.02	47.29
	wvRN	38.13	38.13	38.13	42.13	42.13	42.13	46.59	46.59	46.59
	SHDA+wvRN	40.71	40.71	40.71	45.63	45.63	45.63	48.04	48.04	48.04
	NAMESEA+SCRN	43.26	43.95	44.08	49.89	48.77	49.08	52.82	53.41	52.81
	SCRN	43.36	43.36	43.36	48.43	48.43	48.43	52.54	52.54	52.54
	SHDA+SCRN	46.52	46.52	46.52	50.82	50.82	50.82	54.36	54.36	54.36
Macro-F1(%)	NAMESEA+wvRN	31.62	31.90	32.93	35.98	36.51	36.24	38.53	39.42	39.76
	wvRN	31.53	31.53	31.53	35.03	35.03	35.03	38.43	38.43	38.43
	SHDA+wvRN	33.63	33.63	33.63	37.31	37.31	37.31	41.60	41.60	41.60
	NAMESEA+SCRN	35.88	35.91	35.62	40.69	40.03	39.99	43.45	43.51	43.96
	SCRN	35.58	35.58	35.58	40.16	40.16	40.16	42.12	42.12	42.12
	SHDA+SCRN	38.18	38.18	38.18	42.23	42.23	42.23	44.98	44.98	44.98
汉明损失(%)	NAMESEA+wvRN	19.19	19.03	18.89	18.30	18.09	18.21	17.32	16.72	16.32
	wvRN	19.94	19.94	19.94	18.65	18.65	18.65	17.21	17.21	17.21
	SHDA+wvRN	16.53	16.53	16.53	17.01	17.01	17.01	15.05	15.05	15.05
	NAMESEA+ SCRN	15.14	15.04	15.08	13.90	14.06	13.96	13.52	13.27	13.10
	SCRN	15.83	15.83	15.83	14.39	14.39	14.39	13.49	13.49	13.49
	SHDA+SCRN	14.61	14.61	14.61	13.63	13.63	13.63	12.32	12.32	12.32

表 8　Blogcatalog 数据集上的实验结果

		训练集占数据集比(%)								
		15			25			35		
选择策略个数		5	10	15	5	10	15	5	10	15
Micro-F1(%)	NAMESEA+wvRN	25.68	25.89	26.32	27.32	27.75	28.32	28.89	29.32	29.74
	wvRN	25.39	25.39	25.39	27.38	27.38	27.38	28.46	28.46	28.46
	SHDA+wvRN	25.85	25.85	25.85	28.23	28.23	28.23	29.11	29.11	29.11
	NAMESEA+ SCRN	23.04	23.40	23.89	26.02	26.32	26.89	29.10	29.32	29.67
	SCRN	23.61	23.61	23.61	25.32	25.32	25.32	28.03	28.03	28.03
	SHDA+SCRN	24.26	24.26	24.26	27.91	27.91	27.91	29.39	29.39	29.39
Macro-F1(%)	NAMESEA+wvRN	9.84	10.42	10.49	10.71	10.97	11.11	13.23	14.05	14.22
	wvRN	9.02	9.02	9.02	11.37	11.37	11.37	13.84	13.84	13.84
	SHDA+wvRN	11.34	11.34	11.34	14.91	14.91	14.91	13.68	13.27	13.27
	NAMESEA+ SCRN	9.06	9.15	9.54	10.16	10.54	11.29	12.32	12.49	14.22
	SCRN	9.12	9.12	9.12	10.97	10.97	10.97	11.14	11.14	11.14
	SHDA+SCRN	11.78	11.78	11.78	12.57	12.57	12.57	13.57	13.57	13.57
汉明损失(%)	NAMESEA+wvRN	5.24	5.18	5.21	5.20	5.10	4.99	5.04	4.95	4.90
	wvRN	5.44	5.44	5.44	5.25	5.25	5.25	5.19	5.19	5.19
	SHDA+wvRN	5.48	5.48	5.48	5.10	5.10	5.10	4.91	4.91	4.91
	NAMESEA+ SCRN	5.65	5.36	5.46	5.32	5.15	5.09	4.94	4.92	4.90
	SCRN	5.58	5.58	5.58	5.20	5.20	5.20	4.92	4.92	5.16
	SHDA+SCRN	5.45	5.45	5.45	5.16	5.16	5.16	5.01	5.01	5.01

法相比 wvRN, SCRN 在精度上提高证明了这一点，所以就精度而言 NAMESEA 算法达到了目标。

在汉明损失 (Hamming Loss) 指标上，NAMESEA+wvRN 算法相比 wvRN 下降了 1%, NAMESEA+SCRN 算法相比 SCRN 也下降了 1% 左右，但是相比 SHDA+ wvRN 和 SHDA+SCRN 要高 1%~2%。总体而言，NAMESEA 算法在一定程度上保证了多标签节点预测的真实性。综上所述，NAMESEA 算法解决了实际中遇到的问题，大大降低了在社交网络中运行多标签算法所消耗的时间和内存占有，同时保证了多标签预测的精度。

5 结束语

本文分析了随着社交网络规模扩大和复杂性增强所导致地得到节点信息要消耗大量时间和内存占有等额外系统开销问题，针对现有多标签分类方法上的不足提出了基于 NSGA2 算法的多标签种子节点选择算法(NAMESEA 算法)。基于真实数据集，NAMESEA 算法与其他多标签预测算法在多个真实数据集上进行实验对比，证实运用该算法能够大大节约时间消耗和内存占有且一定程度上提高预测标签的精度。

在未来的下一步工作中，将加入新的目标、方法模型以丰富增强多标签种子节点的功能，使之更好地应用到在线社交网络中。

参 考 文 献

[1] WANG X and SUKTHANKAR G. Multi-label relational neighbor classification using social context features[C]. Proceedings of the 15th ACM SIGKDD International Conference on knowledge Discovery and Data Mining, Chicago, USA, 2013: 464–472.

[2] 吴信东, 赵银凤, 李磊. 基于种子节点选择的网络环境下多标签分类算法研究[J]. 电子学报, 2016, 44(9): 2074–2080. doi: 10.3969/j.issn.0372-2112.2016.09.008.
WU Xingdong, ZHAO Yinfeng, and LI Lei. Multi-label classification in network environments via seed nodes selection[J]. *Acta Electronica Sinica*, 2016, 44(9): 2074–2080. doi: 10.3969/j.issn.0372-2112.2016.09.008.

[3] LI Lei, HE Jianping, WANG Meng, *et al*. Trust agent-based behavior induction in social networks[J]. *IEEE Intelligent Systems*, 2016, 30(1): 24–30. doi: 10.1109/ MIS.2016.6.

[4] 许宇光, 潘惊治, 谢惠扬. 基于最小点覆盖和反馈点集的社交网络影响最大化算法[J]. 电子与信息学报, 2016, 38(4): 795–802. doi: 10.11999/JEIT160019.
XU Yuguang, PAN Jingzhi, and XIE Huiyang. Minimum vertex covering and feedback vertex set-based algorithm for influence maximization in social network[J]. *Journal of Electronics & Information Technology*, 2016, 38(4): 795–802. doi: 10.11999/JEIT160019.

[5] 陈季梦, 陈佳俊, 刘杰, 等. 基于结构相似度的大规模社交网络聚类算法[J]. 电子与信息学报, 2015, 37(2): 449–454. doi: 10.11999/JEIT140512.
CHEN Jimeng, CHEN Jiajun, LIU Jie, *et al*. Clustering algorithms for large-scale social networks based on structural similarity[J]. *Journal of Electronics & Information Technology*, 2015, 37(2): 449–454. doi: 10.11999/JEIT140512.

[6] ZHANG M and ZHOU Z. A k-nearest neighbor based algorithm for multi-label classification[C]. Proceedings of the IEEE International Conference on Granular Computing, Beijing, China, 2005: 718–721.

[7] HULLER E, FURNKRANZ J, CHENG W, *et al*. Label ranking by learning pairwise preferences[J]. *Artificial Intelligence*, 2008, 172(16): 1897–1916. doi: 10.1016/j.artint.2008.08.002.

[8] MACSKASSY S and PROVOST F. A simple relational classifier[C]. Proceedings of the Second Workshop on Multi-Relational Data Mining at ACM SIGKDD, Washington, DC, USA, 2003: 64–76.

[9] BOUTELL M R, LUO Jiebo, SHEN Xipeng, *et al*. Learning multi-label scene classification[J]. *Pattern Recognition*, 2004, 37(9): 1757–1771. doi: 10.1016/j.patcog.2004.03.009.

[10] 刘世超, 朱福喜, 甘琳. 基于标签传播概率的重叠社区发现算法[J]. 计算机学报, 2016, 39(4): 717–729. doi: 10.11897/SP.J.1016.2016.00717.
LIU Shichao, ZHU Fuxi, and GAN Lin. A label-propagation-probability-based algorithm for overlapping community detection[J]. *Chinese Journal of Computers*, 2016, 39(4): 717–729. doi: 10.11897/SP.J.1016.2016.00717.

[11] 邢千里, 刘列, 刘奕群, 等. 微博中用户标签的研究[J]. 软件学报, 2015, 26(7): 1626–1637. doi: 10.13328/j.cnki.jos.004655.
XING Qianli, LIU Lie, LIU Yiqun, *et al*. Study on user tags in Weibo[J]. *Journal of Software*, 2015, 26(7): 1626–1637. doi: 10.13328/j.cnki.jos.004655.

[12] ZHANG Ling and ZHOU Zhihua. A lazy learning approach to multi-label learning[J]. *Pattern Recognition*, 2007, 40(7): 2038–2048. doi: 10.1016/j.patcog.2006.12.019.

[13] TANG L and LIU H. Scalable learning of collective behavior based on sparse social dimensions[C]. Proceedings of the ACM CIKM, Hong Kong, China, 2009: 1107–1116.

[14] 申超波, 王志海, 孙艳歌. 基于标签聚类的多标签分类算法[J]. 软件, 2014, 33(8): 16–21. doi: 10.3969/j.issn.1003-6970.2014.08.004.
SHEN Chaobo, WANG Zhihai, and SUN Yange. A multi-label classification algorithm based on label clustering[J]. *Software*, 2014, 33(8): 16–21. doi: 10.3969/j.issn.1003-6970.2014.08.004.

[15] 郑伟, 王朝坤, 刘璋, 等.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. doi: 10.3724/SP.J.1016.2010.01418.
ZHENG Wei, WANG Chaokun, LIU Zhang, *et al*. A multi-label classification algorithm based on random walk model[J]. *Chinese Journal of Computers*, 2010, 33(8): 1418–1426. doi: 10.3724/SP.J.1016.2010.01418.

[16] 张振海, 李士宁, 李志刚, 等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展, 2013, 50(6): 1177–1184.
ZHANG Zhenhai, LI Shining, LI Zhigang, *et al*. Multi-label feature selection algorithm based on information entropy[J]. *Journal of Computer Research and Development*, 2013, 50(6): 1177–1184.

[17] KALYANMOY D, AMRIT P, SAMEER A, *et al*. A fast and elitist multi-objective genetic algorithm: NSGA-II[J]. *IEEE Transactions on Evolutionary Computation*, 2002, 6(2): 182–197.

[18] 刘晓娟, 闫海兰. 基于 NSGA2 算法的并行机多目标调度问题研究[J]. 物联网技术, 2013, 10(1): 43–47.
LIU Xiaojuan and YAN Hailan. Research on the multi-objective scheduling problem of parallel machine based on NSGA2 algorithm[J]. *Internet of Things*, 2013, 10(1): 43–47.

[19] 孙建龙, 吴锁平, 陈燕超. 基于改进 NSGA2 算法的配电网分布式电源优化配置[J]. 电力建设, 2014, 35(2): 86–90. doi: 10.3969/j.issn.1000-7229.2014.02.017.
SUN Jianlon, WU Suoping, and CHEN Yanchao. Optimal configuration of distributed generation in distribution network based on improved NSGA2[J]. *Electric Power Construction*, 2014, 35(2): 86–90. doi: 10.3969/j.issn.1000-7229.2014.02.017.

[20] 张利. NSGA2 算法及其在电力系统稳定器参数优化中的应用[D]. [硕士论文], 西南交通大学, 2013: 3–9.
ZHANG Li. NSGA2 Algorithm and its application in optimizing power system stabilizer parameters[D]. [Master dissertation], Southwest Jiaotong University, 2013: 3–9.

[21] NEVILLE J, GALLAGHER B, ELIASSI-RAD T, *et al*. Correcting evaluation bias of relational classifiers with network cross validation[J]. *Intelligent Systems*, 2016, 31(1), 24–30. doi: 10.1007/s10115-010-0373-1.

李　磊：男，1981 年生，副研究员，硕士生导师，研究方向为数据挖掘、社会计算、图计算.

楚喻棋：女，1993 年生，硕士生，研究方向为社会计算、多目标规划.

汪　萌：男，1984 年生，教授，博士生导师，研究方向为多媒体信息处理、数据挖掘.

韩　莉：女，1978 年生，硕士生，研究方向为公共安全与应急管理.

吴信东：男，1963 年生，教授，博士生导师，研究方向为数据挖掘、知识库系统、万维网信息探索.