基于 Multi-agent 理论的社会网络文体分类方法①

吴家菁, 王 杨, 闫小敬, 赵传信, 陈付龙

(安徽师范大学 数学计算机科学学院, 芜湖 241000)

摘 要: 针对当前社会网络中的文体分类存在分类效果不理想问题, 结合网络文体的多样性、多归属性及动态性的特征, 提出了一种基于 multi-agent 的属性融合和词库关联的网络文体分类方法. 首先提取网络文体的特征关键词和词义等基本属性, 建立 Multi-agent 的融合分类模型, 并给出了基于 Multi-agent 的社会网络文体融合分类算法. 实验结果表明该方法与传统单分类器以及其他多分类器融合分类方法相比, 不仅可以通过语义特征提取对语义网络中的网络文体进行高精度分类, 而且可以实现社会网络文体分类的自动化, 具有更高的分类精度与稳定性.

关键词: 社会网络; 特征提取; multi-agent 融合; 文体分类

Social Network Style Classification Method Based on Multi-agent Theory

WU Jia-Jing, WANG Yang, YAN Xiao-Jing, ZHAO Chuan-Xin, CHEN Fu-Long

(School of Mathematics & Computer Science, Anhui Normal University, Wuhu 241000, China)

Abstract: Recently, there are some problems like extracting hardly and lacking classification methods in stylistic classification of social networks. Combining network stylistic diversity, multi-attribution and dynamic characteristics .A attribute fusion and thesaurus associated method based multi-agent has been proposed from feature extraction. Firstly, it extracts the basic attributes of keywords and meaning of characteristics. Then, a multi-agent fusion classification model has been established with the interaction of multi-agent and it also gives the algorithm of the model. The experimental results show that this method which compares with the traditional single fusion classification classifier and other multi-classifier fusion classification not only achieves the high-precision network stylistic classification in semantic network through Semantic features extraction but also receives Social Network stylistic classification's automation. The method has a higher accuracy classification and stability.

Key words: social networks; feature extraction; multi-agent fusion; stylistic classification

近年来, 随着 Facebook、Twitter 等网络新媒体的迅速产生和发展, 社会网络成为目前学术界与产业界共同关注的热点之一[1]. 社会网络文化引发了信息产生方式和传播模式的改变, 一方面社会网络呈现出了动态变化、形式多样化的特征; 另一方面, 社会网络衍生出不同的网络文体, 如淘宝体、甄嬛体、陈欧体等[2]. 网络文体是指起源或流行于网络的新文体, 通常是由一个突发奇想的帖子、一次集体娱乐或者是一个热点事件而产生, 网络文体一般形式自由, 特点鲜明, 在一段时间内会引起较高的关注度. 由于网络文体能够引导网络舆情的演化, 因此针对网络文体的相关研究具有重要的理论意义与现实意义[3].

目前, 针对社会网络环境下的文体挖掘鲜有研究. 传统的文本分类方法有以下两种, 从单分类器融合角度主要包括朴素贝叶斯方法[4]、最大熵算法[5]和支持向量机方法[6]等. 为了提高分类器算法的精度和性能, 研

① 基金项目:安徽省自然科学基金(1308085QF118);安徽师范大学创新基金(2012cxjj09);教育部人文社科青年基金(11YJC880119)
收稿时间:2014-03-19;收到修改稿时间:2014-04-29

究者们进一步提出了从多分类器融合角度进行多属性融合的分类方法[7,8]. 如文献[9]提出了一种新的基于神经网络的融合规则, 并以此建立了一个新的多分类器组合模型, 能够一定程度上提高分类的精度和稳定性. 此外, 还有研究者针对文本特征进行相关分类方法的研究, 如文献[10]从文本分类中特征选择方法的比较与改进进行了相关研究.

上述方法从不同角度针对文本的不同特征进行分类, 提高了分类的效果, 但忽略了社会网络文体的多样性、多归属性及动态性的特征, 单一的分类器难以全面的考虑文体的多分类和多样性. 本文在分析综合上述相关研究的基础上, 结合目前对社会网络文体分类方法的不足及 Agent 可以通过语义特征提取进行高精度分类且可以实现社会网络文体分类的自动化的特性, 将 multi-agent 的理论应用于网络文体的分类问题, 以提升分类器的性能.

1 问题描述与相关定义

通常用图 $G=(V,E)$ 来描述社会网络; 节点 $i\in V$ 表示社会网络中的某一个个体对象, 一条边 $e\in E$ 表示节点个体间的关系. 而本文将富含语义的社会网络定义为一个五元组 $U=\{No.,D,N,L,S\}$, 由带有特征标记的结点和带有特征标记的链所组成的相互关联的复杂网络. 结点表示富含语义的节点文体, 边表示节点文体相互之间的语义关系; $No.(number)$ 为待分类的网络文体的编号; $D(Designation)$ 称为分类文体的名称; $N(Nodes)$ 为结点, $L(Lines)$ 为结点之间的边.

社会网络文体文本构成的规格和模式, 是一种反映了文本从内容到形式的整体特点, 但文体内容的特征难以表示, 下面给出空间向量模型[13](VSM)进行的处理表示.

定义 1. 文体空间向量: 将网络文体表示成为一组正交词条的 n 维空间向量, 通过向量的方式来计算相似度. 若将每个文体表示为特征向量, d_i 为其中一个特征矢量:

$$d_i=\{t_1,w_1;t_2,w_2;\cdots;t_k,w_k;\cdots;t_n,w_n\} \tag{1}$$

其中 t_k 表示词语, w_k 表示词语在 t_k 文本 d_i 中的权值. 定义 $w_i(d)$ 为 t_i 在 d 中出现频率 $tf_i(d)$ 的函数, 则:

$$w_i(d)=\psi(tf_i(d)) \tag{2}$$

常用的 ψ 有布尔函数, 对数函数, TFIDF 函数等. 其中 N 为文体总数, n_i 是文体在所有文体集合中出现的次数. 文体的特征及其划分, 往往取决于其层面结构中某些因素的强化、突出或变异. multi-agent 在用于分类时, 首先将单分类器输出的度量值作为初状态输入到各 Agent 通过引入决策共现矩阵, 以及利用分类器之间的决策相关信息, 在 Agent 之间进行信息交流, 指导各个 Agent 向不同类别溯源, 从而通过 Agent 之间的信息交互改变溯源概率, 最终得到群体决策, 即多分类器融合结果. 在进行类别相似度计算时, 仅考虑两个特征向量中所包含的词条的重叠程度, 即:

$$sim(d_k,c_i)=\frac{n\cap(d_k,c_i)}{n\cup(d_k,c_i)} \tag{3}$$

其中, $n\cap(d_k,c_i)$ 是 d_k 和 c_i 具有相同词条的数目, $n\cup(d_k,c_i)$ 是 d_k 和 c_i 具有的所有词条数目.

2 基于multi-agent的网络模型

2.1 语义特征值提取规则

基于 multi-agent 的语义特征值提取社会网络文体分类方法明确了各类网络文体之间的联系和规则, 细化了分类的任务, 实现了社会网络文体分类的自动化.

定义 2. Agent 网络系统空间: 定义为由多个系统空间中的节点 Agent 个体组成, 用集合 $<I,E>$ 表示.

图 1 给出了 Agent 网络系统空间示意图(本文主要采用 Microsoft Visio 绘图软件画图), 包括中心个体 Agent、交互 Agent、控制 Agent 和分析 Agent 等, 它们通过各自不同的功能与特点, 相互合作共同完成对网络的划分和分类任务.

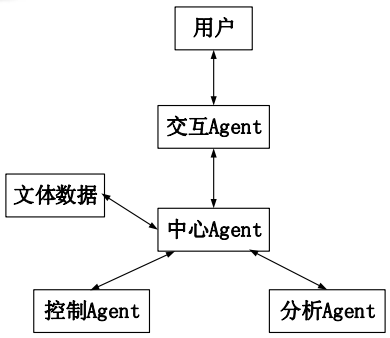

图 1 Agent 网络系统空间

其中 I 为 Agent 内部集合, 包括中心个体 Agent;

$$I=\{[Agent\ i,type],Agent\ i\in G,type\in T\}$$

G 表示任务集, T 表示分类. 中心个体 Agent 是语义

网络的整个系统空间中最核心的部分, 把握整个系统空间的资源和信息, 其他 Agent 在进入系统模型和划分分类完成后退出系统模型都需要由它进行管理, 其他Agent向它发出请求后, 中心个体Agent协调各个模块 Agent, E 为 Agent 外部集合, 包括交互 Agent、控制 Agent 和分析 Agent 等; $E=\{[Agent\ i,A],Agenti\in G,action\in A\}$. G 表示任务集, A 表示个体Agent的动作和行为. 交互Agent是人和计算机组成的一个整体, 交互 Agent 建立后, 通过对人与计算机进行交互设计, 让人与计算机之间建立一种有机关系, 从而可以有效达到人(即使用者)的目标. 控制 Agent 能够自行的控制其状态和行为, 多个控制 Agent 共同完成对语义网络的整个系统空间的划分和分类任务. 分析 Agent 当对语义网络的整个系统空间的划分和分类需要作出结果分析的时候, 就由分析 Agent 对系统工作做出相关分析.

2.2 Multi-agent 融合分类模型

在 multi-agent 模型中, 针对样本 x, 定义分类可信度矩阵 B 以及决策共现矩阵 D 分别表示文体的可信度以节点文体个体相遇时需要交换的信息量. 定义可信度矩阵为 $B(x)=[b_{ki}(x)_{K\times M}]$, $b_{ki}(x)$ 表示第 K 类认为 i 节点文体是该类的可信度. 则:

$$b_{ki}(x)=P(E(x)=i\,|\,e_k(x)=j_k) \tag{4}$$

$e_k(x)=j_k$ 表示分类器 k 的决策为 j_k, $E(x)=i$ 表示融合的最终决策为 i. 若设 Bayes 规则中的混淆矩阵 C 得到样本 x 的分类可信度矩阵为 $B(x)$;

C 为 K 个 $M*M$ 的矩阵 $C^{(k)}(k=1,...,k)$, 矩阵的 $C^{(k)}$ 元素 $C^{(k)}$ 表示分类器 K 将第 i 类样本划分为第 j 类的个数, 则有:

$$b_{ki(x)=P(E(x)=i|e_k(x)=j_k)}=c_{ij_k}^{(k)}\Big/\sum_{i=1}^{M}c_{ij_k}^{(k)} \tag{5}$$

定义决策共现矩阵

$$D=[d_{j_1,j_2,i,k_1,k_2}]_{M\times M\times M\times K\times K} \tag{6}$$

表示两个分类器间的决策相关性.

$$d_{j_1,j_2,i,k_1,k_2}=P(F=i\,|\,f_{k_1}=j_{k_1},f_{k_2}=j_{k_2}) \tag{7}$$

其中 d_{j_1,j_2,i,k_1,k_2} 分类器 k_1 将第 i 类分为第 j_1 类, 同时分类器 k_2 将第 i 类分为第 j_2 类的频率. 将分类器的信息传给 Agent 对应于 multi-agent 系统中, 表示当分类器 f_{k_1} 的决策为 j_1 类时, f_{k_2} 的决策为 j_2 类时, 其两者的 Agent 之间决策为 i 时进行交换的信息量.

3 基于Multi-agent的融合分类算法

3.1 基于智能 Agent 的特征提取算法

通过对社会网络文体进行语义特征提取, 并建立基于 multi-agent 的融合分类模型. 本小节对所获取的基本特征进行遍历、分类和组合, 首先以特征作为计算第 i 层向量之间的相似度 $sim(d_k,c_i)$ 并与阈值 T(Threshold)相比较, 若 $sim(d_k,c_i)<T$, 则停止分类, 若 $sim(d_k,c_i)>T$, 则继续进一步分类, 如此反复, 之后的每一层出现分类精度超过前面最好的分类精度. 图 2 描绘了具体的过程, 算法 1 描述了具体的特征值提取. 在以下实验中, 我们将充分验证这一特征值提取算法的有效性.

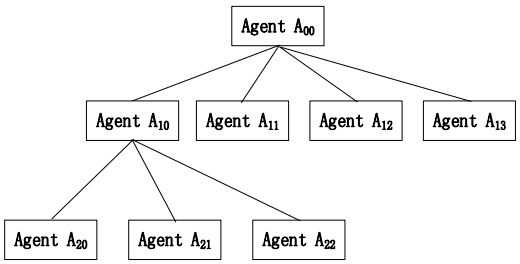

图 2　基于智能 Agent 理论的特征值提取过程

算法 1 描述:

Step1. 对语义网络系统空间进行总分类, 即粗分为第一层 $Agent\ \ A_{00}$;

Step2. 计算第 i 层向量之间的相似度 $sim(d_k,c_i)$ 并与该层的阈值 T(Threshold)相比较, 若 $sim(d_k,c_i)<T$, 则停止分类, 转向 Step5, 若 $sim(d_k,c_i)>T$, 则继续进一步分类, 转向 Step3;

Step3. 将所给文本对所有库进行查询, 应用算法激活该库对应的 Agent, 细分为第 i+1 层 $Agent\ \ A_{ij}$;

Step4. 重复上述过程直至 $sim(d_k,c_i)$ 无限趋近于 T;

Step5. 根据得到的相似度值做排序(从大到小), 得出社会网络衍生出的不同的网络文体, 即文本属性;

Step6. 输出结果反馈.

3.2 Multi-agent 融合分类算法

对数据集进行特征值提取后, 需将其划分成 U_1、U_2 和 U_3 三部分, 用融合分类器对训练集 U_1 设计 K 个不同的分类器, 并对 U_2 和 U_3 上的样本进行决策分类.

算法 2: Muleti-Agent 融合分类算法

Input:

$C^{(k)}$ //混淆矩阵

U_3, D //共现矩阵

Output:

S_i //融合分类后的结果

Begin

Step1. 计算样本 x 的分类可信度矩阵 $B(x)$;

Step2. 定义各种网络文体的类属概率矩阵 $Z=[z_{ki}]_{K\times M}$,其元素 z_{ki} 表示第 k 种文体属于第 i 类文体的概率, 矩阵 Z 的行和为 1; 初始时, 设 $Z=B(x)$;

Step3. 定义 rate 为文体数中某一类别占总文体数目的比例; 若 $rate=1$ 表示所有文体集中于某一个, 初始值 rate 为各个类别的集中程度; 定义 dec 为集体的决策;

Step4. 若 $rate>t$, t 为阈值, 表示各类别已答成共识, 转 Step8, 否则转 Step4;

Step5. 根据 Agent 的自身情况改变矩阵 Z, $Z=z_{ki}+M$; 其中为 k 类别与其他类别在第 i 文体上交换的信息总量;

$$M=\frac{1}{K}\sum_{k_1=1,k_1\neq k}^{K} d_{j,j_1,i,k,k_1}\cdot\sqrt{z_{ki}\cdot z_{k_1,i}} \tag{8}$$

Step6. 对矩阵 Z 进行归一化;

Step7. 重新计算 rate 和 dec, 转 Step4.

Step8. 输出融合分类结果 dec.

End

4 实验与讨论

4.1 数据集

实验采集了中国知网(http://www.cnki.net/)上 5000 篇期刊文献分别作为分类样本进行实验, 采用的分词方法是中国科学院计算技术研究所的 ICTCLAS 系统, 该系统对于分好的词进行了词性的标注; 对样本集划分为前 3000 作为分类器训练集, 另外的 2000 作为两个融合测试集, 适应特定领域分类的应用要求, 进行开放性测试, 采集的语料集统计见表 1, 取其中的工程科技、农业科技、医药卫生科技、哲学人文科学、信息科技和社会科学六个大类作为语料集.

表 1 中国知网的数据集统计

类别	训练集数目	测试集 1	测试集 2
工程科技	600	143	200
农业科技	300	143	100
医药卫生科技	400	143	60
哲学人文科学	650	143	220
信息科技	350	143	150
社会科学	700	142	270

4.2 实验评估

本文通过查全率 r、查准率 p [9]进行查询结果的评价方法, 下面给出计算公式.

$$r=\frac{a}{a+c}, \tag{9}$$

$$p=\frac{a}{a+b}. \tag{10}$$

其中 a 表示分类器认为属于该类别实际也属于的文体数; b 表示分类器认为属于这个类而实际不属于的问题数, c 表示分类器认为属于这个类而不属于这个类的文体数.

4.3 实验结果

在采集的实验数据集上, 对基于 multi-agent 融合算法和特征选择方法的文体分类效果进行整体查询率和准确率的比较实验. 从表 2 和表 3 可以看出, 在当前数据样本下, 由于不同类别的自身具有不同的属性特征, 两种分类算法在工程科技、农业科技和社会科学类别均具有较高的查询率, 其他类别的查询率稍低, 显示了针对查询率低的类别在特征提取上具有一定的难度; 但是查询准确率两种方法均较高, 均在 80%以上. 因此, Multi-agent 融合具有更好的分类效果.

表 2 整体查询率比较示意图

类别	multi-agent 融合	特征选择方法
工程科技	99.45	94.62
农业科技	96.29	90.12
医药卫生科技	58.34	49.83
哲学与人文科学	62.78	54.61
信息科技	58.10	51.42
社会科学	98.38	95.23

表 3 准确率比较示意图

类别	multi-agent 融合	特征选择方法
工程科技	98.23	91.41
农业科技	92.19	85.24
医药卫生科技	90.21	83.32
哲学与人文科学	94.57	87.75
信息科技	91.42	82.22
社会科学	99.21	92.12

针对分布均匀的数据集, 图 3 和图 4 给出了四种单分类器和两种多分类器在分类精度和稳定性的比较. 单分类器, 针对的分类特征不同, 其分类的原理设计差别很大, 分类结果存在较大的偏差, 对于相同

的数据集分类效果往往不同. 在本文的数据集实验中, 单分类器中 BP 算法的分类效果最好, 精度最高达到 93.8%, 相对而言, NB 算法较差, 其最高精度仅为 84.2%; 对于多分类器融合方法来说, 由于设计原理比单分类器考虑因素更多, 分类结果较好, multi-agent 和 MV 算法的精度最高分别达到了 97.4%和 94.2%, 比单分类器精度均高; 同时, 通过对四种单分类器和两种融合算法的分类稳定性进行比较, 可以看出, 四种单分类器, BP 的分类最稳定; 多分类器融合算法中, multi-agent 融合算法的稳定明显较高. 综上, multi-agent 多分类器融合算法在分类精度和稳定性上比大多数单分类器的效果要好, 与传统的多分类器融合算法相比性能也具有一定的提高.

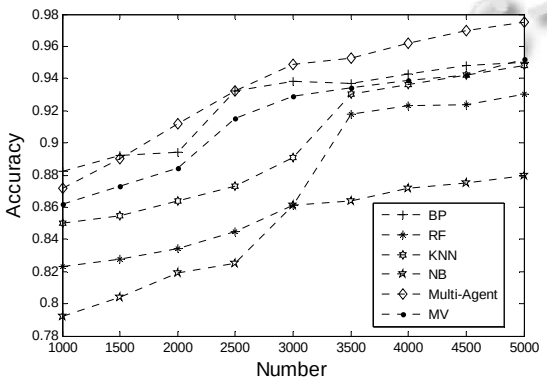

图 3　单分类器与多分类器融合的分类精度比较

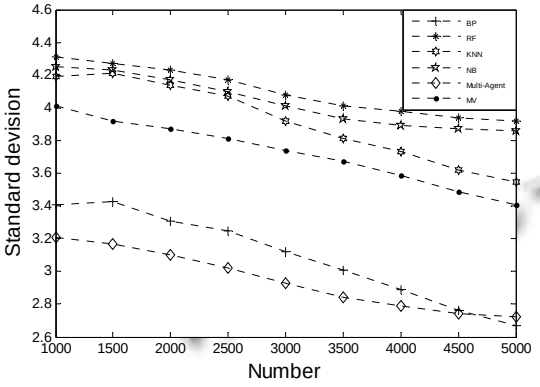

图 4　单分类器与多分类器融合的分类稳定性比较

5　结语

社会网络文体的有效分类是进行信息检索以及高效管理的重要基础, 本文将 multi-agent 理论应用于社会网络的文体分类问题, 通过语义特征提取对语义网络中的网络文体进行高精度分类, 而且可以实现社会网络文体分类的自动化, 具有更高的分类精度与稳定性. 下一步的工作将针对大数据环境下的社会网络文体分类方法的精度以及稳定性问题展开研究.

参考文献

1 崔斌.“社会网络”综述-CCF YOCSEF 学术报告会.中国计算机学会通讯,2011,7(10):74–75.
2 曾丹,吉晖.网络语言研究现状与展望.大连海事大学学报(社会科学版),2009,8(5):103–106.
3 熊伟,周水庚,关佶红.网络数据分类研究进展.模式识别与人工智能,2011,24(4):527–537.
4 Mithehell T. Machine Learning. New York: McGraw-Hill, 1997.
5 Adwait R. Maximum Entropy Models for Natural Language Ambiguity Resolution. Pennsylvania: University of Pennsylvania, 1998.
6 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
7 蒋林波,蔡立军,易叶青.一个新的多分类器组合模型.计算机工程与应用,2008,44(17):131–135.
8 Briem GJ, Benediktsson JA, Sveinsson JR. Multiple Classifies Applied to Multisource Sensing Data. IEEE Trans. Geosci Remote Sensing, 2002, 40(10): 2291–2299.
9 孟佳娜,林鸿飞,李彦鹏.基于特征贡献度的特征选择方法在文本分类中应用.大连理工大学学报,2011,51(4): 611–615.
10 单丽莉,刘秉权,孙承杰.文本分类中特征选择方法的比较与改进.哈尔滨工业大学学报,2011,43(3):319–324.
11 S. Wasserman KF. Social Network Analysis: Methods and Applications. Cambridge. UK: Cambridge University Press, 1994.
12 Aggaarwal CC. Ed. Social Network Data Analytics. NY: Springer, 2011.
13 王继成,潘金贵.Web 文本挖掘技术研究.计算机研究与发展,2000,37(5):513–520.