(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112949713 B
(45)授权公告日 2023.11.21

(21)申请号 202110227294.1
(22)申请日 2021.03.01
(65)同一申请的已公布的文献号
申请公布号 CN 112949713 A
(43)申请公布日 2021.06.11
(73)专利权人 武汉工程大学
地址 430074 湖北省武汉市洪山区雄楚大街693号
(72)发明人 曹倩倩 陈向阳
(74)专利代理机构 湖北武汉永嘉专利代理有限公司 42102
专利代理师 唐万荣
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 40/216(2020.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)
G06F 16/215(2019.01)
G06F 16/242(2019.01)
G06F 16/951(2019.01)
G06F 16/955(2019.01)
G06F 18/2415(2023.01)
G06N 20/20(2019.01)

(56)对比文件
CN 102789498 A,2012.11.21
CN 103116646 A,2013.05.22
CN 103365997 A,2013.10.23
CN 108062331 A,2018.05.22
CN 108733652 A,2018.11.02
CN 108804651 A,2018.11.13

审查员 洪汇隆

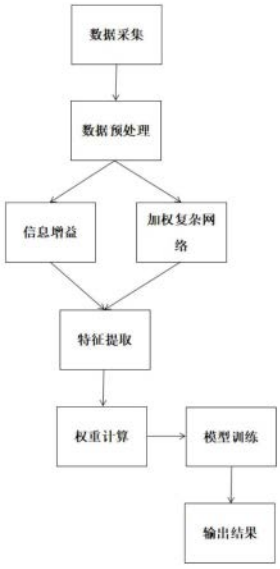
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种基于复杂网络的集成学习的文本情感分类方法

(57)摘要

本发明提供了一种基于复杂网络的集成学习的文本情感分类方法,结合现有的特征提取方法和基于复杂网络的特征选择方法,提高了对中文文本的情感分析的准确率。本发明通过实验使用集成学习结合朴素贝叶斯分类器验证了可行性,对比现有的特征提取技术和情感分类方法,本发明的分类准确率有明显的提高,得到了更好的文本情感分类效果。



CN 112949713 B

1.一种基于复杂网络的集成学习的文本情感分类方法,其特征在于:包括以下步骤:

S0:采集数据并对数据进行预处理得到原始特征集;

S1:通过现有信息增益方法对原始特征集进行特征选择,得到第一特征选择结果集;

S2:基于复杂网络综合特性对原始特征集进行特征选择,得到第二特征选择结果集;

S3:去除第一特征选择结果集与第二特征选择结果集的重复项,取并集融合得到最终特征选择结果集;具体步骤为:

定义信息增益是信息熵的差值,是移除某个变量的不确定性之后的信息量;采用信息增益算法IG计算特征项的不确定性造成的信息熵的差值,用于评价特征项对文档的重要程度,则信息增益公式为:

IG(X,Y)=E(X)-E(X|Y);

设包含特征项w的文档的概率为P(w),不包含特征项w的文档的概率为$P(\bar{w})$,属于C_i类的文档的概率为$P(C_i)$,包含特征项w且属于C_i类的文档的概率为$P(C_i|w)$,不包含特征项w且不属于C_i类的文档的概率为$P(C_i|\bar{w})$,语料库中文档类别的个数为n,则C_i类文档中是否包含特征项w的信息增益为:

$$\begin{aligned}IG(w)&=E(C_i)-E(C_i|w)\\&=-\sum_{i=1}^{n}P(C_i)\log_2 P(C_i)+P(w)\sum_{i}^{n}P(C_i|w)\log_2 P(C_i|w)+P(\bar{w})\sum_{i=1}^{n}P(C_i|\bar{w})\log_2 P(C_i|\bar{w})\\&=P(w)\sum_{i}^{n}P(C_i|w)\log_2\frac{P(C_i|w)}{P(C_i)}+P(\bar{w})\sum_{i=1}^{n}P(C_i|\bar{w})\log_2\frac{P(C_i|\bar{w})}{P(C_i)}\end{aligned}$$

按照信息增益值的降序排列特征项,提取排列靠前的一定数量的词语作为全局特征词,并保存特征词文本文件;

S4:采用TF-IDF方法对最终特征选择结果集中的特征进行权重计算;具体步骤为:

对最终特征选择结果集中的特征词进行排序,将正类放在前面,负类放在后面;采用TF-IDF算法计算特征词在不同类别中的权重,或通过SQL语句计算每一类特征的总权重;

设$n_{i,j}$表示词t_i在文档d_j中出现的次数,$\sum_k n_{i,j}$表示文档d_j中所有k个词次数的总和,定义词频TF是特征词t_i在文档d_j中出现的频率,频率越高对文档越重要,则词频TF的表达式为:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}};$$

设$|\{j:t_i\in d_j\}|$表示词t_i的文档数,定义逆文档频率IDF是包含特征词t_i的文档占总文档D的比重的倒数,用于避免出现频率高但对文档分类作用小的词获得高权重,则逆文档频率IDF的表达式为:

$$idf_i=\log\frac{|D|}{\left|\{j:t_i\in d_j\}\right|};$$

则通过表达式:

$TF\text{-}IDF=tf_{i,j}\cdot idf_i$,

表示词语对于文本的重要性随词频的增大而增大、随文档频率的增加而减小;在当前

文本中出现的次数多，并且在别的文本中出现的次数少的词语对于文本有意义；均匀出现在各个文本中的词语对文本的意义小；

S5：配置环境构建分类训练模型，利用集成学习加强朴素贝叶斯方法对数据进行分类训练并输出结果。

2.根据权利要求1所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S0中，具体步骤为：

S01：创建并运行爬虫程序，基于urllib标准库读取URL标签，利用requests库对服务器发送请求对象，利用BeautifulSoup库解析网页，获得文本数据；

S02：对文本数据进行包括清洗、分词的预处理工作得到原始特征集，并以txt格式存储为文本文档；文本文档包括停用词表、评论文本、分词后的数据，文本文档的保存格式为编号-文本-类别；

S03：采用SQL语句根据查询分析需求对评论文本进行ID编号，通过MySQL数据库对原始特征集增加主键约束。

3.根据权利要求1所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S2中，具体步骤为：

S21：以特征词为节点，连接句子中共现跨度小于或等于2的特征词，合并处在不同句子中的相同特征词节点，根据复杂网络的综合特性对预处理后的文本数据构建加权复杂网络；

S22：分别计算节点n_i的加权度、加权聚集系数和节点介数，并分别进行归一化处理；构造评估函数CF，以函数值作为节点n_i的综合特征值；

S23：对节点的函数值进行排序，选取函数值较大的前m个节点对应的特征词作为文本的关键词。

4.根据权利要求3所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S21中，具体步骤为：

用加权复杂网络的节点代表特征词，设节点的集合为：

$N=\{n_1,n_2,n_3,\ldots,n_k\}$；

用加权复杂网络的边代表特征词之间的包括共现和邻接位置的语义相关关系，设经过预处理之后的原始特征词为n，复杂网络中的结点个数为k，加权复杂网络中边的集合为：

$E=\{e_{i,j}=(n_i,n_j)\mid n_i,n_j\in N\}$；

用边的权值代表特征词的语义相关关系的程度，权值越大，表明特征词之间语义相关关系越紧密，设边e_{ij}的权重为w_{ij}，边的权重集合为：

$W=\{w_{12},w_{13},\ldots,w_{ij},\ldots\}$；

则将文本表示成加权的复杂网络为：

$G=(N,E,W)$。

5.根据权利要求4所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S22中，具体步骤为：

设节点n_i的各部分的权重β_i(1≤i≤3)是可调节的参数，则：

$\beta_1+\beta_2+\beta_3=1$；

设节点n_i与所有邻居节点n_j的边的权值为W_{ij}，则用于反映节点n_j与其他节点的连接强

度的加权度WD_i为：

$$WD_i=\sum_{(n_i,n_j)\in N}W_{ij};$$

设用于表示节点n_i邻接节点间边的权重和的节点n_i的加权聚集度为WK_i，节点n_i的度数为k_i，则加权聚集系数为：

$$WC_i=\frac{2WK_i}{k_i\left(k_i-1\right)};$$

设节点n_i的介数为p_i，以评估函数CF的函数值作为节点n_i的综合特征值，则：

$CF_i=\beta_1WD_i+\beta_2WC_i+\beta_3P_i$。

6.根据权利要求1所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S5中，具体步骤为：

S51：配置环境，确定待分类项组成的测试集，对测试集数据进行包括清洗、分词的预处理，并对测试集的每条文本进行id编号；

S52：假设各特征条件相互独立，对待分类项求解各类别出现的概率并记录为已知概率，构建包括多变量伯努利模型和多项式模型的分类训练模型；

S53：利用集成学习加强朴素贝叶斯方法对测试集数据进行分类训练，根据已知概率提取和计算待分类项的特征属于某一类别的概率，取概率最大的类别为待分类文本的所属类别并输出结果，实现对文本的情感分类。

7.根据权利要求6所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S53中，集成学习融合AdaBoost算法，通过提高前一轮分类器分类错误的样本的权值，降低分类正确的样本权值，产生多个弱分类器；通过多数加权投票组合弱分类器，加大误差率小的分类器，减少误差率大的分类器，提高分类的准确率和效率。

8.根据权利要求6所述的一种基于复杂网络的集成学习的文本情感分类方法，其特征在于：所述的步骤S53中，具体步骤为：

S531：输入数据集$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$，$x_i\in X$，X表示训练样本集空间，$Y_i\in Y=\{1,2\}$是某一类别集；每次迭代的索引为$t=1,2,\ldots,T$，通过AdaBoost算法为每个训练样本分配权重$w_i^t$；初始时，对所有i都有$w_i^1=\frac{1}{N}$；

S532：将AdaBoost算法用于朴素贝叶斯算法；在迭代过程中若训练样本x_i被错误分类，则权重w_i^{t+1}增加；若训练样本x_i被正确分类，则权重w_i^{t+1}减少；将训练样本的权重为w_i^t引入到参数$P(X_k|C_j)$，则朴素贝叶斯公式变为：

$$P\left(X_k|C_j\right)=\frac{w_{jk}e^{w_i^t}+1}{\sum_{i=1}^{|n|}W_{jk}e^{w_i^t}+|n|};$$

样本权重、朴素贝叶斯的先验概率和后验概率随着AdaBoost的迭代而更新，对朴素贝叶斯分类器的分类产生扰动，增加了朴素贝叶斯分类器的相异性。

一种基于复杂网络的集成学习的文本情感分类方法

技术领域

[0001] 本发明属于机器学习分类技术领域，具体涉及一种基于复杂网络的集成学习的文本情感分类方法。

背景技术

[0002] 随着信息科技的飞速发展，越来越多的互联网应用已经渗入到人们生活的方方面面。普通用户与网络应用之间的交互也越来越频繁，互联网用户群体的角色逐渐从互联网内容信息的浏览者演变为创造者。在这一过程中，用户可以在媒体平台提出情感态度型的观点和评论，对其进行检测和分类不仅可以产生巨大的商业价值，还可以维护互联网环境的安全。其中由于微博人口基数大，涉及话题广泛的特点，对人们的日常生活产生了不可估量的影响，而对微博的情感分析，更是有着十分重要的意义。近年来，随着复杂网络的兴起，国内外学者开始研究利用复杂网络来表示文本，根据其小世界特性进行文本挖掘，主要集中在文本的关键词提取领域。Zhu等通过构建词同现网络，利用节点缺失对网络中平均最短路径长度的影响来提取中文文本关键词。Liu等利用基于知网的词语语义相似度构建中文文本网络，结合复杂网络理论和统计方法来进行关键词提取。Huang等利用词语的句法关系建立文本复杂网络来进行关键词抽取。赵鹏等综合考虑文本语言网络中的节点度与聚集系数进行关键词抽取。在文本分类领域，赵洋等将复杂网络的分析理论引入到分类器。

[0003] 复杂网络就是结合网络的视角和基本原理的复杂系统，语言复杂网络就是用复杂网络视图研究的语言结构。Sole认为语言在各个层次上都体现了复杂网络的性质，包括语音、词法、句法和语义。语言复杂网络通常是将语言中的语素(字、词)定义为节点，将语素间的关系定义为边，常见的连接关系有：共现关系(语素同时出现在句子或单词中)，概念同义，句法关系等。

[0004] 语言网络既不是完全随机的，也不是完全规则的，它也具备复杂网络的小世界特性。复杂网络的如下重要统计特性对于语言网络同样适用：

[0005] 1.度与度分布。在复杂网络中，节点的度是指与该节点相连接的节点的数目。度分布函数则反映了网络的统计特性。

[0006] 2.聚集系数。聚集系数是用来衡量网络的集团化程度，节点i的聚集系数C指与该节点邻接的节点之间实际相连接的边数与最大可能连接边数的比值：

[0007] $$C_i = \frac{2e_i}{k_i\left(k_i - 1\right)}。$$

[0008] 其中，k表示节点i的度数，e表示节点i的邻接节点间实际存在的边数，称为聚集度。所有节点的聚集系数平均值即为该网络的聚集系数。聚集系数体现了节点的局部聚集密度及网络的聚集特性。

[0009] 3.介数。介数包括节点介数和边介数。节点的介数指网络中任意两点间的最短路径通过该节点(边)的比例。介数在一定程度上可以体现节点对整个网络信息流动的影响。除此之外，复杂网络还有平均最短路径、正负匹配度等统计特性。

[0010] 特征选择(Feature Selection)的目的是为了在文本预处理的基础上提高文本内容的类别区分能力和减少计算复杂度而对原始特征集合的降维过程。从而减少系统计算的复杂度和提高分类的准确率。常用的特征选择方法有以下几种:特征频度法(TF)、文档频率法(DF)、互信息(MI)、信息增益(IG)、期望交叉熵等。这些现有方法一般以文档频率、词频等统计信息为基础来进行特征词的选取,而忽略了文本中词汇间的语义关联关系,使得特征词的选取结果不能令人满意,从而影响了文本分类的效果。

[0011] 在文本中,离散的字、词语之间通过一定的相互关系组合在一起形成句子,从而构成了语义丰富的文本。基于语言复杂网络的文本特征选择方法结合语义学和句法学的理论、利用汉语词同现网络中的小世界特性,首先通过构造文本加权复杂网络以保留文本中的语义信息及其结构信息,然后利用节点的综合特性寻找关键节点(即中心词语),以此来作为文本的特征词,而去除那些信息量较少的词,以降低文本复杂网络的节点数目,达到降低复杂性的目的。基于复杂网络的特征选择方法考虑了词汇间的语义关联关系,但没有考虑词频等统计信息。

发明内容

[0012] 本发明要解决的技术问题是:提供一种基于复杂网络的集成学习的文本情感分类方法,用于提高对文本情感分析的准确率。

[0013] 本发明为解决上述技术问题所采取的技术方案为:一种基于复杂网络的集成学习的文本情感分类方法,包括以下步骤:

[0014] S0:采集数据并对数据进行预处理得到原始特征集;

[0015] S1:通过现有信息增益方法对原始特征集进行特征选择,得到第一特征选择结果集;

[0016] S2:基于复杂网络综合特性对原始特征集进行特征选择,得到第二特征选择结果集;

[0017] S3:去除第一特征选择结果集与第二特征选择结果集的重复项,取并集融合得到最终特征选择结果集;

[0018] S4:采用TF-IDF方法对最终特征选择结果集中的特征进行权重计算;

[0019] S5:配置环境构建分类训练模型,利用集成学习加强朴素贝叶斯方法对数据进行分类训练并输出结果。

[0020] 按上述方案,所述的步骤S0中,具体步骤为:

[0021] S01:创建并运行爬虫程序,基于urllib标准库读取URL标签,利用requests库对服务器发送请求对象,利用BeautifulSoup库解析网页,获得文本数据;

[0022] S02:对文本数据进行包括清洗、分词的预处理工作得到原始特征集,并以txt格式存储为文本文档;文本文档包括停用词表、评论文本、分词后的数据,文本文档的保存格式为编号-文本-类别;

[0023] S03:采用SQL语句根据查询分析需求对评论文本进行ID编号,通过MySQL数据库对原始特征集增加主键约束。

[0024] 按上述方案,所述的步骤S2中,具体步骤为:

[0025] S21:以特征词为节点,连接句子中共现跨度小于或等于2的特征词,合并处在不同

句子中的相同特征词节点，根据复杂网络的综合特性对预处理后的文本数据构建加权复杂网络；

[0026] S22：分别计算节点n_i的加权度、加权聚集系数和节点介数，并分别进行归一化处理；构造评估函数CF，以函数值作为节点n_i的综合特征值；

[0027] S23：对节点的函数值进行排序，选取函数值较大的前m个节点对应的特征词作为文本的关键词。

[0028] 进一步的，述的步骤S21中，具体步骤为：

[0029] 用加权复杂网络的节点代表特征词，设节点的集合为：

[0030] $N=\{n_1,n_2,n_3,\ldots,n_k\}$；

[0031] 用加权复杂网络的边代表特征词之间的包括共现和邻接位置的语义相关关系，设经过预处理之后的原始特征词为n，复杂网络中的结点个数为k，加权复杂网络中边的集合为：

[0032] $E=\{e_{i,j}=(n_i,n_j)\,|\,n_i,n_j\in N\}$；

[0033] 用边的权值代表特征词的语义相关关系的程度，权值越大，表明特征词之间语义相关关系越紧密，设边e_{ij}的权重为w_{ij}，边的权重集合为：

[0034] $W=\{w_{12},w_{13},\ldots,w_{ij},\ldots\}$；

[0035] 则将文本表示成加权的复杂网络为：

[0036] $G=(N,E,W)$。

[0037] 进一步的，所述的步骤S22中，具体步骤为：

[0038] 设节点n_i的各部分的权重β_i(1≤i≤3)是可调节的参数，则：

[0039] $\beta_1+\beta_2+\beta_3=1$；

[0040] 设节点n_i与所有邻居节点n_j的边的权值为W_{ij}，则用于反映节点n_j与其他节点的连接强度的加权度WD_i为：

[0041] $$WD_i=\sum_{(n_i,n_j)\in N}W_{ij};$$

[0042] 设用于表示节点n_i邻接节点间边的权重和的节点n_i的加权聚集度为WK_i，节点n_i的度数为k_i，则加权聚集系数为：

[0043] $$WC_i=\frac{2WK_i}{k_i\left(k_i-1\right)};$$

[0044] 设节点n_i的介数为p_i，以评估函数CF的函数值作为节点n_i的综合特征值，则：

[0045] $CF_i=\beta_1WD_i+\beta_2WC_i+\beta_3P_i$。

[0046] 进一步的，所述的步骤S3中，具体步骤为：

[0047] 定义信息增益是信息熵的差值，是移除某个变量的不确定性之后的信息量；采用信息增益算法IG计算特征项的不确定性造成的信息熵的差值，用于评价特征项对文档的重要程度，则信息增益公式为：

[0048] $IG(X,Y)=E(X)-E(X|Y)$；

[0049] 设包含特征项w的文档的概率为P(w)，不包含特征项w的文档的概率为$P\left(\overline{w}\right)$，属于$C_i$类的文档的概率为$P(C_i)$，包含特征项w且属于$C_i$类的文档的概率为$P(C_i|w)$，不包含特

征项w且不属于C_i类的文档的概率为$P(C_i|\overline{w})$，语料库中文档类别的个数为n，则C_i类文档中是否包含特征项w的信息增益为：

[0050]
$$\begin{aligned} IG(w) &= E(C_i) - E(C_i|w) \\ &= -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i) + P(w)\sum_{i}^{n} P(C_i|w)\log_2 P(C_i|w) + P(\overline{w})\sum_{i=1}^{n} P(C_i|\overline{w})\log_2 P(C_i|\overline{w}) \\ &= P(w)\sum_{i}^{n} P(C_i|w)\log_2 \frac{P(C_i|w)}{P(C_i)} + P(\overline{w})\sum_{i=1}^{n} P(C_i|\overline{w})\log_2 \frac{P(C_i|\overline{w})}{P(C_i)} \end{aligned}$$

[0051] 按照信息增益值的降序排列特征项，提取排列靠前的一定数量的词语作为全局特征词，并保存特征词文本文件。

[0052] 进一步的，所述的步骤S4中，具体步骤为：

[0053] 对最终特征选择结果集中的特征词进行排序，将正类放在前面，负类放在后面；

[0054] 采用TF-IDF算法计算特征词在不同类别中的权重，或通过SQL语句计算每一类特征的总权重；

[0055] 设$n_{i,j}$表示词t_i在文档d_j中出现的次数，$\sum_k n_{i,j}$表示文档d_j中所有k个词次数的总和，定义词频TF是特征词t_i在文档d_j中出现的频率，频率越高对文档越重要，则词频TF的表达式为：

[0056] $$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}};$$

[0057] 设$|\{j:t_i \in d_j\}|$表示词t_i的文档数，定义逆文档频率IDF是包含特征词t_i的文档占总文档D的比重的倒数，用于避免出现频率高但对文档分类作用小的词获得高权重，则逆文档频率IDF的表达式为：

[0058] $$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|};$$

[0059] 则通过表达式：

[0060] $TF\text{-}IDF = tf_{i,j} \cdot idf_i$，

[0061] 表示词语对于文本的重要性随词频的增大而增大、随文档频率的增加而减小；在当前文本中出现的次数多，并且在别的文本中出现的次数少的词语对于文本有意义；均匀出现在各个文本中的词语对文本的意义小。

[0062] 按上述方案，所述的步骤S5中，具体步骤为：

[0063] S51：配置环境，确定待分类项组成的测试集，对测试集数据进行包括清洗、分词的预处理，并对测试集的每条文本进行id编号；

[0064] S52：假设各特征条件相互独立，对待分类项求解各类别出现的概率并记录为已知概率，构建包括多变量伯努利模型和多项式模型的分类训练模型；

[0065] S53：利用集成学习加强朴素贝叶斯方法对测试集数据进行分类训练，根据已知概率提取和计算待分类项的特征属于某一类别的概率，取概率最大的类别为待分类文本的所属类别并输出结果，实现对文本的情感分类。

[0066] 进一步的，所述的步骤S53中，集成学习融合AdaBoost算法，通过提高前一轮分类器分类错误的样本的权值，降低分类正确的样本权值，产生多个弱分类器；通过多数加权投

票组合弱分类器，加大误差率小的分类器，减少误差率大的分类器，提高分类的准确率和效率。

[0067] 进一步的，所述的步骤S53中，具体步骤为：

[0068] S531：输入数据集$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$，$x_i\in X$，X表示训练样本集空间，$Y_i\in Y=\{1,2\}$是某一类别集；每次迭代的索引为$t=1,2,\ldots,T$，通过AdaBoost算法为每个训练样本分配权重$w_i^t$；初始时，对所有i都有$w_i^1=\frac{1}{N}$；

[0069] S532：将AdaBoost算法用于朴素贝叶斯算法；在迭代过程中若训练样本x_i被错误分类，则权重w_i^{t+1}增加；若训练样本x_i被正确分类，则权重w_i^{t+1}减少；将训练样本的权重为w_i^t引入到参数$P(X_k|C_j)$，则朴素贝叶斯公式变为：

[0070] $$P\left(X_k|C_j\right)=\frac{w_{jk}e^{w_i^t}+1}{\sum_{i=1}^{|n|}W_{jk}e^{w_i^t}+|n|};$$

[0071] 样本权重、朴素贝叶斯的先验概率和后验概率随着AdaBoost的迭代而更新，对朴素贝叶斯分类器的分类产生扰动，增加了朴素贝叶斯分类器的相异性。

[0072] 本发明的有益效果为：

[0073] 1.本发明的一种基于复杂网络的集成学习的文本情感分类方法，结合现有的特征提取方法和基于复杂网络的特征选择方法，提高了对中文文本的情感分析的准确率。

[0074] 2.本发明通过实验使用集成学习结合朴素贝叶斯分类器验证了可行性。

[0075] 3.对比现有的特征提取技术和情感分类方法，本发明的分类准确率有明显的提高，得到了更好的分类效果。

附图说明

[0076] 图1是本发明实施例的流程图。

[0077] 图2是本发明实施例保存的数据预处理的文本文档示意图。

[0078] 图3是本发明实施例的测试结果对比图。

具体实施方式

[0079] 下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0080] 本发明实施例针对微博文本的情感分析进行了研究，结合现有的特征提取方法和基于复杂网络的特征选择方法，首先利用传统的信息增益方法进行特征选择，然后再次对原始的特征集基于复杂网络综合特性提取特征项，最后将两者取并集去除重复项，即为最终的特征选择结果集。最终通过实验使用集成学习结合朴素贝叶斯分类器验证了方法的可行性，对比实验发现，将现有的特征提取方法和基于复杂网络的特征选择方法两者相结合的方法得到的分类效果最好。

[0081] 为满足对数据进行计算、查询、统计、分析各类需求，需要做以下两个方面的工作。

[0082] 1.对所有爬取的数据进行清洗、分词等预处理，存储在数据库当中，利用SQL语句对其增加主键约束，方便之后数据的计算、查询、统计、分析。

[0083] 2.对存储的数据集进行权重之前先刚其进行排序，从而可以提高计算、查询、统

计、分析的效率。

[0084] 参见图1,本发明实施例的一种基于复杂网络的集成学习的文本情感分类方法,包括以下步骤:

[0085] S1:创建爬虫程序,基于urllib标准库读取特定的URL标签,接下来利用requests库对服务器发送请求对象,最后利用BeautifulSoup库解析网页,最终获得所需要的文本数据信息;

[0086] 在本发明实施例中,数据集通过爬取新浪微博官网评论文本获取;也可以使用情感分类公开数据集。对于数据爬取,配置Python运行环境之后,使用pip来安装Requests类库。具体操作:在Windows平台下,运行cmd命令窗口,输入“pip3 install requests”并按Enter键,即可以安装Requests类库。

[0087] S2:对爬取的文本数据进行清洗、分词等预处理工作;使用文本文档保存已处理过的数据集。所保存的文本文档包括:哈工大停用词表、评论文本、分词后的数据,参见图2;

[0088] 预处理后的文本文档保存格式为编号-文本-类别。

[0089] 存储的数据集是txt格式。

[0090] 在对数据进行清洗,主要的操作是去除含有URL的链接,URL中带有的有用信息很少,一般都是为了广告的导向和用户的定位。

[0091] S3:使用传统的方法进行信息增益特征提取,采用传统的特征提取方法提取处理后的数据的信息增益特征;

[0092] 提取之前,使用SQLyog利用SQL语句给每条评论文本加上编号,也就是添加主键;

[0093] 其中,使用MySQL数据库对数据集增加主键约束。在本发明实施例中,训练数据集有13712条,测试集有1509条。

[0094] 为每个评论文本加上主键约束为根据查询分析需求对评论文本进行ID编号。

[0095] S4:根据复杂网络中的加权复杂网络的特性对预处理后的文本构建该网络提取特征;

[0096] 利用复杂网络的综合特性,对预处理后的文本数据构建加权复杂网络,其中节点代表特征词,边代表特征词之间的语义相关关系,在文本中体现为特征词的共现及邻接位置关系,而边的权值代表特征词的语义相关关系的程度,权值越大,表明特征词之间语义相关关系越紧密;

[0097] 复杂网络的综合特性包括复杂网络的加权度、加权聚集系数、节点介数。

[0098] 在本发明实施例中,利用评估函数CF以函数值作为节点n_i的综合特征值。$CF_i=\beta_1 WD_i+\beta_2 WC_i+\beta_3 P_i$;其中$\beta_i$(1≤i≤3)是可调节的参数,代表相应部分的权重;且有$\beta_1+\beta_2+\beta_3=1$,$p_i$节点$n_i$的介数;对节点的函数值进行排序,选取函数值较大的前m个节点对应的特征词作为文本的关键词。

[0099] 注意:其中加权度$WD_i=\sum_{(n_i,n_j)\in N} W_{ij}$是节点$n_i$与所有邻居节点$n_j$的边的权值$W_{ij}$和为该节点的加权度$WD_i$,反映了该节点与其他节点的连接强度。其中加权聚集系数$WC_i=\frac{2WK_i}{k_i(k_i-1)}$,$WK_i$为节点$n_i$的加权聚集度,表示节点$n_i$邻接节点间边的权重和,$k_i$表示节点$n_i$的度数。为了得到较好的实验效果,经过反复实验,$CF_i$中$\beta_1$取0.4、$\beta_2$取0.3、$\beta_3$取0.3。

[0100] S5:对传统方法以及加权复杂网络提取的特征进行融合取并集;

[0101] 将信息增益提取到的特征和复杂网络提取的特征进行融合取两者的并集作为最终的特征提取结果。配置python运行环境pycharm,下载使用的Python库,将其与SQL Server数据库服务器连接,将特征集合存放在数据库当中;

[0102] 信息增益(IG)算法其基本思想是计算某一变量的不确定性的存在与否造成的信息熵的差值并以此来评价该特征项对文档的重要程度。信息增益是信息熵的差值,是某个变量的不确定性在移除之后的信息量,定义如下公式:

[0103] IG(X,Y)=E(X)-E(X|Y);

[0104] 信息增益在情感分析问题中则转化为以特征项为研究对象,用特征项W在C_i类中是否出现的情况所带来的信息量,定义如下式:

[0105]
$$\begin{aligned} IG(w) &= E(C_i) - E(C_i|w) \\ &= -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i) + P(w)\sum_{i}^{n} P(C_i|w)\log_2 P(C_i|w) + P(\overline{w})\sum_{i=1}^{n} P(C_i|\overline{w})\log_2 P(C_i|\overline{w}) \\ &= P(w)\sum_{i}^{n} P(C_i|w)\log_2 \frac{P(C_i|w)}{P(C_i)} + P(\overline{w})\sum_{i=1}^{n} P(C_i|\overline{w})\log_2 \frac{P(C_i|\overline{w})}{P(C_i)} \end{aligned}$$

[0106] 公式中:P(w)为包含特征项w的文档概率;$P(\overline{w})$为不包含特征项w的文档的概率;P(C_i)为属于C_i类的文档的概率;P(C_i|w)为包含特征项w并属于C_i类的文档的概率;$P(C_i|\overline{w})$为不包含特征项w并不属于C_i类的文档的概率;n为语料库中文档类别的个数。

[0107] 计算出特征项的信息增益之后,按照信息增益值降序排列,提取前500个词语为全局特征词,保存特征词文本文件。

[0108] S6:对提取的特征进行权重计算,使用的是TF-IDF的方法;

[0109] 对特征数据集进行权重计算,权重计算之前,对有标签的微博进行排序,将正类的放在前面,负类的放在后面。利用TF-IDF权重计算可以计算每个特征词的权重值,也可以通过SQL语句计算出来每一类的总权重;

[0110] TF-IDF中的TF是指某个给定的词t_i在文档d_j中出现的频率,频率越高对文档越重要,IDF是指包含该词t_i的文档占总文档D的比重的倒数;逆文档频率的出现是为了避免一些类似“我”、“的”、“他”等出现频率很高但是对文档分类作用较小的词获得高权重。

[0111] 为了更大程度的保留文本信息,体现文本的结构和语义特征,本发明将文本表示成加权的复杂网络结构。文本加权复杂网络由许多节点和边构成,其中节点代表特征词,边代表特征词之间的语义相关关系,在文本中体现为特征词的共现及邻接位置关系,而边的权值代表特征词的语义相关关系的程度,权值越大,表明特征词之间语义相关关系越紧密;文本加权复杂网络形式化的表示为G=(N,E,W),其中N表示节点的集合N={n_1,n_2,n_3,...,n_k},n代表经过预处理之后的原始特征词,k表示复杂网络中的结点个数,E表示加权复杂网络中边的集合E={$e_{i,j}$=(n_i,n_j)|n_i,n_j∈N},W表示边的权重集合W={w_{12},w_{13},...,w_{ij},...},w_{ij}表示边e_{ij}的权重。

[0112] 基于加权复杂网络的特征选择算法通过分析加权文本复杂网络中节点的综合特性,即综合考虑节点的加权度、加权聚集系数和边介数来衡量特征词在文本中的重要性,通过构造评估函数反应节点的综合特性,体现节点的连接状况、局部密集程度、以及对网络全局的影响,从而进行文本关键词选取,以此达到特征选择的目的。

[0113] 具体的算法如下：

[0114] Step1:对文档d进行预处理

[0115] Step2:建立文本加权复杂网络,以特征词为节点,将句子中共现跨度小于或等于2的特征词连接,并将处在不同句子中的相同特征词节点进行合并处理。

[0116] Step3:分别计算节点n_i的加权度、加权聚集系数及节点介数,并分别进行归一化处理,然后构造评估函数CF,以函数值作为节点n_i的综合特征值。

[0117] Step4:对节点的函数值进行排序,选取函数值较大的前m个节点对应的特征词作为文本的关键词。

[0118] 在一些可选的实施方案中,所述方法还包括权重的计算。TF-IDF是常见的权重计算方法,考虑了问题和反文档频率的影响。对于内容中出现的次数较多的词语有相对比较大的权重。但权重计算是一个全局性的信息,没有分辨哪个特征项在哪个类别中相对比较重要的功能,不能作为区分类别的方法。一般情况下用来表示特征词语在文本中是否重要或者重要程度。通过词频和文档频率,我们采用TF-IDF权重计算方法计算出特征词在不同类别中的权重。

[0119] 词频(TF)是指某个给定的词t_i在文档d_j中出现的频率,频率越高对文档越重要,数学表达式如下：$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$;

[0120] 其中,$n_{i,j}$表示词t_i在文档d_j中出现的次数,$\sum_k n_{i,j}$表示文档d_j中所有k个词次数的总和。

[0121] 逆文档频率(IDF)是指包含该词t_i的文档占总文档D的比重的倒数。逆文档频率的出现是为了避免一些类似“我”、“的”、“他”等出现频率很高但是对文档分类作用较小的词获得高权重。数学表达式如下所示：

[0122] $$idf_i=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|},$$

[0123] 其中,$|\{j:t_i\in d_j\}|$表示词t_i的文档数。

[0124] $TF\text{-}IDF=tf_{i,j}\cdot idf_i$

[0125] 表示词语对于文本的重要性,对于词频增大的时候也随之增大,随文档频率的增加而减小。也就是对于在当前文本中出现的次数多,并且在别的文本中出现的次数少的词语对于文本有意义。均匀出现在各个文本中的词语对文本的意义小。

[0126] S7:配置环境,进行分类模型训练,利用集成的朴素贝叶斯方法对数据进行分类训练。

[0127] 在一些可选方案中构建分类训练模型,利用集成学习加强朴素贝叶斯,提高分类准确率和效率。朴素贝叶斯算法NB(Naïve Bayes)是一种非常简单的分类算法,以贝叶斯算法为基础。其基本的思想是:对于给出的待分类项,求解在此项出现的条件下各个类别出现的概率,取概率最大的那个,就认为此待分类文本属于哪一个类别。假设各特征条件相互独立。常用的模型为多变量伯努利模型和多项式模型,本篇采用多项式模型。

[0128] 该算法分为三个步骤：

[0129] Setp1:确定测试集,对测试集做和训练样本一样的预处理(去停用词,分词,每条文本进行id编号)。

[0130] Sept2:计算已知概率。程序通过之前提供的测试样本,统计每一个类别在训练样本中出现的概率,计算出每个特征出现的概率,然后将其记录,作为已知概率。

[0131] Sept3:计算分类,有了之前得到的概率,通过提取所输入文本的特征,计算这些特征属于某一类别的概率,然后通过对概率进行判断,返回概率大的结果,实现情感分类。

[0132] 融合AdaBoost算法做法的核心问题,是通过提高前一轮分类器分类错误的样本的权值,降低分类正确的样本权值,对于那些没有本分类正确的样本会得到后面分类器更多的关注。然后可以产生很多的弱分类器,通过多数加权投票组合这些弱分类器,加大误差率小的分类器,减少误差率大的分类器,使其在表决中起到较少的作用。

[0133] 算法:输入数据集,$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$,$x_i\in X$,X用来表示训练样本集空间,$Y_i\in Y=\{1,2\}$是某一类别集。每次迭代的索引为t=1,2,...,T,AdaBoost算法在训练样本上维护一套权重分布w,其中,每个训练样本都对应一个权重w_i^t,初始时,对所有i都有$w_i^1=\frac{1}{N}$。

[0134] 在本发明实施例中,将AdaBoost算法用于朴素贝叶斯算法,每次迭代过程中,训练样本x_i如果被错误分类,权重w_i^{t+1}将增加,否则w_i^{t+1}将减少。AdaBoost在迭代的时候,会为每个训练样本分配的权重为w_i^t,然后将其引入到参数$P(X_k|C_j)$中,则之前的朴素贝叶斯公式会变为:

[0135] $$P\left(X_k|C_j\right)=\frac{w_{jk}e^{w_i^t}+1}{\sum_{i=1}^{|n|}W_{jk}e^{w_i^t}+|n|}$$

[0136] 因此,随着AdaBoost的每次迭代,样本权重每次都会更新,朴素贝叶斯的先验概率和后验概率都有变化,对朴素贝叶斯分类器的分类产生了扰动,增加了朴素贝叶斯分类器的相异性。

[0137] 表1.本发明中含有URL的微博数量统计

[0138]

统计项	结果
含有URL的微博数量	1756(总量:4780)
平均引用次数	2.72

[0139] 表2.本发明数据集信息

[0140]

训练集	测试集	特征词	类别
13712	1509	383	2

[0141] 表3.本发明前9条分类结果呈现表

[0142]

id	文本（微博内容）	类别
1	别样,不二,麦子,粽,小葱,二饼,专业,生日快乐	1
2	今天,拿下,第一,功臣,回,江苏,涨工资	1
3	妈,房间,隔音,效果,差	0
4	喜欢,表个态,瞬间,乐翻	1
5	涝,不愿旱,下,耶	1
6	看到,很多,朋友,留言,说,昨天晚上,还,没,看够,今晚,继续,分享,下集,咯,新一轮,倒计时	1
7	大,夜里,洗衣服,最,感觉	0
8	不由,孙辈们,肃然起敬	0
9	意外,收到,朋友,美国,带,回来,礼物	1
.....	.....	...

[0143] 表4.本发明与传统的方法比较结果图

[0144]

类别	IG			CN			IG-CN		
	精准率（%）	召回率（%）	F1(%)	精准率（%）	召回率（%）	F1(%)	精准率（%）	召回率（%）	F1(%)
0	0.5103	0.8156	0.8980	0.6164	0.8491	0.7142	0.7226	0.8944	0.9443
1	0.8765	0.6259	0.7699	0.8748	0.7139	0.7862	0.9256	0.7928	0.8844
平均值	0.6934	0.72075	0.83395	0.7456	0.7815	0.7502	0.8241	0.8436	0.91435

[0145] 参见图3和上述表格实验结果可以看出，本方法与传统的方法相比，取得了很明显的优势，并且每一步的实施都是必不可少的。

[0146] 以上实施例仅用于说明本发明的设计思想和特点，其目的在于使本领域内的技术人员能够了解本发明的内容并据以实施，本发明的保护范围不限于上述实施例。所以，凡依据本发明所揭示的原理、设计思路所作的等同变化或修饰，均在本发明的保护范围之内。

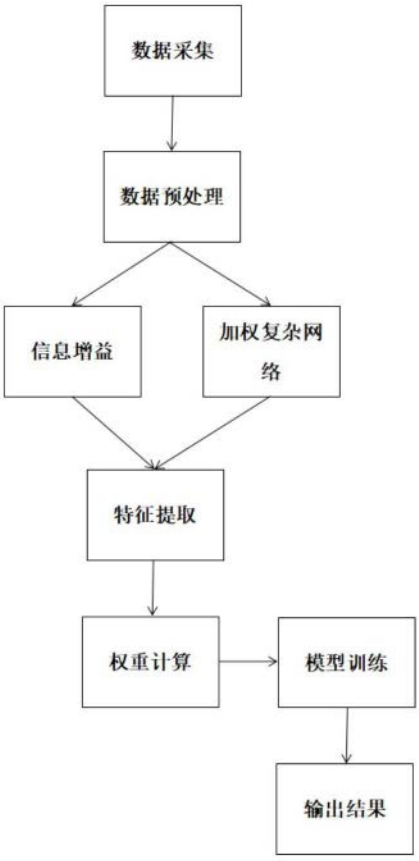


图1

neg_1.txt	2020/8/30 17:46	文本文档	560 KB
neg1_div.txt	2020/8/30 17:53	文本文档	447 KB
pos_1.txt	2020/8/30 17:46	文本文档	555 KB
pos1_div.txt	2020/8/30 17:54	文本文档	454 KB
train_all.txt	2020/8/30 17:48	文本文档	0 KB
train_div.txt	2020/8/30 17:56	文本文档	901 KB

图2

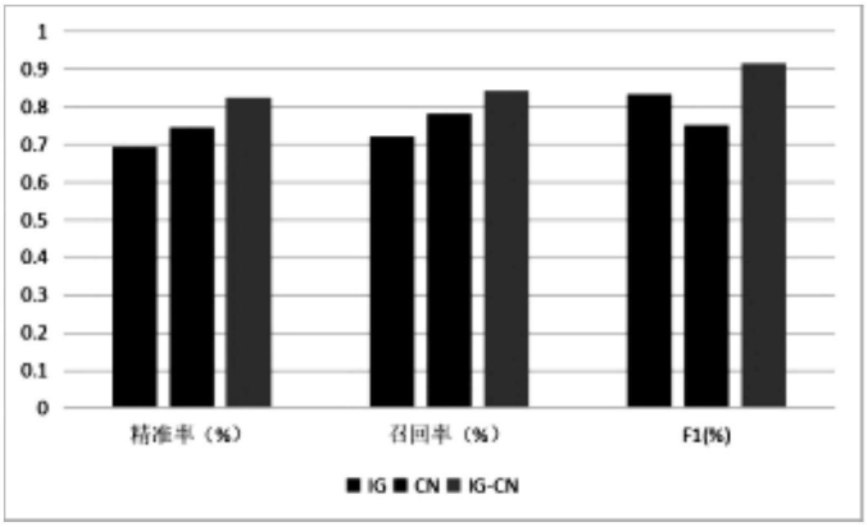


图3