第 42 卷　第 2 期　　计　算　机　工　程　　2016 年 2 月
Vol. 42　No. 2　　Computer Engineering　　February 2016

·人工智能及识别技术·　　文章编号：1000-3428(2016)02-0184-06　　文献标识码：A　　中图分类号：TP18

基于类序列规则的中文微博情感分类

郑　诚[1,2]，沈　磊[1,2]，代　宁[1,2]

（1. 安徽大学计算机科学与技术学院，合肥 230601；
2. 计算智能和信号处理教育部重点实验室，合肥 230601）

摘　要：研究中文微博文本的情感分类问题，介绍一种基于类序列规则的微博情感分类方法。通过情感词典和机器学习的方法获得微博文本中每个句子的 2 个潜在的情感标签，将每条微博文本看作是一个数据序列，从数据集中挖掘出类序列规则，从挖掘出的规则中提取出的有效特征并结合文本其他特征来训练分类器。在 COAE 会议提供的微博数据集上的实验结果表明该方法的有效性。

关键词：情感分类；微博文本；类序列规则；情感词典；机器学习；文本特征

中文引用格式：郑　诚，沈　磊，代　宁. 基于类序列规则的中文微博情感分类[J]. 计算机工程，2016，42(2)：184-189，194.

英文引用格式：Zheng Cheng，Shen Lei，Dai Ning. Chinese Microblog Emotion Classification Based on Class Sequential Rules[J]. Computer Engineering，2016，42(2)：184-189，194.

Chinese Microblog Emotion Classification Based on Class Sequential Rules

ZHENG Cheng[1,2]，SHEN Lei[1,2]，DAI Ning[1,2]

（1. School of Computer Science and Technology，Anhui University，Hefei 230601，China；
2. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education，Hefei 230601，China）

【Abstract】This paper studies the problem of emotion classification in Chinese microblog texts. It introduces a novel approach based on class sequential rules for emotion classification of microblog texts. This approach obtains two potential emotion labels for each sentence in a microblog text by using an emotion lexicon and a machine learning approach respectively，and regards each microblog text as a data sequence. It mines class sequential rules from the dataset. It derives new effective features from the mined rules for emotion classification of microblog texts and other text features to train classifier. Experimental results on a COAE dataset show its validity compared with the traditional methods.

【Key words】emotion classification；microblog text；class sequential rule；emotion lexicon；machine learning；text feature

DOI：10.3969/j.issn.1000-3428.2016.02.033

1　概述

情感分析和观点挖掘是自然语言处理领域中最活跃的研究内容之一[1]。近年来，已有大量的研究集中在对社会媒体的情感分析和观点挖掘上。由于社会媒体网站所产生文本的固有特点，如长度限制和非正式的表达，对其进行情感分析也是一项极具挑战的任务[2]。目前的研究主要集中在基于词典和基于机器学习的方法[3]。基于词典方法的结果依赖于情感词典的质量，而机器学习的方法如 SVM 和朴素贝叶斯，其结果依赖于提取的特征，应用最广泛的是基于 n 元文法和基于词典的特征。

为了更好地利用文本的顺序和话语结构信息来进行微博文本的情感分类，本文提出利用类序列规则为有监督的情感分类提取新的有效的特征。

2　相关工作

早在 2000 年，情感分析已经成为自然语言处理领域中最活跃的研究领域之一。之前的研究主要集中在评论、论坛讨论和博客。

情感分类的目的是将一篇文档分成 3 个极性类，即“积极的”、“消极的”或“中性的”。文献[4]首先应用机器学习技术来决定一个评论是积极的还是消极的。除了仅仅将一个文本简单地分为积极的和消极的，一些研究已经可以识别文本的情绪了，如生气、高兴等；文献[5]在博客文本中，利用 SVM 训

基金项目：安徽省高校自然科学基金资助重点项目(KJ2013A020)；安徽省自然科学基金资助项目(11040606M133)。
作者简介：郑　诚(1964－)，男，副教授，博士，主研方向为智能语义信息检索、数据挖掘；沈　磊、代　宁，硕士研究生。
收稿日期：2015-01-27　　**修回日期：**2015-02-27　**E-mail：**943846578@qq.com

练出一个 132 种心情的情感分类器;文献[6]用 SVM 和 CRF 基于机器学习的技术研究了 Web 博客语料库的情感分类。

随着社会媒体的发展,已有大量的研究着手于 Twitter 和中文微博的情感分析问题。文献[7]利用 Twitter 的哈希标签和笑容符作为情感标签来获得数据;Barborsa 和文献[8]利用带噪音标签的 3 种资源作为训练数据,然后用 SVM 训练出一个分类器;文献[9]利用情绪词和表情图片 2 种情绪知识对大规模微博非标注语料进行筛选并自动标注,用自动标注好的语料作为训练集构建微博情感文本分类器,对微博文本进行情感极性自动分类;文献[10]用基于转发、评论和共享关系特征的输入的中文微博集合构建一个图,来决定每条微博文本的情感;文献[11]用一元文法和二元文法的混合模型来训练分类器,对 Tweet 进行情感分析;文献[12]基于人工标注的数据训练出一个语言模型,然后在一个不同的角度用噪音情感符数据做平滑。

本文研究中文微博文本的情感分类问题,提出利用机器学习方法中的类序列规则在文档级上进行情感分类。

3　微博文本实例

在本文关注于中文微博(如新浪微博)的情感分类上,意在将每条中文微博进行多类别分类("生气"、"困扰"、"恐惧"、"高兴"、"喜欢"、"悲伤"、"惊讶"和"无情感")。本文任务的重点是文档级的情感分析,这比句子级更具挑战性。已有的基于词典和基于机器学习的方法总是将一条微博文本看作是一个词袋或一个句袋,这就并没有考虑到一条微博中文本的顺序和话语结构,导致不能得到很满意的情感分析结果。一条中文微博一般包含若干个句子,最多可能有 140 个汉字,且一条微博文本的情感分类一般是由文本中句子的情感分类序列和句子间的话语关系来决定。表 1 给出一个微博文本的例子,包含 3 个句子和它们的情感标签。句子 1 的情感标签是无情感,句子 2 的情感标签是悲伤,句子 3 的情感标签是高兴,最后 2 个句子由一个转折连词"但是"相连接。基于情感序列和连词序列,可以将整篇微博文本的情感标签看作为"高兴"。

表 1　包含 3 个句子的微博文本实例

句子	情感
1. 今天考试成绩出来了。	无情感
2. 分数考的不高[流泪]!	悲伤
3. 但是进步很大[嘻嘻]。	高兴

本文的方法至少有 2 个优点:

(1)本文的方法考虑到了一条微博文本中句子的顺序和句子间的话语关系。

(2)尽管方法依赖于句子级的情感分类结果,但是由单个方法引起的错误影响较小,因为本文用2 种不同的方法来获得每个句子的情感标签,并且充分利用到这 2 个情感标签。

4　基本方法

本节将介绍基本的基于词典和基于学习的方法来做文档级和句子级的情感分类。一方面,这 2 种方法将作为整个微博文本情感分类的基准。另一方面,本文将用这 2 种方法来获得一条微博文本中每个句子的情感标签,然后利用这个句子级的情感标签。

4.1　基于词典的方法

基于词典的方法依赖于情感词典的质量。在本文的实验中,从 4 个方面的资源中构建了一个中文情感词典:(1)在本文研究中,采用了大连理工大学的情感领域本体,包括 7 个情感类别。实验中去除掉一些对语料库不适合的情感词。(2)将 COAE2014 提供的发现新词任务的答案加入构建的情感词典中。(3)收集到一些对情感分类有用的常见的俚语。(4)从新浪微博中搜集到一些表情符号来扩充情感词典。表 2 显示了本文构建词典中每个情感类别词的总数。

表 2　实验中每个情感类别词的数量

情感类别	数量
生气	591
困扰	9 621
恐惧	1 278
高兴	2 976
喜欢	10 237
悲伤	2 337
惊讶	218

本文使用汉语分词系统 NLPIR 来进行对微博的分词,基于已经构建好的情感词典,统计每个情感类别的词在一个文本中出现的数量,该文本的情感标签仅仅取决于文本中出现情感类别词最多的那一类。如果一个文本不包含情感词,将被标注为"无情感"。

上述过程同样可以应用在一个句子上来得到一个句子级的情感标签。

4.2　基于 SVM 的方法

在之前的研究中,由于 SVM 在情感分类方面显示出了它的优越性,本文采用 SVM 作为基于学习方法的学习模型。本文主要采用台湾大学的 Libsvm 工具包来进行多情感分类。对于文档级和句子级的情感分类,在实验中使用以下基于文本的特征。

(1)情感词特征。将文本或句子中出现某个情感类别的词数量作为特征。比如,在一个文本中仅

有 3 个“高兴”类别的词和 1 个“困扰”类别的词出现,其相关的特征值就是:0(生气),1(困扰),0(恐惧),3(高兴),0(喜欢),0(悲伤),0(惊讶)。

(2)标点符号特征。一些标点符号序列可以反映特定的情感类别,本文收集了一些像这样的标点符号序列作为特征。比如,“???”可能反映的是“生气”情感;“!!!”可能反映的是“惊讶”情感。

(3)句子构成特征。加入了微博消息的句子构成特征,包括首句的情感极性、尾句的情感极性、各类情感(7 类)的句子数。

5 本文方法

本文提出方法开始是利用句子级的类序列规则来做情感或情绪的分类任务。(1)分别用基于词典和基于 SVM 的方法来获得一条微博文本中每个句子的 2 个情感标签。(2)将微博文本转化为包含句子级情感标签和连词的序列。(3)从这些序列中挖掘出类序列规则。(4)从类序列规则中提取出特征,然后用它们对整个文本采用基于 SVM 的情感分类。在此方法中,类序列规则表现了句子顺序和句子间的话语关系的信息,而这些信息对情感分类是一种隐藏的、有用的模式。

本文首先描述类序列规则和类序列规则挖掘算法,然后介绍本文方法。

5.1 类序列规则挖掘

由文献[2]中的描述,可令 $I=\{i_1,i_2,\cdots,i_n\}$ 是一组项,一个序列是项集的一个有序表,一个项集是非空的一组项。假设一个句子由 $<a_1,a_2,\cdots,a_i,\cdots,a_r>$ 来表示,a_i 是一个项集,也被称作为 s 的一个元素,且 a_i 由 $\{x_1,x_2,\cdots,x_j,\cdots,x_k\}$ 表示,$x_j\in I$ 是一个项,一个项在一个句子的一个元素中只能出现一次,但是在不同元素中可以出现多次。一个序列 $s_1=a_1a_2\cdots a_r$ 是另一个序列 $s_2=b_1b_2\cdots b_m$ 的子序列,如果存在整数 $1\leqslant j_1<j_2<\cdots<j_{r-1}<j_r\leqslant m$,这样 $a_1\in b_{j1},a_2\in b_{j2},\cdots,a_r\in b_{jr}$,也可以说 s_2 包含 s_1。

现有例子有 $I=\{1,2,3,4,5,6,7\}$,序列 $<\{2\}\{4,5\}>$ 包含于 $<\{6\}\{2,7\}\{4,5,6\}>$,因为 $\{2\}\subseteq\{2,7\},\{4,5\}\subseteq\{4,5,6\}$;但是 $<\{3,8\}>$ 并不包含于 $<\{3\}\{8\}>$,反之亦然。

用于挖掘的输入的序列数据 D 是成对出现的,$D=\{(s_1,y_1),(s_2,y_2),\cdots,(s_n,y_n)\}$,$s_i$ 是一个序列且 $y_i\in Y$ 是一个类标签,Y 是所有类的一个集合。在本文中,$Y=$ {生气,困扰,恐惧,高兴,喜欢,悲伤,惊讶,无情感}。一个类序列规则是这样的一个隐含式:

$X\rightarrow y$,X 是一个序列,且 $y\in Y$

定义的支持度与置信度如下:

定义 1 如果 X 是 s_i 的一个子序列且 $y_i=y$,(s_i,y_i)就满足这个类序列规则,规则的支持度是所有 D 中实例满足这个规则的部分。

定义 2 如果 X 是 s_i 的一个子序列,一个数据实例(s_i,y_i)要覆盖类序列规则,规则的置信度是既要覆盖这个规则又要满足这个规则的实例所占的比例。

表 3 给出一个 5 个序列和 2 个类别 c_1 和 c_2 的序列数据库实例,用 20% 的最小支持度和 40% 的最小置信度,发现的一个类序列规则如下:

$<\{1\},\{3\},\{7,8\}>\rightarrow c_1$[支持度 =2/5,置信度 =2/3]

数据序列 1 和数据序列 2 满足这个规则,数据序列 1,2,5 覆盖了这个规则。

表 3 一个挖掘类规则序列数据库实例

序列	序列规则	类别
1	$<\{1\},\{3\},\{5\},\{7,8,9\}>$	c_1
2	$<\{1\},\{3\},\{6\},\{7,8\}>$	c_1
3	$<\{1,6\},\{9\}>$	c_2
4	$<\{3,5\},\{6\}>$	c_2
5	$<\{1\},\{3\},\{4\},\{7,8\}>$	c_2

给定一个已标注的序列数据集 D、最小支持度和最小置信度的阈值,利用类序列规则挖掘算法就可以挖掘出满足最小支持度和最小置信度条件的规则。由于篇幅的关系,算法的具体细节在此就省略了。

5.2 微博文本中挖掘的类序列规则

将每篇微博文本转化为一个序列,包括训练数据集和测试数据集。本文整个微博项 I 包含了所有的情感标签和连词,微博中的每个句子表示为 1 个项集(包含 1 个或 2 个情感标签),每个连词也表示为一个项集。这样一条微博就被表示成为一个项集的序列。

上文介绍了如何用基于词典的方法和基于学习的方法来获取每个句子的 2 个情感标签。每个句子由 2 个情感标签的优势在于,可以用 2 个情感标签进行序列模式特征匹配提高性能。

比如表 1 中的例子,这条微博包含了 3 个句子。首先,用基于词典的方法来分别获得句子的情感如“无情感-悲伤-高兴”;然后用基于 SVM 的方法来获得情感如“悲伤-悲伤-高兴”,注意第 3 句的开始有个转折连词“但是”;最后,将这条微博表示成如下序列:

<{无情感,悲伤},{悲伤},{但是},{高兴}>

这里要在序列中加入连词的原因是连词总是能反映出句子间的话语之间的联系(比如承接关系、转折关系、因果关系等),而且句子间的话语之间的联系对整篇微博的情感会有非常大的影响。本文总是在句子的开头找到连词,一个连词能表示句子前后

之间的联系;这样将连词加入到序列中就显得很有用。本文共收集并使用 46 个连词,如表 4 所示。

表 4　连词表

关系	连词
并列关系	和,跟,与,既,同,及
承接关系	则,乃,就,便,于是,然后,说到,此外,像,如,一般,比方,接着
转折关系	却,虽然,但是,尽管,然而,而,偏偏,只是,不过,至于,致,不料,岂知
因果关系	原来,因为,由于,以便,因此,所以,是故,以致
递进关系	不但,不仅,而且,何况,并,且

构建微博文本中的序列数据库的具体步骤如下:

(1)对于测试集微博中的每个句子,同时利用基于词典和基于 SVM 的方法来判别句子的情感。如果 2 种方法得到的情感相同,句子就得到一个情感标签;相反,句子将得到 2 个情感标签。对于训练集微博中的每个句子,直接使用人工标注的句子的情感标签。这里应注意在训练集中,手工标注每个句子可能有一个最初的情感和第二情感,但是本文使用最初的情感作为句子的情感标签。

(2)结合每个句子的情感标签和句首的连词来将一篇微博文本转化为一个序列。

(3)在训练集中,每条微博的情感标签联系一个相关序列作为一个类。

比如,如果知道表 1 中整篇微博的情感标签是“高兴”,就可以得到如下的数据实例:

(<{无情感,悲伤}{悲伤}{但是}{高兴},高兴>)

基于这个从训练集中构建的数据实例,就可以用类序列规则挖掘算法来挖掘出满足最小支持度和最小置信度的类序列规则,这些规则表示了不同情感类别的特定模式。

这里应该注意有些连词和情感出现的比较频繁,还有一些出现的比较罕见,这样仅仅通过一个最小支持度来控制类序列规则的生成过程是不够的。因为为了挖掘出包含罕见的连词和情感的模式,就需要使最小支持度的值非常小,这会使得频繁出现的连词和情感生成大量的多余的模式而导致过拟合,所以本文采用多最小支持度的策略[1]。在这个策略中,一个规则的最小支持度取决于规则中的训练集项的最小频率和参数 τ 的乘积。这样最小支持度会随着数据中项的实际频率而改变,频繁项规则的最小支持度会很高,而罕见项规则的最小支持度会很小。

5.3　微博文本的情感分类

从训练集中挖掘出一些类序列规则之后,用在每个规则 $X \rightarrow y$ 中的序列模式 X 作为一个特征。如果一条微博对应的模式(序列)包含 X,其相应的文本特征值置为 1,反之置为 0。此外,还用到情感词特征、标点符号特征和句子结构特征。最后用台湾大学的 Libsvm 工具包来进行模型训练和测试。整体流程如下:

输入　微博文本集 D

输出　每条微博 a_i 分类结果

(1) $D = \{a_1, a_2, \cdots, a_n\}$。

(2)获取每条微博 a_i 中情感词特征 $SentiNumi = \{num_1, num_2, \cdots, num_7\}$。

(3)统计每条微博中出现的特殊标点符号特征。

(4)依次获取首句的情感极性(FirstSenPlo)、尾句的情感极性(LastSenPlo)、7 种类别情感句子数($z_1, z_2, z_3, z_4, z_5, z_6, z_7$)。

(5)利用挖掘算法,获取类规则序列特征 $X \rightarrow y$,X 是一个序列,且 $y \in Y$。

(6)根据上述提取的特征,训练 LibSVM 分类器,得到每条微博的情感类别 c。

6　实验结果与分析

6.1　数据集

本文采用的数据集是 COAE(中文倾向性分析评测)2014 会议上提供的微博数据,COAE 由中国中文信息学会信息检索专业委员会从 2008 年开始举办。每届评测国内外大约有 20 多家科研单位参加。评测任务主要包含面向新闻的情感关键句抽取与判定、微博倾向性分析和微博观点句要素抽取等。采用了其中的部分数据,训练数据包含 4 000 条微博和 14 526 个句子,每个文本被标注有一个最初的情感标签,每个句子标注有一个最初的情感标签和可能的第 2 个情感标签。测试集包含 5 000 条微博和 16 785 个句子,每条微博最初的情感类型均已被标注,所有的这些微博文本均来自新浪微博。表 5 展示了文档级每个情感类别的数量分布。

表 5　训练和测试数据集的每个情感类别的数量分布

情感类别	训练数据集	测试数据集
生气	235	218
困扰	425	467
恐惧	49	51
高兴	371	558
喜欢	597	779
悲伤	388	372
惊讶	112	168
无情感	1 832	2 437
总计	4 000	5 000

分词工具是中科院的 NLPIR 汉语分词系统,分类工具为台湾大学的 LIBSVM 工具包。本文的实验

环境为 HP Elite 7100 PC, Windows 7, Intel Core i5 650 @ 3.20 GHz,4 GB 内存。

6.2 评价指标

准确率、召回率和 F 值一般都用作为评价指标,当处理多类别问题时,采用宏平均和微平均作为评价标准。宏平均和微平均的准确率、召回率和 F 值计算公式如下:

宏准确率

$$= \frac{1}{7}\sum_i \cdot \frac{\text{数据集中正确分类的微博数量(情感 = } i\text{)}}{\text{数据集中已分类的微博数量(情感 = } i\text{)}}$$

宏召回率

$$= \frac{1}{7}\sum_i \cdot \frac{\text{数据集中正确分类的微博数量(情感 = } i\text{)}}{\text{人工标注的微博数量(情感 = } i\text{)}}$$

$$\text{宏 } F \text{ 值} = \frac{2 \times \text{宏准确率} \times \text{宏召回率}}{\text{宏准确率} + \text{宏召回率}}$$

微准确率

$$= \frac{\sum_i \text{数据集中正确分类的微博数量(情感 = } i\text{)}}{\sum_i \text{数据集中已分类的微博数量(情感 = } i\text{)}}$$

微召回率

$$= \frac{\sum_i \text{数据集中正确分类的微博数量(情感 = } i\text{)}}{\sum_i \text{人工标注的微博数量(情感 = } i\text{)}}$$

$$\text{微 F 值} = \frac{2 \times \text{微准确率} \times \text{微召回率}}{\text{微准确率} + \text{微召回率}}$$

其中,i 表示 7 种情感类别中的一种,包含生气、困扰、恐惧、高兴、喜欢、悲伤和惊讶,这里没有考虑无情感的这种情况来作对比。

6.3 对比方法

本文的方法(仅有类序列规则):采用基于 SVM 的方法来进行微博的分类,特征只用到提取的类序列规则。

本文的方法(类序列规则 + 情感词特征):采用基于 SVM 的方法来进行微博的分类,特征用到提取的类序列规则和情感词特征。

本文的方法(类序列规则 + 情感词特征 + 标点符号特征):采用基于 SVM 的方法来进行微博的分类,特征用到提取的类序列规则、情感词特征和标点符号特征。

本文的方法(类序列规则 + 情感词特征 + 标点符号特征 + 句子结构特征):采用基于 SVM 的方法来进行微博的分类,特征用到提取的类序列规则、情感词特征、标点符号特征和句子结构特征。

基于表情符号的方法:对微博文本中找出所有类别的情感符号,数量最多的那类表情符即为整个微博的情感。

基于词典与规则结合的方法:随机抽取若干个情感词组合识别并分析,得到几种以情感词为中心的、与情感表达有关的组合模式,每个模式按特定公式计算情感。

基于 SVM 传统的方法(传统的极性特征):采用机器学习的方法,训练分类器,特征即链接、情感词典、情感短语和表情符。

在本文实验中,参数的值是根据在训练数据上的五折交叉验证过程来设定的。参数最小置信度为 0.01,$\tau = 0.05$,SVM 训练选择多项式核函数。

6.4 对比结果

表 6 展示了本文方法与其他方法对比情况。

表 6 实验对比结果

方法	宏平均			微平均		
	准确率	召回率	F 值	准确率	召回率	F 值
表情符号	0.213 6	0.248 0	0.229 5	0.135 6	0.274 9	0.181 6
情感词典与规则结合	0.303 6	0.341 5	0.321 4	0.216 9	0.257 8	0.235 6
传统 SVM 方法	0.367 1	0.328 9	0.346 9	0.425 2	0.408 5	0.416 7
本文的方法(仅类序列规则)	0.406 4	0.426 7	0.416 3	0.392 4	0.488 2	0.435 1
本文的方法(类序列规则 + 情感词特征)	0.408 5	0.431 6	0.419 7	0.393 3	0.497 2	0.439 2
本文的方法(类序列规则 + 情感词特征 + 标点符号特征)	0.412 1	0.445 9	0.428 3	0.395 6	0.489 6	0.437 6
本文的方法(类序列规则 + 情感词特征 + 标点符号特征 + 句子结构特征)	0.419 4	0.493 6	0.491 9	0.472 8	0.455 5	0.464 0

本文方法在宏 F 值和微 F 值上都超过了其他方法,而且在同时有类序列规则、情感词特征、标点符号特征和句子结构特征作为特征训练时取得的效果最好,结果显示出了基于类序列规则特征的有效性。比较使用其他特征和仅仅使用类序列规则特征,前者有一个更好的增长效果,表明了文本中其他相关的特征的有效性,且句子结构特征对分类结果影响最大。本文方法比传统的基于 SVM 的方法效果好,是因为传统的方法中特征未考虑到文本的顺序和句子间的话语联系。基于词典与规则相结合的方法效果不理想的原因是其过度依赖于构建的词典。

图 1 展示了各个方法在 7 种情感类别上的F 值,可以看到本文方法在“生气”、“困扰”、“恐惧”、“喜欢”和“悲伤”上要明显优于其他方法。

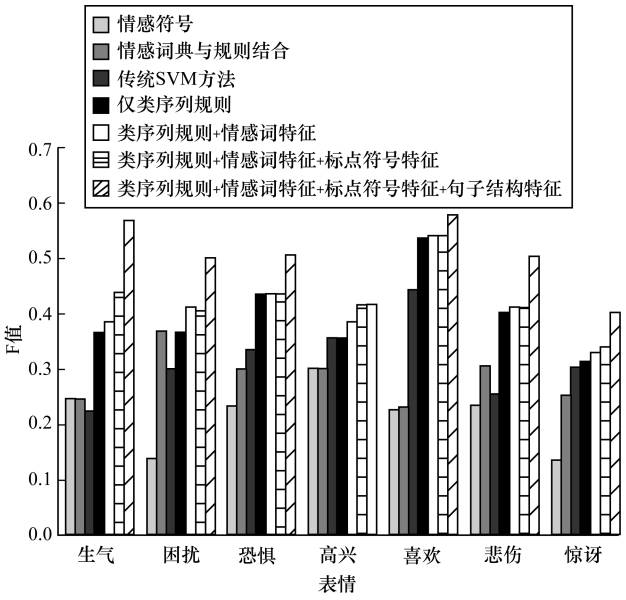

图 1　不同方法在 7 种情感类别上的 F 值

6.5　参数研究

为研究最小置信度阈值如何影响情感分类的效果,本文用不同的最小置信度进行了实验。图 2 展示了本文方法的性能变化。

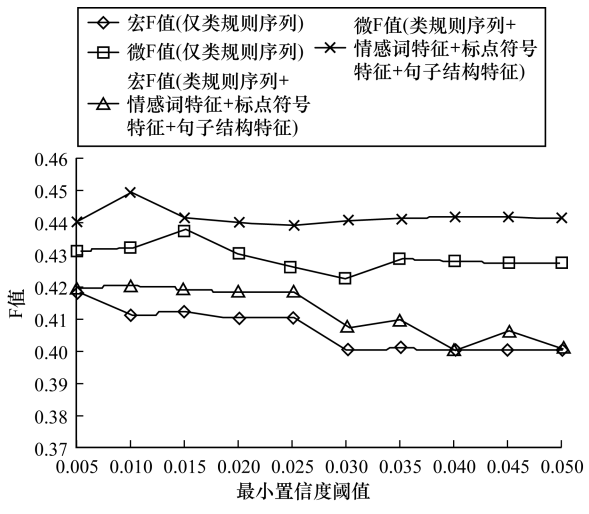

图 2　最小置信度的阈值

在图 2 中,最小置信度从 0.005 ~0.05 之间变化,可以看到微 F 值在很大的范围中都很稳定,而宏 F 值随着最小置信度值越大下降很明显,这是由于挖掘出来的类序列规则越来越少的缘故。总之,置信度的值对这种方法的影响不是特别大。

研究多类最小支持度参数 τ 影响情感分类的结果,本文做了不同值 τ 下的实验。τ 值也是从 0.005 ~0.5 之间变化,可以看到随着 τ 值的变化,宏 F 值和微 F 值变化都很稳定,如图 3 所示。

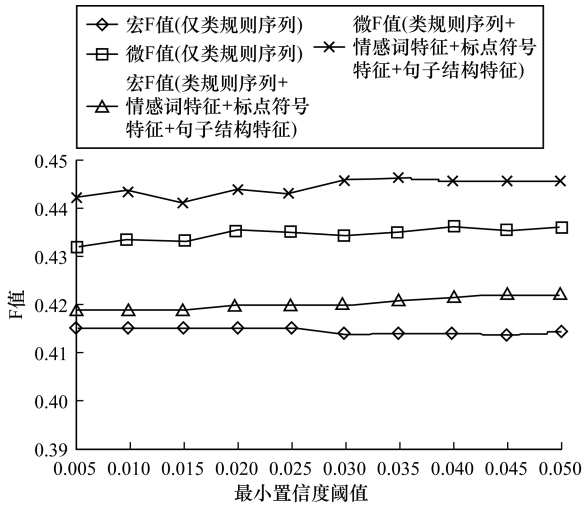

图 3　多类别最小支持度参数

7　结束语

本文研究了微博文本中情感分类的问题,并提出一个基于类序列规则的方法,利用机器学习方法中的类序列规则在文档级上进行情感分类,COAE 数据集上的实验结果表明本文方法具有有效性和鲁棒性。在下一步的工作中,将考虑到一些句子的依存关系和规则相结合,更充分地利用语篇信息和社交网络信息提高分类的效果。

参考文献

[1] Liu B. SentimentAnalysis and Opinion Mining:Synthesis Lectures on Human Language Technologies [M]. [S.1.]:Morgan & Claypool Publishers,2012.

[2] 赵妍妍,秦　兵,刘　挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.

[3] 谢丽星,周　明,孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报,2012,26(1):73-83.

[4] Pang B,Lee L,Vaithyanathan S. Thumbs up?:Sentiment Classification Using Machine Learning Techniques[C]//Proceedings of ACL Conference on Empirical Methods in Natural Language Processin. [S.1.]:Association for Computational Linguistics,2002:79-86.

[5] Mishne G. Experiments with Mood Classification in Blog Posts [C]//Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access. New York,USA:ACM Press,2005:19.

[6] Yang C,Lin K H,Chen H H. Emotion Classification Using Web Blog Corpora[C]//Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence. Washington D.C.,USA:IEEE Press,2007:275-278.

(下转第 194 页)

5.3 检测正确率对比实验

选择 3 组 KDDCUP 数据集中 1 000 条数据记录,实验环境和 5.2 节相同,BP 神经网络、遗传算法优化的 BP 神经网络(BP-GA)、人工蜂群算法优化的 BP 神经网络(BP-ABC)检测正确率对比如表 4 所示。

表 4 3 组数据检测正确率对比 %

网络模型	第 1 组数据	第 2 组数据	第 3 组数据
BP	70.5	69.4	77.3
BP-GA	94.1	84.0	95.1
BP-ABC	98.0	97.6	97.4

从表 4 中可以看出,人工蜂群优化的 BP 神经网络在入侵检测中具有更高的正确率。因为 BP 神经网络在满足一定误差要求时就会停止,有可能这个误差不是全局最小,而人工蜂群优化的 BP 神经网络是在最大迭代次数内找到更低的误差去代替有可能已经满足要求的误差,所以后者会有更高的正确率。

6 结束语

本文用人工蜂群算法优化 BP 神经网络中的权值和阈值,避免 BP 神经网络陷入局部最优,加快了收敛速度。这种集成方法充分利用神经网络学习能力强的优势,同时也利用了人工蜂群算法全局寻优的特点。通过在入侵检测中的应用表明了该优化算法的有效性。今后的工作中会使用更多的数据集进行评估,找出该优化算法在不同数据集中的应用优势,并研究此方法在入侵检测分类中的应用。

参考文献

[1] 王 玲. 基于 BP 算法的人工神经网络建模研究[J]. 装备制造技术,2014,(1):162-164.

[2] 魏 海,杨华舒,苏志敏,等. 前馈神经网络导数特性分析[J]. 计算机工程,2014,40(7):198-201.

[3] 聂 铭,周冀衡,杨荣生,等. 基于粒子群优化 BP 神经网络的烤烟钾氯比预测模型[J]. 烟草科技,2014,(6):49-53.

[4] Lim S, Zhu J. Integrated Data Envelopment Analysis: Global vs. Local Optimum[J]. European Journal of Operational Research,2013,229(1):276-278.

[5] 章 纯,刘 锋,廖国维,等. 基于 WCA 优化算法的空间桁架结构优化设计[J]. 建筑钢结构进展,2014,(1):34-41.

[6] 上官海洋,向铁元,张 巍,等. 基于智能优化算法的 FACTS 设备多目标优化配置[J]. 电网技术,2014,38(8):2193-2199.

[7] Hua Haiyan, Lin Shuwen. New Knowledge-based Genetic Algorithm for Excavator Boom Structural Optimization[J]. Chinese Journal of Mechanical Engineering,2014,27(2):392-401.

[8] 刘三阳,张 平,朱明敏. 基于局部搜索的人工蜂群算法[J]. 控制与决策,2014,(1):124-128.

[9] 张永俊,牟 琦,毕孝儒. 基于云模型的增量 SVM 入侵检测方法[J]. 计算机应用与软件,2013,30(3):311-314.

[10] 夏淑华. 基于 BP 神经网络的入侵检测系统设计[J] 计算机与现代化,2011,(4):47-49.

[11] Karaboga D, Ozturk C. A Novel Clustering Approach Artificial Bee Colony(ABC) Algorithm[J]. Applied Soft Computing,2011,11(1):652-657.

[12] Karaboga D, Basturk B. A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony(ABC) Algorithm[J]. Journal of Global Optimization,2007,39(3):459-471.

[13] KDD Cup 1999 Data. The UCI KDD[EB/OL]. (1999-10-28). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

编辑 顾逸斐

(上接第 189 页)

[7] Davidov D, Tsur O, Rappoport A. EnhancedSentiment Learning Using Twitter Hashtags and Smileys[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Washington D. C., USA: IEEE Press,2010:241-249.

[8] Barbosa L, Feng J. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Washington D. C., USA: IEEE Press,2010: 36-44.

[9] 庞 磊,李寿山,周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程,2012,38(13):156-158,162.

[10] Liu Q, Feng C, Huang H. Emotional Tendency Identification for Micro-blog Topics Based on Multiple Characteristics[C]//Proceedings of the 26th PACLIC'12. Washington D. C., USA: IEEE Press,2012:269-278.

[11] Go A, Huang L, Bhayani R. Twitter Sentiment Analysis[C]//Proceedings of ACL'09. Washington D. C., USA: IEEE Press,2009:235-243.

[12] Liu K L, Li W J, Guo M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis[C]// Proceedings of AAAI'12. Washington D. C., USA: IEEE Press,2012:125-134.

编辑 索书志