基于模糊集和支持向量机的文本流派分类方法

朱艳辉[1],阳爱民[2],杨伟丰[1]
ZHU Yan-hui[1],YANG Ai-min[2],YANG Wei-feng[1]

1.湖南工业大学 计算机与通信学院,湖南 株洲 412008
2.国防科学技术大学 计算机学院,长沙 410073
1.Institute of Computer & Communication,Hunan University of Technology,Zhuzhou,Hunan 412008,China
2.Institute of Computer,National University of Defense Technology,Changsha 410073,China
E-mail:swayhzhu@163.com

ZHU Yan-hui,YANG Ai-min,YANG Wei-feng.Text genre classification method based on fuzzy set and Support Vector Machine.Computer Engineering and Applications,2008,44(11):145-147.

Abstract: In terms of poor performance of current genre classifications,the paper proposes a text genre classification method based on fuzzy set and Support Vector Machine(SVM),which combines advantages of both SVM and fuzzy set theory.Experiments give comparative classifying effect under different text feature generation methods and different feature number using movie reviews as data.The author also compare our method with SVM.The comparative result indicates that its micro-average-precision is better than that of SVM.Our method is proved from theory and experiment that it gains a better classifying performance.
Key words: fuzzy theory;Support Vector Machine(SVM);text genre classification

摘 要:针对目前流派分类技术分类性能不够好的问题,将支持向量机和模糊集理论的优点结合起来,提出了一种基于模糊集和支持向量机的文本流派分类方法。并以电影评论作为数据集,比较和分析了该方法在不同文本特征生成方法、不同特征数目下的分类效果,并与SVM方法进行了比较,实验结果表明其微平均查准率要优于SVM方法。理论和实验都证明了提出的方法可以取得较好的分类性能。
关键词:模糊理论;支持向量机;文本流派分类
文章编号:1002-8331(2008)11-0145-03 **文献标识码:**A **中图分类号:**TP301

1 引言

互联网技术飞速发展,每天产生大量的文本文档,各种智能文本自动处理技术也随即产生,如:文本检索技术可以自动地将某些用户需要的信息检索出来;文本分类技术可以将待分文本自动地归为预先设定的文本类别。然而现有的文本分类与检索技术大多只是针对文本主题进行分类检索,如果能按文本的风格(如正面和方面,健康和不良等)进行分类,则可进一步提高检索性能。

文本的风格又称为文本的流派,文本流派分类就是按照文本的风格进行分类,而普通的文本分类是按照文本的主题进行分类的。对文本内容的分类,近十多年来的研究已经比较深入。但是,对文本流派的分类研究还处于一个比较初级的阶段。近年来,国外已有不少学者在文本流派分类方面进行了研究[1-3],取得了一些有效的方法。然而研究表明,目前所有的流派分类技术,在查全率和查准率等分类性能上比传统的基于主题的分类技术要差,所以如何提高流派分类的分类性能成为了目前该领域的研究热点。本文以电影评论(英文文本)作为数据集,将分类性能较好的支持向量机和模糊集合理论的优点结合起来,提出了一种基于模糊集和支持向量机的文本流派分类方法(Text Genre Classification Method Based on Fuzzy set and Support Vector Machine,TGCMBFSVM)。实验结果表明TGCMBFSVM取得了不错的分类效果。

2 TGCMBFSVM结构设计

TGCMBFSVM结构框图如图1所示。其中分类器是整个TGCMBFSVM的核心部分,这里采用的是基于模糊集和支持向量机的流派分类器,具体工作原理在第3章介绍。

采用较常用的向量空间法表示文本,一篇文档表示为特征空间中的一个向量$(w_1,w_2,\cdots,w_n)^T$,其中w_i为第i个特征项的权值,n表示文本的全部特征总数。

2.1 文本特征的确定

常用的文本特征有词、词性、短语和N-Gram项等。

基金项目:湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.05JJ40101);湖南省教育厅资助科研项目(No.07B014)。
作者简介:朱艳辉(1968-),女,副教授,CCF高级会员,主要研究领域为:智能信息处理、文本分类;阳爱民(1970-),男,研究员,博士后,CCF高级会员,主要研究领域为:智能计算、机器学习。
收稿日期:2007-07-27 **修回日期:**2007-10-22

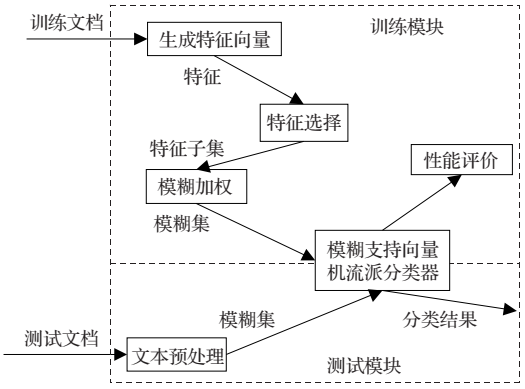

图 1 TGCMBFSVM 结构框图

(1)词语。对于英文来说,不存在分词的问题,因为词语之间已经使用空格隔开。但是,英语存在词形的变化,如:名词的单复数、动词的时态变化等,所以需进行词干抽取。

(2)词性。即词语的语法类别,如:名词、动词、形容词等。形容词常用于表达情感特征,文献[1,2]论述了形容词在文本流派分类中的应用,所以本文将形容词作为一种文本特征。

(3)*N*-Gram 项。对于英文来说,*N*-Gram 项一般由相邻单词构成,例如:从"Republic of China"中提取 2-Gram 项,可得到"Republic of"、"of China"两个 2-Gram 项。使用 *N*-Gram 项进行文本分类,最基本的要求是所选择的 *N*-Gram 项能够覆盖文档中的词,这就涉及如何选择参数 *N* 的问题,一般英文文档常取 1-Gram、2-Gram 和 3-Gram 项。

2.2 权值计算方法

(1)TFIDF 方法

第 i 个文本中特征项 k 的权值由式(1)计算:

$$w_{ik}=\frac{tf_{ik}\times\log(\frac{N}{n_k})}{\sqrt{\sum_k(tf_{ik}\times\log(\frac{N}{n_k}))^2}} \tag{1}$$

其中,tf_{ik} 表示特征 k 在文本 i 中出现的次数;N 是训练集中所含文本的总数;n_k 是出现特征 k 的文本数。

(2)布尔模型

$$w_{ik}=\begin{cases}1 & \text{特征在文档中出现时}\\ 0 & \text{特征在文档中不出现时}\end{cases}$$

2.3 特征选择

特征选择是从原始的属性集中选择一个子集。特征选择的目的是识别属性的重要性,进而删除无关的属性或多余属性,选择尽量少而又能尽量好地区分待分类文本的一个子集。特征选择的优点在于一方面减少学习算法的运算时间,另一方面可以提高结果的准确率。

文本特征选择的常用方法有文档频次(DF)、互信息方法(MI)、信息熵方法(IG)、$\chi2$ 统计量方法(CHI)。根据文献[4],$\chi2$ 统计量方法取得了较好的效果,这里采用其进行特征选择。

2.4 模糊加权

在实际应用中,不同的训练样本对分类结果的影响是不同的。一般来说,训练集中存在某些点对分类结果的影响很大,而另一些点对分类结果的影响小,因此更需要将重要样本正确分类而不在意其他样本(例如噪声)是否被错分。

这样,训练样本不再严格属于两类中的某一类,某一个样本 90%可能属于某一类,10%可能不属于这一类。换言之,对于每个样本 x_i,存在一个与之对应的模糊隶属度 s_i,$0<s_i<1$,s_i 可以看作是 x_i 隶属于某一类的程度,而 $1-s_i$ 则是 x_i 不属于某一类的程度。由于需要为每个样本 x_i 指定一个模糊隶属度 s_i,这时训练集便转化为模糊训练集。应选择一个合适的函数来给每一个输入样本赋予一个模糊隶属度,称这样的函数为隶属函数。通过隶属函数求出模糊隶属度 s_i 得到模糊训练集的过程称为模糊加权。

隶属函数可以是任意形式的曲线,隶属函数的选择是设计 TGCMBFSVM 一个很关键的任务之一。比较常见的模糊集隶属函数有:高斯型、三角形、sigmoid 函数型等。

3 基于模糊集和支持向量机的流派分类器 TGCBFSVM

3.1 SVM

SVM 方法基于结构风险最小化原理,明显优于传统的基于经验风险最小化原理的常规分类方法。SVM 方法本质上是一种两类分类器,非常适合用于具有两类的文本流派分类问题。

SVM 学习问题[5]描述如下:假定大小为 l 的训练样本集 $\{(x_i,y_i),i=1,2,\cdots,l,x_i\in R^n,y_i\in\{+1,-1\}\}$,对于本文介绍的流派分类问题,$y_i$ 由正类和反类两种类别组成。SVM 的主要目的是构造一个分类超平面以分割两类不同的样本,使得分类间隔最大,最优分类超平面问题描述如式(2)所示:

$$\begin{aligned}&\min\frac{1}{2}(\omega\cdot\omega)+C\sum_{i=1}^{l}\xi_i\\&\text{s.t. } y_i(\omega\cdot x_i+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,\cdots,l\end{aligned} \tag{2}$$

采用拉格朗日乘子法求解这个具有线性约束的二次规划问题,得到对偶最优化问题,如式(3):

$$\begin{aligned}&\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i,x_j)\\&\text{s.t. } 0\leqslant\alpha_i\leqslant C,\sum_{i=1}^{l}\alpha_i y_i=0\end{aligned} \tag{3}$$

式中 α_i 为与每个样本对应的 Lagrange 乘子,对应 $\alpha_i>0$ 的样本点 $\boldsymbol{x}_i$ 被称作支持向量。

对于非线性问题,可以通过非线性变换转化为某个高维空间中的线性问题,在变换空间求最优分类面,仍可用线性分类器完成分类。目标函数变为式(4):

$$\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\,k(x_i,x_j) \tag{4}$$

其中 $K(x_i,x_j)$为满足 Mercer 条件的核函数。应用较多的核函数有:多项式核函数、径向基核函数、神经网络核函数。决策函数和参数 b 分别如式(5)和式(6)所示:

$$\begin{aligned}&f(x)=\text{sgn}(\sum_{i=1}^{l}y_i\alpha_i K(x_i,x_j)+b)\\&\text{s.t. } b=\frac{1}{N_{NSV}}\sum_{x_i\in JN}(y_i-\sum_{x_j\in J}\alpha_j y_j K(x_j,x_i))=0\end{aligned} \tag{5}$$

其中 N_{NSV} 为标准支持向量数,JN 为标准支持向量的集合,J 为支持向量的集合。

3.2 TGCBFSVM

在等式(2)中,等式右边的第 2 项实际上是惩罚项,较大的

C 意味着为错误项指定了较大的惩罚值,从而减少了错误分类的数据点;另一方面,较小的 C 值,意味着忽略了一些"微不足道"的错误分类点,因而可以得到较大的分类间隔。无论 C 的值是大还是小,在 SVM 的训练过程中这个参数的值始终是固定的,也就是说,所有训练点都是被同等对待的,这样就导致了 SVM 对某些特殊情形的过分敏感,例如孤立点与噪声,这即是所谓的"过学习"现象。通过对训练样本进行模糊加权后,训练集便转化为模糊训练集。引入模糊参数的训练样本集为:$\{(x_i, y_i, s_i), i=1,2,\cdots,l, x_i \in R^n, y_i \in \{+1,-1\}, \sigma<s_i<1\}$,$\sigma$ 是一个充分小的正数。选择高斯型隶属函数,即:

$$s_{i+}=\exp\left(-\frac{(x_i-x_{c+})^2}{2\delta_+^2}\right)\ x_i \in y_+\quad s_{i-}=\exp\left(-\frac{(x_i-x_{c-})^2}{2\delta_-^2}\right)\ x_i \in y_- \tag{6}$$

其中:

$$x_{c+}=\frac{1}{N_+}\sum_{i=1}^{l} x_i\ \ x_i \in y_+, x_{c-}=\frac{1}{N_-}\sum_{i=1}^{l} x_i\ \ x_i \in y_- \tag{7}$$

式中 N_+,N_- 分别表示正类和负类样本的数目。这样,支持向量机的原始优化问题变为:

$$\min \frac{1}{2}(\omega\cdot\omega)+C_+\sum_{i=1}^{l} s_i\xi_i+C_-\sum_{i=1}^{l} s_i\xi_i$$
$$\text{s.t. } y_i(\omega\cdot x_i+b)\geqslant 1-\xi_i, \xi_i\geqslant 0, i=1,\cdots,l \tag{8}$$

式中 ξ_i 是松弛变量,表示 SVM 解的错误度量,因此 $s_i\xi_i$ 就是衡量对于重要性不同的变量错分程度的度量。其对偶问题:

$$\max \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i,x_j)$$
$$\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i=0\ \ 0\leqslant\alpha_i\leqslant C_+ s_i, 0\leqslant\alpha_i\leqslant C_- s_i, i=1,2,\cdots,l \tag{9}$$

由式(9)可知,与传统 SVM 不同的是,在 TGCBFSVM 中将惩罚项模糊化,以降低不太重要的数据对分类结果的影响。在 TGCBFSVM 中,α_i 的上界是一个动态的模糊隶属度。通过不同的 s_i,可以控制所需要的样本。

4　实验及结果分析

4.1　数据集和文本预处理

从 http://www.imdb.com 网站选择电影评论作为数据集,该网站的每篇评论都按"loved 或 hated"进行了分类,将标为"loved"的评论文章看作正面的,而将"hated"的看成反面的,无需再手工标注。从电影网站下载了 300 篇正面评论文章和 300 篇反面评论文章,均为英文文档,去掉 HTML 标记,整理成 600 个纯文本文件。选择训练集和测试集的方法如下:将这些标注好正面或反面文章的样本集平均分成十份,选择其中一份作为测试集,剩余的九份作为训练集。这样每一份都依次轮流作为测试集,运行分类算法,共执行 10 次分类操作。用 VC++6.0 实现本文算法,在 256 M 内存、Windows XP 的环境下进行实验。分别使用形容词和 N-Gram 两种方法生成文本特征,使用 TFIDF 方法和布尔模型计算特征权值,使用 $\chi 2$ 方法提取文档特征。C 值取 0.95,核函数选择径向基核函数。

从以下几个方面考察 TGCMBFSVM 的性能,使用的性能评价指标为国际上常用的,查准率和查全率的平衡点(Break-even Point)处的微平均查准率(Micro-Averaging-Precision)。

(1)分别使用形容词和 N-Gram 两种方法生成文本特征时分类器的性能;

(2)选不同数量的特征时分类器的性能;

(3)不同权值计算方法对分类其性能的影响;

(4)与 SVM 分类器的性能比较。

4.2　不同权值计算方法下 TGCMBFSVM 的分类性能

在特征数目从 1 000~3 500 的情况下,采用 1-Gram 特征表示方法,分别使用布尔模型和 TFIDF 方法式(1)计算特征权值,对分类器的微平均查准率进行了测试和比较。如图 2 所示。

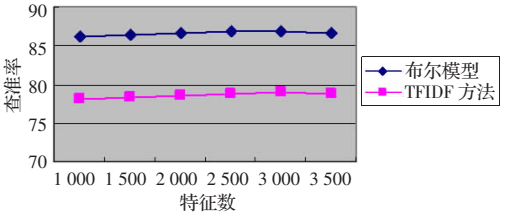

图 2　不同权值计算方法下 TGCMBFSVM 的微平均查准率比较

从图 2 可知,无论特征数目取多少,使用布尔模型作为特征权值计算方法都比 TFIDF 方法的性能要好。这个结果与基于主题的文本分类正好相反,可见,文本流派分类与基于主题的文档分类有着不同的特征。基于此,下面的实验都以布尔模型作为特征权值计算方法。

4.3　不同特征生成方法下 TGCMBFSVM 的分类性能

在特征数目从 1 000~3 500(形容词特征数从 300~1 500)的情况下,分别使用形容词和 N-Gram 两种文本特征生成方法,权值计算采用布尔模型,对 TGCMBFSVM 进行了测试。分类器的微平均查准率如表 1 所示。表 1 中,1-Gram 表示只取 1-Gram 项,2-Gram 同此;1/2-Gram 表示既取 1-Gram 项,又取 2-Gram 项。

表 1　不同特征生成方法下微平均查准率比较表

特征数目	1-Gram	2-Gram	1/2-Gram	adjectives
1 000(300)	86.25	80.15	85.85	80.65
1 500(500)	86.33	80.29	85.90	80.74
2 000(800)	86.65	80.45	85.98	80.78
2 500(1 100)	86.89	80.51	85.09	80.01
3 000(1 300)	86.75	80.72	85.28	80.12
3 500(1 500)	86.73	80.75	85.31	80.11

从表 1 中可以得到如下结论:

(1)使用 1-Gram 特征生成方法的性能要优于 adjectives 特征和其他 N-Gram 方法。

(2)随着特征数目的增加,分类器的微平均查准率逐步提高;当达到一定的数目后,查准率不再升高,反而有所下降,但下降不是很明显。

(3)只取形容词作为文本特征时,分类器的分类性能要比 N-Gram 方法差。

4.4　与 SVM 分类器的性能比较

将 TGCMBFSVM 与文献[3]中介绍的分类性能最好的 SVM 方法进行比较,特征数目选择 3 000,微平均查准率结果如图 3所示。从图 3 可知,TGCMBFSVM 分类方法的性能要优于 SVM方法。

5　结束语

文章针对目前流派分类技术分类性能较差的问题,提出了

(下转 157 页)