基于边界可信度相似的快速文本分类方法

杨林波[1],王士同[1,2]
YANG Lin-bo[1],WANG Shi-tong[1,2]

1.江南大学 信息工程学院,江苏 无锡 214122
2.江南大学 创新多媒体中心,江苏 无锡 214122
1.School of Information,Jiangnan University,Wuxi,Jiangsu 214122,China
2.Creative Multimedia Center,Jiangnan University,Wuxi,Jiangsu 214122,China
E-mail:seekingyang@163.com

YANG Lin-bo,WANG Shi-tong.Fast text categorization approach based on similarities between text boundaries.Computer Engineering and Applications,2009,45(4):156-158.

Abstract: Center and boundaries are important characters of a class in text analysis.Using the center and boundaries as the criterion for text categorization,a fast text categorization approach based on the similarities between boundaries had been presented in this paper.By adjusting the similarity of a text to its class based on the similarity of the boundaries,the disadvantages of the imbalance of the classes and the distribution of the samples can be overcome such that the performance of text categorization may be enhanced.The experimental results demonstrate the advantage of the proposed approach in accuracy and robustness,especially in speed.

Key words: text categorization;similarity;fast categorization

摘 要:类别的中心和边界是类别的重要特征.利用训练样本的中心和边界作为分类准则,提出了一种基于边界可信度相似的快速文本分类算法。通过类别边界可信度调整文本与类别的相似性,克服了数据集类别间样本分布不均衡和类别中样本密度不均的缺点,提高了分类性能。实验结果表明该算法提高了文本分类的效果,显示出了较好的鲁棒性,并显著提高了文本分类效率。

关键词:文本分类;相似度;快速分类
DOI:10.3778/j.issn.1002-8331.2009.04.044 **文章编号**:1002-8331(2009)04-0156-03 **文献标识码**:A **中图分类号**:TP18

1 引言

随着信息技术的飞速发展和互联网技术在全球的普及,当代生活中文本信息越来越彰显出海量特征,如何高效快捷地管理和利用这些信息也越来越被人们关注。文本分类(Text Categorization,TC)是自然语言处理中的一个重要的问题,也是信息检索和文本挖掘的重要基础,其主要任务是根据已知类别的文本信息判定未知类别的文本信息所属的类别,以便于人们管理利用。目前,文本分类已经被应用到信息检索、机器翻译、自动文摘、信息过滤、邮件分类等许多领域中。

传统文本分类方法主要有决策树法(Decision Tree,DT)、支持向量机法(Support Vector Machine,SVM)、线性最小二乘法(Linear Least Squares Fit,LLSF)、K 最近邻法(K-Nearest Neighbor,KNN)、朴素贝叶斯方法(Naive Bayes,NB)、神经网络法(Neural Net,NN)及中心向量法等[1-2]。其中,KNN 和 SVM 分类效果较好[2],但处理大规模的数据集时效率显著降低;中心向量法进行文本分类时效率最高,适用于数据集各类之间差异较明显的分类问题,对于较复杂的情况,其分类效果对数据集的分布较敏感。为了克服以上矛盾,一些传统分类方法的改进算法被提出,它们大致分为两类:一类引入索引技术采用快速搜索算法;另一类删减压缩原始的数据集,减小训练集,提高效率[3]。但这类方法或是降低了分类效果或是只在一定程度上提高了效率。

在分析了中心向量法和 KNN 法的分类特性后, 本文提出了一种新的快速文本分类方法,通过对原始训练样本集的训练生成代表样本,并获得类别间的边界信息,通过代表样本和边界信息能更准确地描述原始类别信息。实验证明这种方法不仅提高了分类效果,而且大大提高了分类的效率。

2 中心向量法和 KNN 法的文本分类方法及特性

2.1 中心向量法文本分类及其特性

中心向量法的基本思想是:通过对训练集进行训练得到每一个已知类别的中心,称之为类中心向量,分类过程中将待分文档与已知的类中心向量进行相似度比较,判定规则为相似度最大的类中心向量所代表的类别为待分文档的类别。常用的文

基金项目:国家教育部科学技术重点研究项目(the Key Technologies Project of the Ministry of Education of China No.105087)。
作者简介:杨林波(1983-),男,硕士研究生,研究方向为人工智能、模式识别;王士同(1964-),男,教授,博士生导师,研究方向为人工智能、模式识别、数据挖掘、神经网络及生物信息学。
收稿日期:2008-01-09 **修回日期**:2008-04-02

本相似性度量有欧氏距离、夹角余弦距离等。中心向量法最初用于信息检索,现已广泛应用于文本分类,其过程描述如下:

令 $C=\{c_i\}_{i=1}^{m}$ 代表训练集所包含的 m 个类。

(1)训练

对每一个类 c_i,计算该类中所有文档向量的算术平均作为该类的类中心向量 $\boldsymbol{v}(c_i)$。

(2)分类

①给定一个待分类文档 d,计算 d 与所有类中心向量 $\boldsymbol{v}(c_i)$ 的相似度 $Sim(d,\boldsymbol{v}(c_i))$;

②返回 $c(d)=\arg\max Sim(d,\boldsymbol{v}(c_i))$。

设整个训练集的文档数为 N,类别数为 m,则训练阶段的时间复杂度为 $O(N)$;分类阶段对每一个待分文档计算 m 个相似度值,时间复杂度为 $O(m)$。

中心向量法的分类特性是:当训练集中各类别间大小相对均衡,且同类别文档分布稠密时,分类效果较好;而训练集中各类别间大小不均衡,且同类别文档分布稀疏时,分类效果较差,即对训练集各类别的大小和分布敏感。如图 1,当 c_i、c_j 两类大小不均衡时 $d_1>d_2$,c_i 类边缘文本易被误分至 c_j 类中。

2.2　KNN 法文本分类及其特性

KNN 法文本分类的基本思想是:考察训练集中与待分文本距离最近(最相似)的 k 篇文本,根据这 k 篇文本的类别来判断待分类文本的类别,通常采用基于个数、相似度累加或者加权平均的投票策略作为判别规则。KNN 是一种基于统计的惰性学习算法,没有训练阶段,简单易行且分类效果较好。

KNN 法文本分类的特性是:一是对 k 值的选择依赖性较大,二是靠近类中心区域的训练样本对分类结果影响小,而每类训练样本在边缘区域分布的稠密对分类结果影响明显[3]。如图 2 所示,文档 d 本属于类别 2,但由于类别 1 中靠近文档 d 的边缘分布密度大,当选择 d 的 k 个近邻作判别时就会误分为类别 1。

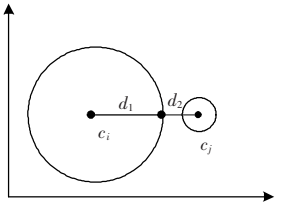

图 1　中心向量法分类示例

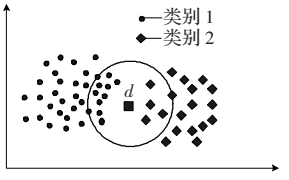

图 2　KNN 法分类示例

针对上述两种方法的分类特性,提出了一种基于边界可信度相似的快速文本分类方法,通过各类别的中心和边界来描述类别的信息与特征,并以此为依据进行待分类文档的所属类别的判断。

3　基于边界可信度相似的快速文本分类方法

3.1　基本概念

为了对训练样本的分布进行描述,引入以下概念。给定训练样本集 $D=\{x_1,\cdots,x_l\}$,其中 $x_i\in R^n$,$i=1,\cdots,l$;样本所属的类别 $C=\{c_1,\cdots,c_m\}$;各类别的中心 $V=\{v(c_1),\cdots,v(c_m)\}$,其中 $v(c_i)\in R^n$,$i=1,\cdots,m$。定义:

定义 1　设 $Sim(x_i,x_j)$ 表示两个文档 x_i 与 x_j 之间的相似度(距离),类别 c_a 与 $c_b(a\neq b)$ 的边界距离定义为:

$Verge(c_a,c_b)=Sim(v(c_a),x_i)$,其中 x_i 满足

$$\max_{x_i\in c_a,x_j\in c_b,a\neq b} Sim(x_i,x_j) \tag{1}$$

定义 2　边界代表类别真实分布的可信度 λ。显然有限的训练样本不可能完全表示所属类别中样本的真实分布,故 $0<\lambda<1$。

3.2　基于边界可信度的相似性度量

给定训练集 D,训练样本所属类别数 $|C|$,则可按照边界距离大小将每个类别划分为 $|C|$ 个边界区域(可重合),设类别中心区域为 0 边界区域,最大边界距离及以外的区域为 $|C|$ 边界区域,则任意文档必属于某一类别的边界区域。显然,对于给定文档 d,其所处的边界区域 $N\in[0,|C|-1]$;如果它越靠近中心所在的 0 边界区,则它与类别 c_i 的相似性越大,反之,如果它远离中心甚至超过最大边界距离,则它与类别 c_i 的相似性越小;而边界表现出了一定的可信度 λ,则文档 d 属于类别 c_i 的基于边界可信度的相似性度量为:

$$Sim(v(c_i),d)\leftarrow Sim(d,v(c_i))\times\lambda^N \tag{2}$$

3.3　基于边界可信度相似的文本分类算法

基于边界可信度相似的文本分类算法描述如下:

(1)训练

输入:所有训练集文档 $D=\{x_1,\cdots,x_l\}$,其中 $x_i\in R^n$,$i=1,\cdots,l$;样本所属的类别 $C=\{c_1,\cdots,c_m\}$。

输出:类中心向量集 $V=\{\boldsymbol{v}(c_1),\cdots,\boldsymbol{v}(c_m)\}$,其中 $\boldsymbol{v}(c_i)\in R^n$,$i=1,\cdots,m$;边界矩阵 $Verge(c_i,c_j)$,$i\neq j$。

步骤:

①计算类别中所有文档向量的算术平均作为该类的类中心向量 $\boldsymbol{v}(c_i)$;

②计算满足条件 $\max_{x_i\in c_a,x_j\in c_b,a\neq b} Sim(x_i,x_j)$ 的边界距离:

$Verge(c_a,c_b)=Sim(v(c_a),x_i)$

③返回 $V=\{v(c_1),\cdots,v(c_m)\}$,$Verge(c_i,c_j)$,$i\neq j$。

(2)分类

输入:待分文档 d。

输出:d 所属文档类别 $c(d)$。

步骤:

①计算 d 与类中心向量 $\boldsymbol{v}(c_i)$ 的相似度 $Sim(\boldsymbol{v}(c_i),d)$;

②比较 $Sim(d,\boldsymbol{v}(c_i))$ 与 $Verge(c_i,c_j)$ 的大小,计算 d 所处的边界区域 N;

③利用式(2)计算文档 d 属于类别 c_i 的基于边界可信度的相似性;

④返回 $c(d)=\arg\max Sim(v(c_i),d)$。

4　实验设计与分析

为检查算法的性能,在 Celeron ® 2.1 GHz,248 MB,WinXP 环境下用 VC++ 6.0 实现本文算法,并与中心向量法分类器和 KNN 法分类器进行了比较。实验采用中、英文两种数据集:中文数据集来自中文自然语言处理开放平台[1],共包括 10 个类别 2 816 个文档,其中训练集 1 882 个文档,测试集 934 个文档;英文数据集采用国际标准的路透社新闻数据集 Reuters-21578[2],

[1] 李荣陆.http://www.nlp.org.cn/docs/download.php?doc_id=1023。

[2] Lewis D D.Reuters-21578 text categorization test collection.http://www.daviddlewis.com/resources/testcollections/reuters21578/。

它有 21 578 个文档，但是并非所有的文档都有人工标注的类别，我们选择其中常用的 7 个类别共 8 398 篇文档，其中训练集 6 010 篇，测试集 2 388 篇，这其中有一些文档被标注了多个类别，使用它的第一个类主题作为它的类标号。

4.1 实验设计与评价

文档预处理过程中，采用信息增益[4]的特征选择算法和基于文档统计的概率估算方法，去掉一些停用词，并针对英文文档采用 stemming[5]技术进行词根还原，最后文档采用 VSM[6]表示。文档 D_i 与 D_j 的相似性度量，本文采用常用夹角余弦距离[7]表示，计算方法为：

$$Sim(D_i,D_j)=\frac{\sum_{k=1}^{M} w_{ik}\times w_{jk}}{\sqrt{\left(\sum_{k=1}^{M} w_{ik}^2\right)\left(\sum_{k=1}^{M} w_{jk}^2\right)}} \tag{3}$$

其中，M 为特征向量的维数，w_n 为向量的第 n 维分量。

评价分类性能[8]的两种常用指标是准确率 p(Precision)和召回率 r(Recall)。对于给定的某个类别，令 a 表示被正确分到该类的实例的个数，b 表示被误分到该类的实例的个数，c 表示属于该类但被误分到其他类别的实例的个数，则准确率 p 和召回率 r 分别被定义为：

$$r=\frac{a}{a+c} \tag{4}$$

$$p=\frac{a}{a+b} \tag{5}$$

另外一个常用的评估指标被称为 F-指标，它的定义为：

$$F_\beta(r,p)=\frac{(\beta^2+1)pr}{\beta^2 p+r} \tag{6}$$

其中参数 β 用来为准确率 p 和召回率 r 赋予不同的权重，当 β 取 1 时，准确率和召回率被赋予相同的权重。本文采用 F1 指标来衡量不同算法在每一个类别上的分类性能。为了评估算法在整个数据集上的性能，有两种平均的方法可供使用，分别称为宏平均(macro_average)和微平均(micro_average)。宏平均是每一个类的性能指标的算术平均值，而微平均是每一个实例(文档)的性能指标的算术平均。本文采用宏平均来评估分类算法在整个数据集上的性能。

4.2 实验结果与分析

本实验中，相关参数设置如表 1 所示：

表 1 三种算法在两种数据集上的参数设置

数据集	特征向量维数	算法		
		中心向量法	K 最近邻算法 k 值	边界可信算法 λ 值
中文数据集(934 篇)	1 000	-	35	0.97
英文数据集(2 338 篇)	1 000	-	45	0.75

中文数据集中，各类别样本分布情况如表 2 所示。三种算法在数据集各类别上分类性能(F1 值)比较如表 3 所示。

从表 2、表 3 的比较中不难看出，数据集中样本在各类别间的分布相对均衡，各类别中样本的分布密度较高，因此三种分类算法都表现出较好的分类效果，但总体上，基于边界可信度相似的快速文本分类算法的分类效果要优于另外两种算法。

英文数据集中，各类别样本分布情况如表 4 所示。

三种算法在数据集各类别上分类性能(F1 值)比较如表 5 所示。

表 2 中文数据集中各类别训练与测试样本分布

类别	用途		类别	用途	
	测试样本数	训练样本数		测试样本数	训练样本数
交通	71	143	教育	73	147
体育	149	301	环境	67	134
军事	83	166	经济	108	217
医药	68	136	艺术	82	166
政治	167	338	计算机	66	134

表 3 三种算法在中文数据集各类别上的 F1 比较

类别	算法		
	中心向量法	K 最近邻算法	边界可信算法
交通	0.985 7	0.948 9	0.985 7
体育	0.947 3	0.960 2	0.947 3
军事	0.871 1	0.874 1	0.883 4
医药	0.937 5	0.938 4	0.945 7
政治	0.913 4	0.920 4	0.916 6
教育	0.965 5	0.916 6	0.965 5
环境	0.929 5	0.883 7	0.929 5
经济	0.909 0	0.869 1	0.904 1
艺术	0.935 6	0.944 0	0.935 6
计算机	0.920 8	0.928 0	0.927 5
Macro-F1	**0.931 5**	**0.918 3**	**0.934 1**

表 4 英文数据集中各类别训练与测试样本分布

类别	用途	
	测试样本数	训练样本数
acq	719	1 650
corn	56	181
crude	189	389
earn	1 087	2 877
interest	131	347
ship	89	197
trade	117	369

表 5 三种算法在英文数据集各类别上的 F1 比较

类别	算法		
	中心向量法	K 最近邻算法	边界可信算法
acq	0.853 8	0.905 0	0.898 6
corn	0.894 7	0.800 0	0.884 9
crude	0.845 9	0.782 3	0.859 4
earn	0.870 7	0.945 8	0.935 7
interest	0.746 9	0.860 8	0.876 9
ship	0.775 9	0.670 9	0.756 7
trade	0.800 0	0.813 6	0.864 8
Macro-F1	**0.826 8**	**0.825 5**	**0.868 1**

从表 4、表 5 的比较中可以看出，数据集中样本在各类别间的分布不均衡，各类别中样本的分布密度较低，此时传统的两种分类算法都显示出一定的不足，对不同的类别表现出的分类效果差异较大；但基于边界可信度相似的快速文本分类算法的分类效果明显较另外两种分类算法稳定，总体分类效果也明显优于另外两种算法，显示了较好的鲁棒性。

表 6 三种算法在两种数据集上的分类时间比较 ms

数据集	算法		
	中心向量法	K 最近邻算法	边界可信算法
中文数据集(934 篇)	125	28 438	141
英文数据集(2 338 篇)	234	223 282	250

(下转 201 页)