基于改进分类模型的文本分类系统实现*

吕 佳
(重庆师范大学 数学与计算机科学学院 运筹学与系统工程重庆市市级重点实验室, 重庆 400047)

摘要:提出一种基于改进的分类模型的文本分类系统来实现文本的自动分类。针对传统的特征提取算法不能很好区分特征词在类内和类间分布情况的缺陷,该系统利用方差对该算法作了改进,用改进的特征提取算法量化各个特征词的权重,为了降低特征向量的维数,采用为每个类建分类器的分类模型,利用遗传算法来修正各个类特征词的权重,直到为每个类训练出能够代表本类的特征向量,最后用这些类的特征向量进行分类。通过在同一数据集上进行对比实验,说明本文提出的改进分类模型的文本分类系统是正确可行的。

关键词:文本分类系统;特征词;特征提取算法;分类模型;遗传算法;KNN 算法
中图分类号:TP391 文献标识码:A 文章编号:1672-6693(2009)02-0079-05

随着网络和信息技术的飞速发展,人们可获得的知识越来越多,面对如此庞大而且不断增长的信息,如何有效地组织和管理,以及快速地找到用户需要的信息,是当代信息科学和技术领域面临的一大挑战。文本分类能够处理大量的文本,可以较大程度解决信息紊乱的现状,方便用户准确地定位所需要的信息。文本分类作为信息检索、信息过滤、搜索引擎、文本数据库、数字化图书馆等领域的技术基础,有着广泛的应用前景[1-3]。

文本数据的半结构化甚至于无结构化的特点,使得表示文本数据的特征向量高达几万维甚至于几十万维。即使经过初始化筛选处理(使用停用词表、稀有词处理、单词归并等),还会有很多高维数的特征向量留下。高维的特征对分类机器学习未必全是至关重要的,有益的。高维的特性可能会大大增加机器学习的时间而仅产生与之小得多的特征子集一样的分类结果。因此,在进行文本分类中,特征选择显得至关重要。特征提取算法能够删除对分类贡献不大的词条,选择出能够代表文本或类别特征的词条,一方面减少了文本向量的维数,另一方面使特征向量更好代表文本或者类别的特征。文本维数的减少,有利于分类算法的运用,使各种各样的分类算法能够运用到文本分类中,为选择更好的分类算法提供了条件。特征向量更忠实于原文本的特征,能够提高文本分类的精度[4]。

本文实现了一种在改进分类模型上的文本分类系统,将遗传算法运用到特征提取算法中,抛弃了为每个文档进行特征提取的传统方法,而是为每个类进行特征提取。首先用改进的特征提取算法量化各个特征词的权重,然后用遗传算法来修正特征词的权重,直到为每个类训练出能够代表本类的特征向量(又叫分类器),最后用这些类的特征向量进行分类。

1 特征提取算法

特征提取算法 TFIDF 的主要思想是:一个有效的特征项若既能体现所属类别的内容,又能将该类别同其它类别相互区分,则认为该特征项具有很好的类别区分能力,适合用来进行分类。其计算公式[5]为

$$\mathrm{Weight}_{\mathrm{TFIDF}}(t) = tf(t) \times \mathrm{idf}(t) \qquad (1)$$

式中,t 是特征项,$tf(t)$ 指特征项频率,是特征项在文档中出现的次数,$\mathrm{idf}(t)$ 指反文档频率,是特征项在文档集分布情况的量化[5],它能弱化一些在大多数文档中都出现的高频特征项的重要度,同时增强一些在小部分文档中出现的低频特征项的重要度,$\mathrm{idf}(t)$ 计算方法为

$$\mathrm{idf}(t) = \log(N/n) \qquad (2)$$

其中 N 为文档集中的总文档数,n 为出现特征项 t 的文档数。

* 收稿日期:2008-10-10
资助项目:重庆市教委科学技术研究项目(No. KJ070802);运筹学与系统工程重庆市市级重点实验室开放课题(No. YC200804)
作者简介:吕佳,女,副教授,博士研究生,研究方向为数据挖掘、最优化技术。

然而传统的特征提取算法存在未考虑到特征词在类间与类内分布情况的缺陷,具体分析如下:1)算法没有考虑特征词在类间的分布,如果一个特征词,在各个类间分布比较均匀,这样的词对分类基本没有贡献;但是如果一个特征词比较集中地分布在某个类中,而在其它类中几乎不出现,这样的词却能够很好代表这个类的特征,然而传统的特征提取算法不能够区分这两种情况;2)同样地,算法也没有考虑特征词在类内部文档中的分布情况。在类内部的文档中,如果特征词均匀分布在其中,则这个特征词能够很好地代表这个类的特征,如果只在几篇文档中出现,而在此类的其它文档中不出现,显然这样的特征词不能够代表这个类的特征。由此可见,在没有考虑到特征项在类间和类内分布的比例情况[6]时,单纯使用特征提取算法会导致很大误差。

此外,为了提高运行效率,往往还需要对文档向量进行压缩处理,仅保留权值较高的特征项,从而形成维数较低的文档向量。这样一来,低频的词条就很有可能被删除。但是有的低频词只是在某一类别的文档出现,这样的低频专指这个类别,传统的特征提取算法未加任何处理,忽略了这些重要低频高权特征项的分类作用。

2 特征提取算法的改进

方差是描述随机变量分布情况的指标,本文用方差来描述特征词在类间的分布情况。如果特征词方差小,说明其在类间的分布比较均匀,这样的特征词对分类贡献不大,可以用方差来降低该特征词的权重;而特征词在类内部的分布情况也可以用方差来描述,与类间分布不同的是,特征词在类内部分布方差越小,即在类内部分布越均匀,特征词越能代表此类,因此在修正特征提取算法公式时,应该将其值调大。

设总共有 n 个类,$tf_i(t)$ 代表词条 t 在 C_i 类的出现频率,$\overline{tf(t)}$ 代表词条 t 的在各个类的平均词频,计算公式为 $\overline{tf(t)} = \frac{1}{n}\sum_{i=1}^{n} tf_i(t)$。令 t 用在各个类间的平均偏差平方为 D_e,则 t 的平均偏差平方计算公式为

$$D_e = \frac{1}{n}\sum_{i=1}^{n}(tf_i(t) - \overline{tf(t)})^2 \tag{3}$$

用 D_e 修正 TFIDF 公式得到

$$\mathrm{Weight}_{\mathrm{TFIDF}}(t) = tf(t) \times \mathrm{idf}(t) \times D_e \tag{4}$$

显然,当 t 均匀分布在各个类间时,由于 D_e 等于0,故 $\mathrm{Weight}_{\mathrm{TFIDF}}(t) = 0$,词条 t 对分类没有贡献。

下面分析词条 t 在各个类内部分布情况,设 C_i 类中总的文档数为 m,将 t 在各个文档的词频看作是 t 在各个文档中的取值,$\overline{tf_i(t)}$ 表示 t 在类 C_i 文档中的平均词频,其计算公式为

$$\overline{tf_i(t)} = \frac{1}{m}\sum_{j=1}^{m} tf_{ij}(t) \tag{5}$$

用 D_{ii} 表示 t 在类 C_i 中的文档中平均偏差平方,则

$$D_{ii} = \frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t) - \overline{tf_i(t)})^2 \tag{6}$$

为了便于表示,将 D_{ii} 增加一个分母,使其值小于1,得到

$$D'_{ii} = \frac{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t) - \overline{tf_i(t)})^2}{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t))^2} \tag{7}$$

同上分析,如果词条 t 在 C_i 类的文档中分布越均匀,D_{ii} 则越小,而 t 却能够代表 C_i 类,相应的 $1 - D'_{ii}$ 就越大,因此可以用 $1 - D'_{ii}$ 来修正 TFIDF 公式,得到

$$\mathrm{Weight}_{\mathrm{TFIDF}}(t_{ik}) = tf(t_{ik}) \times \mathrm{idf}(t_k) \times D_e \times (1 - D'_{ii}) \tag{8}$$

3 分类模型

传统的文本分类,其分类模型如图1所示,都是在所有的训练集中训练一个分类器,而测试集则直接用分类器来决定其类别,如果在所有的候选集中训练一个分类器,特征词条的总数就相当大,特征向量的维数就相当高,每一个特征词条对应遗传算法的一个编码,因而染色体的长度就相当大,用遗传算法训练处理起来就慢,为此,本文为每个类单独的训练一个分类器。

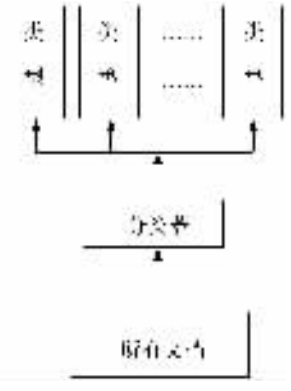

图1 传统的分类模型

本文采用的分类模型如图 2 所示。每一个类都有一个分类器叫做类分类器,要判断一个文档的类别,让每个类分类器判断此文档是否属于该分类器所代表的类。本文采用此种分类模型,首先在每个类的训练文本中,用遗传算法来为该类训练出一个类分类器(类文本向量),因为遗传算法具有很好的寻优能力,借用遗传算法这一特点,在训练文档中训练出一个能够代表这个类的文本作为类分类器。当类分类器训练好后,要看测试文档是否属于该类,比较测试文档和类分类器的相似程度,将其分类到与之相似度最大的分类器所对应的类中。

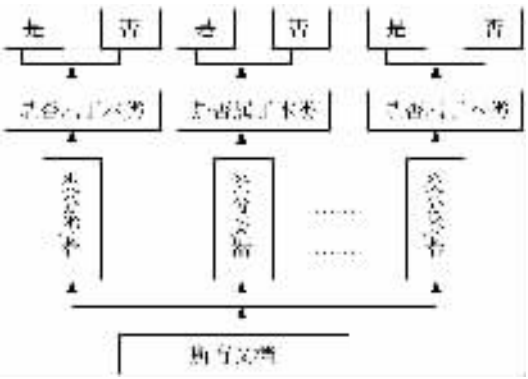

图 2　改进的分类模型

4 用遗传算法进行特征提取的文本分类系统

4.1 系统模型

本文在改进特征提取算法上应用遗传算法,实现了中文文本自动分类系统,该系统由两大模块组成:训练模块和测试模块。其系统模型见图 3。

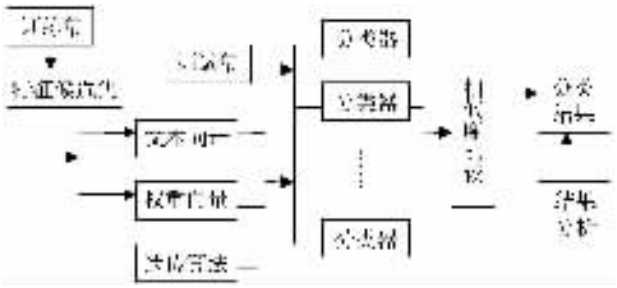

图 3　系统结构图

4.2 系统描述

文本分类系统功能模块如图 4 所示。

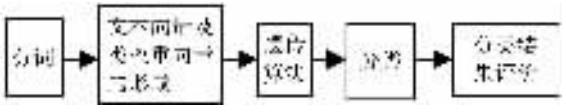

图 4　系统功能模块图

在训练模块中,首先用分词算法对训练文本进行分词处理,分词处理后,这些词、短语成为了候选的特征集合,接着用改进特征提取算法量化这些候选的特征项,再从每个训练文档中,选出前 100 个权重大的特征项组成类的权重向量,同时根据权重向量中特征词在文本中出现与否,将文本表示成文本向量的形式,最后,通过遗传算法训练权重向量,使其能够正确分类此类的训练文本,形成分类器。

测试模块中出现了多个分类器,因为本文在训练的时候,为每个类单独训练了一个分类器,因此分类的时候,训练文本需要和每个分类器比较相似度,最终把文档归到相似度最大的那个分类器所代表的类中。具体做法是:首先将测试文本分词处理,使其成为词或者短语的集,在根据特征权重向量中,特征是否在文本出现,将文本表示成文本向量,用相似度公式,计算文本向量与每个分类器的相似度,将其归类到与之相似度最大的分类器所对应的类中。

分类结果评价模块:包括对文本分类总体效果的查全率、查对率、$F1$ 值计算,以及各个类的查全率、查对率、$F1$ 值计算。

5 实验结果

5.1 实验说明

为了说明改进特征提取算法的优越性,首先将改进特征提取算法与原算法做了对比实验,为了简化分类算法,采用简单易行的 KNN 算法来进行分类[8]。

其次在将遗传算法用于分类的过程中,首先用了改进特征提取算法来量化特征词条的权重,再用遗传算法来修正从每个类选择出来的特征词条组成的权重向量中每个特征词条的权重,最后用训练出的每个类的权重向量来分类,为了检验遗传算法的分类效果,将遗传算法结合改进特征提取算法的分类效果,与改进的特征提取算法结合 KNN 分类算法的最好的分类效果进行比较。

本文中实验数据来源于复旦大学的语料库,其中训练样本和测试样本分别都有 10 个类,训练样本共有 1 882 个文档,而测试样本有 934 个文档,其详细信息见表 1。从表中可以看出,各个类中,训练文档和测试文档的比例大约在 2∶1 左右。

表 1　实验数据介绍

类型	环境	计算机	交通	教育	经济	军事	体育	医药	艺术	政治
训练集	134	134	143	147	217	166	301	136	166	338
测试集	67	66	71	73	108	83	149	68	82	167

通过反复实验试算并结合相关经验,遗传算法的相关参数确定如下:算法采用实数编码,用改进的特征提取算法衡量每个特征项的权重,在根据特征权重排序,从每个训练文档中,抽取了100个特征项,最后经过去重复,每个类的特征项总数便是编码的长度;初始种群为100;交叉概率取0.75;变异概率0.05,变异公式为:$w_k' = \alpha w_k + (1-\alpha) * \Delta$,其中$w_k$为选中的基因位的值,$\alpha$为一个经验值,一般取0.5,$w_k'$为变异后的基因位,$\Delta$为一个随机数;迭代次数为1 000;相似度阈值取0.78。

5.2 实验结果分析

经过反复试验,笔者发现,传统的特征提取算法在k为8和10的时候,分类效果最好,而改进的特征提取算法在k为18的时候分类效果最好。分类结果的评价指标采用查全率、查对率、$F1$值等。文本分类效果总体评价对比情况见表2。从表2中可以看出,改进的特征提取算法的分类效果,无论是查全率、查对率、$F1$值,都要好于传统的特征提取算法。而遗传算法的总体分类效果,从整体上都要优于KNN算法。

表2 分类效果总体评价对比表 %

方法	查全率		查对率		$F1$	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
传统特征提取算法($k=8$)	87.592	88.865	90.817	88.865	88.87	44.50
改进特征提取算法($k=8$)	90.31	92.58	93.33	90.58	91.56	46.87
传统特征提取算法($k=10$)	87.259	88.865	90.866	88.865	88.67	44.39
改进特征提取算法($k=10$)	90.737	91.257	93.223	92.257	91.79	46.69
KNN算法($k=18$)	89.526	90.792	92.270	90.792	90.63	45.35
遗传算法($k=18$)	93.495	93.221	95.332	93.221	92.24	48.65

各个类的分类情况分别见图5,图6和图7。图中横坐标是测试文档的类别编号,分别对应:环境,医药、军事、经济、教育、体育、艺术、政治、计算机、交通类;纵坐标表示分类结果。图5是传统和改进特征提取算法提取的特征用于KNN(当$k=8$时)分类时各个类的分类结果。图6是传统和改进特征提取算法提取的特征用于KNN(当$k=10$时)分类时各个类的分类结果。图7是改进的特征提取算法提取的特征词用于KNN(当$k=18$时)分类结果和遗传算法作为特征提取的分类结果比较。

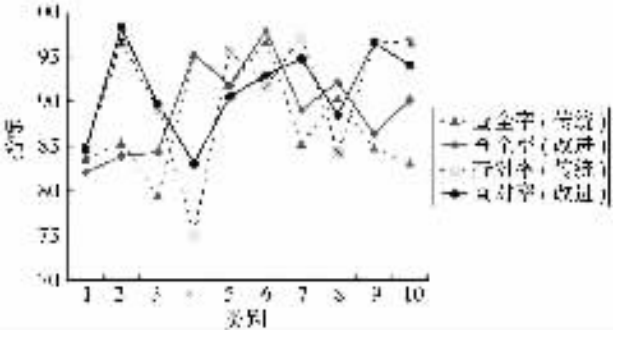

图5 传统和改进特征提取算法分类结果对比图($k=8$)

从图5可以看出,在查全率方面,传统的算法有两个类的值比改进的高,有两个类的值和改进的相等,其余类的值都比改进的低。查对率方面,传统算法有四个类比改进的高,其余都比改进的低。

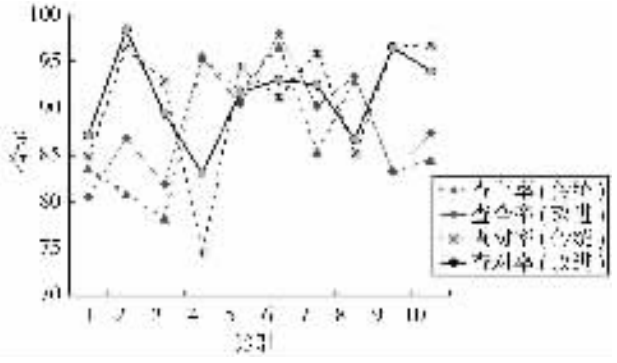

图6 传统和改进特征提取算法分类结果对比图($k=10$)

从图6可以看出,在查全率方面,传统的算法有两个类比改进的值高,两个类和改进的相同,其余都低于改进的值;在查对率方面,传统的有四个类比改进的值高,一个类和改进的相同,其余比改进的低。

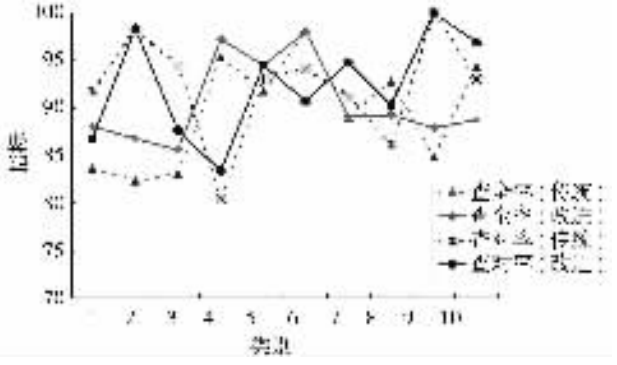

图7 KNN和遗传算法分类结果对比图

从图7可以看出,在查全率方面,KNN方法有两个类比遗传算法的值高,两个类相同,其余类都比遗传算法的值低;在查对率方面,KNN方法有三个类比遗传算法的值高,一个类相同,其余都比遗传算法的值低。

6 结论

本文在改进分类模型的基础上,实现了文本分类系统,能够自动对中文文本进行分类,并对分类结果进行评价。将遗传算法应用在改进特征提取算法上的分类效果,总体的分类正确的文档数,总体分类的查全率,查对率,$F1$值都比KNN分类算法的分类效果略好,从每个类的分类结果来看,遗传算法除个别类的分类效果比KNN分类结果差外,大多数都比KNN分类效果好。这说明本文改进的特征提取算法是正确可行的,且遗传算法在一个类范围内进行特征提取的策略也是可行的。

参考文献:

[1] 吕佳. Web日志挖掘技术应用研究[J]. 重庆师范大学学报(自然科学版),2006,23(4):39-43.

[2] 吕佳. 基于免疫聚类的Web日志挖掘[J]. 重庆师范大学学报(自然科学版),2007,24(2):32-35.

[3] 李晓黎,刘继敏,史忠植. 概念推理网及其在文本分类中的应用[J]. 计算机研究与发展,2000,37(9):1032-1038.

[4] 唐焕玲,孙建涛,陆玉昌. 文本分类中结合评估函数的TEF-WA权值调整技术[J]. 计算机研究与发展,2005,42(1):47-53.

[5] 寇莎莎,魏振军. 自动文本分类中权值公式的改进[J]. 计算机工程与设计,2005,26(6):1616-1618.

[6] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2004,18(1):26-32.

[7] 鲁松,李晓黎,白硕,等. 文档中词语权重计算方法的改进[J]. 中文信息学报,2000,14(6):8-13,20.

[8] 张宁,贾自艳,史忠植. 使用KNN算法的文本分类[J]. 计算机工程,2005,31(8):171-172,185.

Realization of Text Classification System Based on Improved Classification Model

LÜ Jia

(Chongqing Key Lab. of Operations Research and System Engineering, College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: Text classification is to automatically classify an unknown class text into its corresponding text class. With the increasing growth of information, as an important research task in information-processing fields, automatic text classification has nowadays become a research hotspot. A text classification system based on improved classification model presented in this paper is used to realize automatic text classification. The traditional feature selection algorithm doesn't take the distribution of feature terms in inter-class and intra-class into consideration, which makes the algorithm can't effectively weigh the distribution proportion of feature terms. In order to solve the problem, variance in inter-class and intra-class which describes the distribution of feature terms is used to revise weight of the feature term. Then genetic algorithm is applied to feature selection. The traditional idea that selection was done in every document is not adopted here, instead the idea that selection was done in every category is adopted. That is, genetic algorithm is used to modify the weight of the feature term until feature vector trained for every category can represent the feature of this category. Finally, the feature vector trained is used in automatic classification. After some experiments, it has proved that the method proposed is proper and feasible.

Key words: text classification system; feature term; feature selection algorithm; classification model; genetic algorithm; KNN algorithm

(责任编辑 游中胜)