ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.630−639
DOI: 10.3724/SP.J.1001.2008.00630

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向文本分类的混淆类判别技术*

朱靖波+, 王会珍, 张希娟

(东北大学 信息科学与工程学院,辽宁 沈阳 110004)

Confusion Class Discrimination Techniques for Text Classification

ZHU Jing-Bo+, WANG Hui-Zhen, ZHANG Xi-Juan

(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

+ Corresponding author: Phn: +86-24-83672481, E-mail: zhujingbo@mail.neu.edu.cn, http://www.nlplab.com

Zhu JB, Wang HZ, Zhang XJ. Confusion class discrimination techniques for text classification. *Journal of Software*, 2008,19(3):630−639. http://www.jos.org.cn/1000-9825/19/630.htm

Abstract: This paper analyzes confusion class phenomena existing in text classification procedure, and studies further confusion class discrimination techniques to improve the performance of text classification. In this paper, firstly a technique for confusion class recognition based on classification error distribution is proposed to recognize confusion class sets existing in the pre-defined taxonomy. To effectively discriminate confusion classes, this paper proposes an approach to feature selection based on discrimination capability in the procedure of which each candidate feature's discrimination capability for class pair is evaluated. At last, two-stage classifiers are used to integrate baseline classifier and confusion class classifiers, and in which the two output results from two stages are combined into the final output results. The confusion class classifiers in the second stage could be activated only when the output class of the input text assigned by baseline classifier in the first stage belongs to confusion classes, then the confusion class classifiers are used to discriminate the testing text again. In the comparison experiments, Newsgroup and 863 Chinese evaluation data collection are used to evaluate the effectiveness of the techniques proposed in this paper, respectively. Experimental results show that the methods could improve significantly the performance for single-label and multi-class classifier (SMC).

Key words: text classification; confusion class discrimination; feature selection; classification error distribution; machine learning; natural language processing

摘 要: 分析了文本分类过程中存在的混淆类现象,主要研究混淆类的判别技术,进而改善文本分类的性能.首先,提出了一种基于分类错误分布的混淆类识别技术,识别预定义类别中的混淆类集合.为了有效判别混淆类,提出了一种基于判别能力的特征选取技术,通过评价某一特征对类别之间的判别能力实现特征选取.最后,通过基于两阶段的分类器设计框架,将初始分类器和混淆类分类器进行集成,组合了两个阶段的分类结果作为最后输出.混淆类分类器

* Supported by the National Natural Science Foundation of China under Grant No.60473140 (国家自然科学基金); the National 985 Project of China under Grant No.985-2-DB-C03 (国家 985 工程项目); the Program for New Century Excellent Talents in University of China under Grant No.NCET-05-0287 (新世纪优秀人才计划); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z154 (国家高技术研究发展计划(863))

Received 2006-07-02; Accepted 2006-10-10

的激活条件是:当测试文本被初始分类器标注为混淆类类别时,即采用混淆类分类器进行重新判别.在比较实验中采用了 Newsgroup 和 863 中文评测语料,针对单标签、多类分类器.实验结果显示,该技术有效地改善了分类性能.

关键词: 文本分类;混淆类判别;特征选取;分类错误分布;机器学习;自然语言处理

中图法分类号: TP181 **文献标识码:** A

简单来说,文本分类任务可以定义为:根据文本内容赋予 1 个或多个预定义类别或主题类别.核心任务为:根据给定的训练数据,构造高性能的分类器,实现对新文本的自动分类过程.

在实际应用中,根据预定义类别的数量不同,分类系统可以分为两种:两类分类器(binary classifier)和多类分类器(multi-class classifier)[1].其中,两类分类器主要针对正/负或 *A*/*B* 两类分类问题;多类分类器主要针对多类别分类问题,即预定义的类别个数超过两个.如果从文本所属类别的个数来看,文本分类技术又可以分为单标签(single-label)和多标签(multi-label)两种[2−6].单标签分类技术只是给当前文本赋予一个预定义的类别(类别标注);反之,多标签分类技术可能给当前文本赋予多个预定义的类别(多类别标注).本文主要针对单标签多类分类器(single-label and multi-class classifier,简称 SMC)的构建技术,研究如何改善 SMC 的分类性能*.

在文本分类过程中,通过特征降维技术不但可以降低分类系统的时空代价,在一定程度上还可能改善分类系统的性能[7](主要原因在于用于分类器训练的带标样本的数量有限[8].但对于基于支持向量机的分类模型来说,实验结果显示,特征选取技术有损分类性能,但特征抽取技术有可能改善分类性能).特征选取技术(feature selection)的关键在于寻求一个评价函数,根据特征在训练语料中的统计分布信息,对该特征的重要性进行评价.目前常用的特征选取技术[9]有信息增益(information gain,简称 IG)、文档频率(document frequency,简称 DF)、CHI 统计、互信息(mutual information,简称 MI)、TF*IDF、熵(entropy)等.特征抽取技术(feature extraction)[10]主要通过将原始特征进行变换运算,形成新的特征,其中,新的特征表示在物理上不好直接解释.常用的特征抽取技术有潜在语义分析(latent semantic analysis,简称 LSA)、主成分分析(principal component analysis,简称 PCA)、核主成分分析(kernel PCA)、多元判别分析(multiple discrimination analysis,简称 MDA)、独立成分分析(independent component analysis,简称 ICA)等.

本文主要通过分析预定义类别中存在的混淆类(confusion classes)现象,研究混淆类的判别技术,进而改善分类性能.主要研究内容包括:1) 分析了混淆类的特性,并提出了一种基于分类错误分布(classification error distribution,简称 CED)的混淆类识别技术,识别预定义类别中的混淆类集合;2) 提出了一种基于判别能力的特征选取技术,通过评价某一特征对类别之间的判别能力来选择重要特征,其中,认为判别能力强的特征为重要特征,并采用该特征选取技术参与混淆类的判别过程;3) 最后,设计了基于两阶段的分类器设计框架(two-stage classifier).其中,第一阶段分类器称为初始分类器,第二阶段分类器称为混淆类分类器,最后通过组合两个阶段的分类结果作为输出.也就是说,在第一阶段分类结果中,如果该文档被标注为属于混淆类的类别,则在第二阶段中进行混淆类的判别(分类)处理.在比较实验中,采用了两个公开并广泛被用于分类性能评测的语料来测试上述方法的性能,包括 Newsgroup 和 863 中文评测语料.实验结果显示,本文提出的方法很好地改善了分类性能.

1 混淆类识别技术

1.1 混淆类

通常,文本分类系统的构建是基于预定义的类别体系.本文实验中采用的 Newsgroup 语料[11]包含 20 个类别.在文本分类结果中发现,Newsgroup 语料中属于 comp.*讨论组的 5 个类别的测试文本相互误判的情况非常严重,这 5 个类别包括 *comp.graphics*,*comp.os.ms-windows.misc*,*comp.sys.ibm.pc.hardware*,*comp.sys.mac.hardware*

* 本文提出的混淆类判别技术也可以用于改善其他种类分类器的性能,包括单标签两类分类器(single-label and binary classifier,简称 SBC)、多标签两类分类器(multi-label and binary classifier,简称 MBC)和多标签多类分类器(multi-label and multi-class classifier,简称 MMC).

和 *comp.windows.x*.也就是说,分类器对这些类别的判别能力不强,我们称这些类别为混淆类(混淆类相当于一个大类(超类).为了论述方便,文中混淆类和混淆类集合两种说法属于同一个意思).很明显,这些混淆类的存在造成了文本分类性能的下降.因此,本文的研究重点在于混淆类的识别和判别技术,并寻求高性能分类器的构建机制,改善分类性能.

实际上,对混淆类给出一个简单、清晰、明确的定义是不容易的事情.这里,本文尝试分析混淆类的一些特性.混淆类指的是容易混淆的类别集合.换句话说,对于当前分类系统来说,属于混淆类的类别之间存在着严重的误判现象.混淆类的存在与当前给定的训练带标数据相关.例如,Reuters-21578 语料中 *corn*,*grain* 和 *wheat* 类别属于混淆类的原因是由于它们包含的一些训练样本本身同时属于上述多个类别.Newsgroup 语料中属于 comp.*讨论组的 5 个类别中的训练样本的内容主题非常相关,造成了分类器难以准确判别该 5 个类别.

从 Newsgroup 语料的分类实验结果**中,本文分析了混淆类的一些相关特性:

1) 混淆类判定依据应该依赖于分类错误分布,而非类别之间的相似程度.根据直觉考虑,可以通过分析类别相似性计算来判定是否属于混淆类.常用的技术可以采用相似性计算函数(如 Cosine 或 KL 距离)计算两个类别的中心向量的相似度,超过预定义阈值的两个类别或者选择最相似的两个类别,被认为是容易混淆类别.也有一些研究人员通过聚类技术将最相似的类别聚类成一个大类(超类),并认为属于该大类的类别为混淆类[12].其中,基本思想只是考虑各个类别所对应的训练样本的内容来识别混淆类.上述方法采用了一个假设:相似的类别存在严重相互误判的现象,并降低了分类性能.也就是说,相似的类别属于混淆类,不相似的类别不属于混淆类.但是从实验结果中可以发现,该假设不一定成立.存在 31%的类别 *misc.forsale* 分类错误(58 个错判文本中的 18 个文本)属于误判为类别 *comp.sys.ibm.pc.hardware*,两者属于混淆类,而类别 *comp.sys.ibm.pc. hardware* 和 *misc.forsale* 属于不相似类别.

2) 混淆类的识别与分类模型是相关的.类别混淆关系不等同于类别相似关系.混淆类的研究目的是通过减少分类错误来改善分类性能,因而混淆类的识别依赖于分类器的分类错误分布信息(混淆类的特性 1).

3) 混淆类具有整体闭环特性,而不是简单的一对一单向特性.实际上,类别之间的混淆关系(分类误判关系)并非绝对双向的.如分类实验显示,35%的类别 *sci.electronics* 分类错误属于误判为类别 *comp.graphics*;而只有 4%的类别 *comp.graphics* 分类错误属于误判为类别 *sci.electronics*.也就是说,类别 *sci.electronics* 容易误判为类别 *comp.graphics*,反之则不然.实际上,具体两个类别的混淆关系可能是单向的,也可能是双向的.也就是说,类别 *A* 可能存在到类别 *B* 的单向混淆关系,类别 *B* 可能存在到类别 *C* 的单向混淆关系,类别 *C* 可能存在到类别 *A* 的单向混淆关系,最终形成一个闭环的混淆关系,因此可以认为,类别 *A*,*B* 和 *C* 组成一个混淆类集合.单独两个类别由于缺乏双向混淆关系,无法形成闭环的混淆关系,不能组合为混淆类集合.也就是说,评判某个类别集合的混淆程度应该基于整个集合的整体综合评价,并非简单考虑和累加两个类别之间的单向混淆程度.

4) 对于给定的类别体系,可能存在多个不同的混淆类集合,但不同的混淆类集合之间相互独立,没有交集(对于具有交集的混淆类识别和判别技术过于复杂,在本文的研究工作中,只是研究针对不存在交集的混淆类识别技术).

1.2 基于分类错误分布的混淆类识别

本文的研究工作主要针对预定义类别体系中存在的混淆类进行识别和判别.混淆类识别技术的关键在于定义一个混淆评价函数,用于评价某一类别子集的混淆程度.从混淆类特性 1)可以得出,混淆评价函数(confusion evaluation function,简称 CEF)的构建依赖于分类错误分布.本文提出了一种基于分类错误分布(classification error distribution,简称 CED)的混淆类识别技术.一般来说,分类错误分布信息来源于分类实验结果,在本文的工作中,首先将原来的训练数据分为两部分,80%的训练数据用于构建分类器,剩下的 20%训练数据当作测试数据,用于生成分类错误分布矩阵(classification error distribution matrix,简称 CEDM).

** 本文采用针对 Newsgroup 语料分类实验来分析混淆类的特性,该分类实验采用多项式朴素贝叶斯模型(multinomial naïve Bayes)[6]构造分析器,采用所有词汇作为特征(其中去掉禁用词).

为了论述方便,首先引入一些基本概念定义.

假设预定义类别为 $C=\{c_1,c_2,\ldots,c_n\}$,其中包含 n 个类别.分类错误分布矩阵 CEDM 可以表示为 $n\times n$ 的二维关系矩阵,每个元素表示为 $CEDM[i,j]=ErrorRate(c_i,c_j)$,其中,$c_i$ 和 c_j 分别表示第 i 和第 j 个类别.$ErrorRate(c_i,c_j)$是一个错误率函数,计算方法是

$$ErrorRate(c_i,c_j)=\frac{R_{err}(c_i\rightarrow c_j)}{R(c_i)} \tag{1}$$

其中,$R_{err}(c_i\rightarrow c_j)$表示类别 c_i 的文本被误判为类别 c_j 的个数;$R(c_i)$表示类别 c_i 的文本个数.

错误率函数 $ErrorRate(c_i,c_j)$具有如下特性:

1) $ErrorRate(c_i,c_i)=0$;

2) $0\le ErrorRate(c_i,c_j)\le 1$;

3) $ErrorRate(c_i,c_j)$与 $ErrorRate(c_j,c_i)$不一定相等;

4) 如果 $R(c_i)=0$,则针对所有 $j\ne i$,$ErrorRate(c_i,c_j)=0$.

不失一般性,对于给定的任意一个类别子集 $CS_i=\{c_{i1},c_{i2},\ldots,c_{im}\}\subseteq C$,其中包含 $m(\le n)$个类别,评价该类别子集 CS_i 的混淆程度的混淆评价函数 $CEF(CS_i)$定义为

$$CEF(CS_i)=\frac{\sum_{c_l\in CS_i,c_k\in CS_i,l\ne k}ErrorRate(c_l,c_k)}{m} \tag{2}$$

实际上,采用全搜索的方式寻求所有混淆类的方法,由于计算复杂度过高,在实际应用中是不现实的.为此,根据混淆类的第 2 个和第 3 个特性,为了找到具有整体闭环特性的混淆类,本文提出的基于 CED 的混淆类识别算法描述如图 1 所示.

Input: The set of classes $C=\{c_1,c_2,\ldots,c_n\}$, Classification Error Distribution Matrix CEDM;
The process of recognition:
1) Take c_i as the seed category, find candidate subset CS_i that has the maximum confusion. The calculating process is as follows:
$CS_i=\{c_i\}$;
For each class $c\in C$ Do
$c^*=\text{argmax}CEF(CS_i\cup\{c\})$;
If $CEF(CS_i\cup\{c^*\})>CEF(CS_i)$ Then
$CS_i=CS_i\cup\{c^*\}$
Else
STOP
Endif
Endfor
2) In the n candidate confusion subclasses $CSSet=\{CS_1,CS_2,\ldots,CS_n\}$, the criteria condition to determine confusion class CS^* is:
a. $CS^*\in CSSet$;
b. For all classes belonging to $CS^*c_i\in CS^*$, the candidate confusion subset must satisfy: $CS_i=CS^*$. That is, confusion class CS^* must satisfy complete closed-loop characteristic.
Output: All confusion classes set.

Fig.1 Description of CED-based confusion class recognition algorithm

图 1 基于 CED 的混淆类识别算法描述

2 基于判别能力的特征选取技术

2.1 判别能力评价函数

分类器是根据文本内容分析,对每个候选类别赋予一个权重,并根据权重大小进行排序.在 SMC 体系中,采用直接等级排序法(direct rank ordering,简称 DRO)[13]进行排序.常用的方法是构建一个判别函数(discrimination function)来实现类别排序过程.为了改善分类器的性能,较好的方法是增强分类器对类别的判别能力.为了达到这个目的,本文重点研究了基于判别能力的特征选取方法,选择对类别具有较强判别能力的特征参与文本分类过程.

基于判别能力的特征选取方法的关键技术在于寻求特征判别能力的评价函数[14].本文将采用 Kullback-Leibler 距离来评价特征对类别的判别能力.Kullback-Leibler 距离经常被用于两个概率分布的距离计算,距离越大表示越不相似.Kullback-Leibler 距离的定义如下[15]:

$$KL(c_i,c_j)=\int_{\Omega}p(x\mid c_i)\log\frac{p(x\mid c_i)}{p(x\mid c_j)} \tag{3}$$

考虑到 Kullback-Leibler 距离的不对称性,本文采用对称型的 Kullback-Leibler 距离来实现特征判别能力的评价,定义为

$$D_{ij}=D(c_i,c_j)=KL(c_i,c_j)+KL(c_j,c_i) \tag{4}$$

在特征选取过程中,在引入特征之间条件独立假设的前提下,式(4)可以定义为

$$D_{ij}(X)=\sum_{k=1}^{N}D_{ij}(x_k) \tag{5}$$

其中,X 表示特征集合,$D_{ij}(x_k)$表示第 k 个特征 $x_k\in X$ 的对类别 c_i 和 c_j 的判别能力.从式(5)可以看出,特征判别能力函数 D_{ij} 具有单调性,即

$$(x_1\notin X,x_2\notin X)\wedge(D_{ij}^{x_1}\geq D_{ij}^{x_2})\Rightarrow D_{ij}(X\cup x_1)\geq D_{ij}(X\cup x_2) \tag{6}$$

2.2 基于判别能力的特征选取

本文采用基于判别能力的特征选取技术进行特征选取,去掉一些判别能力较差的特征,不仅为了实现特征降维的目的,而且力求增强分类器对混淆类的判别能力.

不失一般性,特征选取过程假设 Y 为原始特征集合.为了获取包含 d 个特征的最佳子集 $X(\subseteq Y)$,首先需要定义一个特征选取评价函数 $J(X)$.评价函数的 J 值越大,表示该特征子集越好.因此,包含 d 个特征的最佳特征子集 $X^*(\subseteq Y)$可以采用如下公式进行构建[7]:

$$J(X^*)=\max_{X\in Y,|X|=d}J(X) \tag{7}$$

但是,该方法需要非常耗时的搜索过程,甚至会造成组合爆炸现象[7].在实际应用中,即使是用很少的特征,也是不现实的.所以,在实际特征选取过程中常常会引入特征条件独立假设,避免疯狂搜索过程[14].

在本文的特征选取过程中采用式(5)来实现特征判别能力评价,代替式(7)中的评价函数 J.但是从式(5)可以看出,该评价函数只能评价特征的局部判别能力.为了能够评价特征的全局判别能力,假设总共有 K 个类别,本文采用了如下 3 种方法:

1) 最大法(max).基本思想是:针对每个特征 x,对所有类别采用式(5)评价该特征的判别能力,选择最大的评价值作为该特征的全局判别能力.特征 x 的全局判别能力的评价方法是

$$D_{\max}(x)=\arg\max_{1\leq i,j\leq K,i\neq j}D_{ij}(x) \tag{8}$$

2) 平均法(average).基本思想是:针对每个特征 x,对所有类别采用式(5)评价该特征的判别能力,选择平均评价值作为该特征的全局判别能力.特征 x 的全局判别能力的评价方法是

$$D_{Ave}(x)=\frac{2}{K(K-1)}\sum_{i=1}^{K}\sum_{j>1}D_{ij}(x) \tag{9}$$

3) 最小法(min).基本思想是:针对每个特征 x,对所有类别采用式(5)评价该特征的判别能力,选择最小的评价值作为该特征的全局判别能力,然后参与到从大到小的特征排序过程中.特征 x 的全局判别能力的评价方法是

$$D_{\min}(x)=\arg\min_{1\leq i,j\leq K,i\neq j}D_{ij}(x) \tag{10}$$

3 两个阶段的分类器设计

为了实现对混淆类的有效判别,改善分类性能的目的,本文采用基于两个阶段的分类器设计框架.在该框架中,可以有效地集成多个不同的分类器.所谓不同的分类器,可能采用不同的分类模型,也可能针对的类别体系

不同,如针对不同的混淆类.从混淆类识别过程可以发现,一个复杂的预定义类别体系中有可能存在多个不相交的混合类集合.在分类过程中,一个混合类集合的判别需要构建专门的分类器,并集成到基于两个阶段的分类器框架中,纠正第 1 阶段的分类错误,最终改善整体分类性能.在该框架中,本文称第 1 阶段的分类器为初始分类器(baseline classifier),第 2 阶段的分类器根据存在的混淆类进行构建,因而称为混淆类分类器(confusion class classifier).

在本文提出的基于两个阶段的分类器设计框架中,首先并非所有类别之间都可能存在混淆关系,相对来说,在给定的预定义类别体系中,不同混淆类集合的个数不会太多.并且,只有当候选类别排序中首位的类别属于混淆类别时(看作默认激活条件),才会激活第 2 阶段的分类器.基于两个阶段的分类器设计主要分为 3 步:第 1 阶段,采用多项式朴素贝叶斯模型[6]构建初始分类器(朴素贝叶斯模型有两种,多项式模型(multinomial model)和多变量伯努利模型(multi-variate Bernoulli model).根据 McCallum 等人的实验结果[6],多项式模型的分类性能优于多变量伯努利模型,因此,本文采用多项式朴素贝叶斯模型来构建分类器).第 2 阶段,如果当前预定义的类别体系中存在 k 个混淆类集合 $CSSet=\{CS_1,CS_2,\ldots,CS_k\}$ 则首先针对每一个混淆类,利用所包含类别的训练语料,类似第 1 阶段初始分类器的构建过程,独立构建一个相应的混淆类贝叶斯分类器,因此可以得到 k 个混淆类分类器;如果不存在混淆类,第 2 阶段就可以跳过不执行.第 3 阶段,将前两个阶段的分类结果集成.

4 实验分析

在本文的比较实验设计中,两个公开标准语料被用于评测和比较分析本文提出的方法,分别为 Newsgroup 语料[11]和 863 中文评测语料[16].

1) Newsgroup 语料.Newsgroup 语料大约包含 20 000 个新闻文本,约平均分为 20 个不同的类别.在语料文本预处理过程中去掉 UseNet Headers、禁用词和在数据集中只出现过一次的词汇,整个预处理过程采用 McCallum 等人开发的 Rainbow 工具(可以从 http://www.cs.cmu.edu/~mccallum/bow/rainbow/下载 McCallum 的 Rainbow 工具[17])完成.McCallum 等人的研究工作显示,Stemming 的处理可能有损于分类性能[6],因此,语料预处理不采用 Stemming 选项.经过语料预处理后,剩余的词汇个数为 62 264.

2) 863 中文评测语料.该语料来源于 2004 年国家 863 中文文本分类评测的语料,其中,采用中图法构建分类体系,共 36 类(原始 863 评测语料的预定义类别体系共包括 38 类,在本文实验中,去掉了 T(工业技术)和 Z(综合性图书)两类,主要原因在于这两类的训练数据构建标准存在一定的争议),每类包含 100 篇中文文本.在语料预处理过程中,分词工具采用东北大学自然语言处理实验室开发的分词工具 NEUCSP(该工具可以从 http://www.nlplab.com/download/CIP/neucsp.zip 下载),去掉禁用词和仅在语料中出现过 1 次的词汇后,剩下的词汇个数为 53 407.

在分类实验过程中,采用 5 次交叉检验的方法,80%语料作为训练语料,剩下的 20%语料作为测试语料,将 5 次交叉检验的分类性能指标取平均值作为最后分类性能评价.实验中,贝叶斯分类器的构建和分类性能评价都采用 Rainbow 工具完成,其中采用正确度(accuracy)作为分类性能评价方法.

实验 1. 混淆类的识别实验结果.

本实验采用基于分类错误分布的混淆类识别技术(如图 1 所示),自动识别 Newsgroup 语料和 863 评测语料中存在的混淆类,并将用于后续实验中.混淆类识别结果见表 1 和表 2.

Table 1 Confusion classes in Newsgroup

表 1 Newsgroup 语料的混淆类

The set of confusion classes (including six classes)		
comp.graphics	*comp.os.ms-windows.misc*	*comp.sys.mac.hardware*
comp.sys.ibm.pc.hardware	*comp.windows.x*	*misc.forsale*

Table 2 Confusion classes in 863 Chinese evaluation corpus

表 2 863 中文评测语料的混淆类

The set of confusion classes (including nine classes)								
TB	TG	TH	TJ	TK	TL	TM	TN	TP

实验 2:基于判别能力的混淆类分类实验.

在此分类实验中,基于多项式贝叶斯模型构建分类器,分别采用 3 种基于判别能力的特征选取方法,包括最大法(max)、平均法(average)和最小法(min).针对 Newsgroup 语料和 863 评测语料中的混淆类,通过分类实验来比较分析 3 种基于判别能力的特征选取技术对混淆类判别的性能.

从图 2 和图 3 的实验结果可以看出,最小法明显比最大法和平均法的性能要差,主要原因在于,如果某一特征对某一特定的类别对的判别能力很弱,那么,即使它对其他类别对的判别能力较强,也无法作为重要特征被选择使用.因而在特征数目少的时候,可能会造成重要特征的丢失.

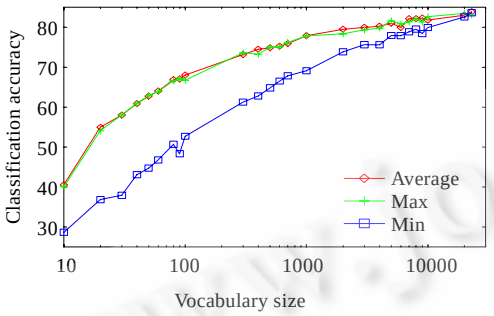

Fig.2 Experimental results of confusion class discrimination on Newsgroup

图 2 Newsgroup 的混淆类判别结果

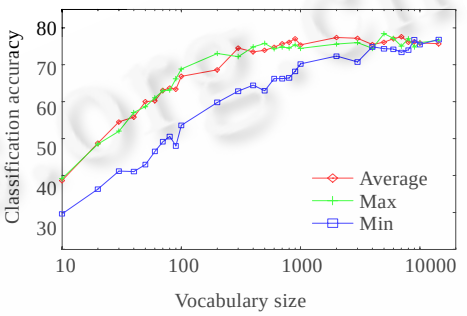

Fig.3 Experimental results of confusion class discrimination on 863 evaluation corpus

图 3 863 评测语料的混淆类判别结果

从最大法和平均法的比较分析中发现一个非常有趣的现象.在 Yang 的论文[9]中显示,基于最大法的 CHI 统计和互信息方法性能优于基于平均法.对于本文提出的基于判别能力的特征选取方法,同样为了考虑全局特征选取性能,采用了 3 种方法:最大法、平均法和最小法.但从图 2 可以看出,在 Newsgroup 语料的混淆类判别中,平均法的性能稍优于最大法.而在图 3 的实验结果中,最大法和平均法的性能曲线相互交错.本文采用 t-检验方法,在给定显著水平α=5%的前提下,最大法和平均法的性能没有显著差异.由此可以得出,在 863 评测语料的混淆类判别中,最大法和平均法的特征选取方法可以看作具有相同的性能,性能曲线相互交错的原因可以理解为是由于在交叉检验中采用语料本身差异所造成的.基于图 2 和图 3 的实验结果综合分析,在下面的分类实验中将采用平均法来实现基于判别能力的特征选取方法,用于混淆类分类器的实现中.

实验 3:基于两个阶段的分类实验.

在本实验中,第 1 阶段的初始分类器采用多项式贝叶斯模型构建,特征选取分别采用 4 种比较常用的技术:信息增益、文档频率、CHI 统计、互信息***.第 2 阶段的混淆类分类器同样采用多项式贝叶斯模型构建,其中,基于判别能力的特征选取方法采用平均法.

其中,第 2 阶段混淆类分类器的激活条件(简称默认激活条件)是:判断第 1 阶段初始分类器的分类结果(排序第 1 位的类别标注,top1)是否属于混淆类,如果是,则激活第 2 阶段的混淆类分类器进行重新分类判别,否则,作为最后分类结果输出.

从图 4 和图 5 的实验结果可以看出,4 种特征选取方法针对基于贝叶斯模型的初始分类器(one-stage)来说,最佳是信息增益 IG,其次是 CHI 统计和文档频率 DF,最后是互信息 MI.其中,互信息比其他 3 种方法的性能差很多,信息增益稍优于 CHI 统计.该结论同样体现在两个阶段的分类器实验结果中,即 two-stage-ig>two-stage-chi >two-stage-df>two-stage-mi.也就是说,two-stage 分类器的性能好坏也与 one-stage 的初始分类器的性能密切相关.从 two-stage 分类器与 one-stage 初始分类器进行比较分析可以发现,针对给定的初始分类器来说,混淆类分

*** 为了获得 CHI 统计和互信息特征选取的全局性能,Yang 的实验结果显示,采用基于最大法的 CHI 统计和互信息性能优于基于平均法的方法,因此,在本文实验中将采用基于最大法的 CHI 统计和互信息实现全局特征选取方法,详细内容参见 Yang 的论文[9].

类器可以明显改善分类性能,即 two-stage-ig>one-stage-ig,two-stage-chi>one-stage-chi,two-stage-df> one-stage-df 和 two-stage-mi>one-stage-mi.但从图 4 的性能曲线可以看出,针对 Newsgroup 语料,two-stage-mi 只是稍优于 one-stage-mi,这一点不同于 863 评测语料上的实验结果.主要原因可能在于,基于互信息的贝叶斯分类器在特征数目较少的情况下,针对 Newsgroup 语料的分类性能太差,因为本文提出的混淆类的识别过程是在采用所有特征的前提下,考虑分类错误分布进行识别的.图 4 和图 5 中分类器设计和特征选取方法见表 3 所示.

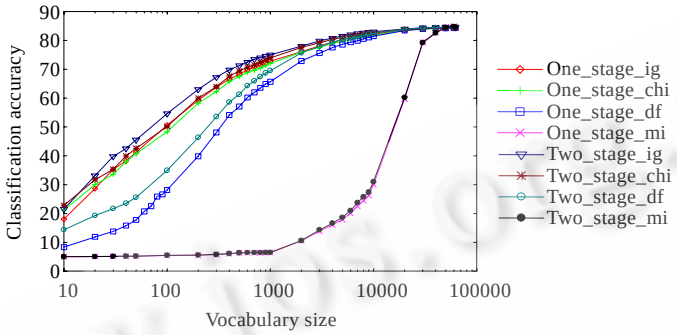

Fig.4 Experiments of two-stage classification on Newsgroup corpus

图 4 在 Newsgroup 语料上的基于两个阶段的分类实验

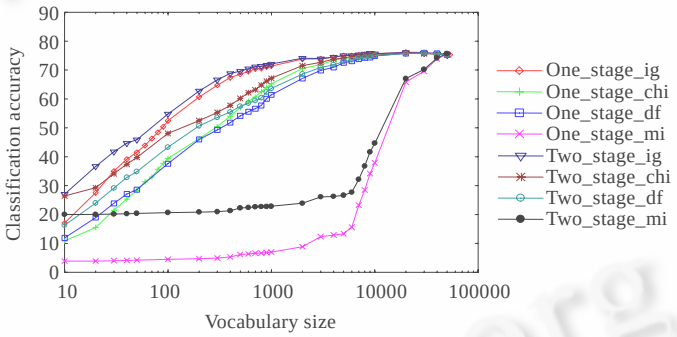

Fig.5 Experiments of two-stage classification on 863 Chinese evaluation corpus

图 5 在 863 中文评测语料上的基于两个阶段的分类实验

Table 3 Description of classifiers in Fig.4 and Fig.5

表 3 图 4 和图 5 中分类器设计表示方法

Symbol in figure	Classifier building	Feature selection method
One-stage-ig	Baseline classifier (multinomial Bayes model)	Information gain
One-stage-chi	Baseline classifier (multinomial Bayes model)	CHI statistic
One-stage-df	Baseline classifier (multinomial Bayes model)	Document frequency
One-stage-mi	Baseline classifier (multinomial Bayes model)	Mutual information
Two-stage-ig	The first stage: Baseline classifier (multinomial Bayes model) The second stage: Confusion class classifier (multinomial Bayes model)	Information gain Average
Two-stage-chi	The first stage: Baseline classifier (multinomial Bayes model) The second stage: Confusion class classifier (multinomial Bayes model)	CHI statistic Average
Two-stage-df	The first stage: Baseline classifier (multinomial Bayes model) The second stage: Confusion class classifier (multinomial Bayes model)	Document frequency Average
Two-stage-mi	The first stage: Baseline classifier (multinomial Bayes model) The second stage: Confusion class classifier (multinomial Bayes model)	Mutual information Average

实验 4:不同激活条件的分类实验.

实际上,第 2 阶段的激活条件可以采用不同的方法.在本实验中,放宽该默认激活条件,改为在第 1 阶段初始分类器的输出类别排序中,如果前 n(=1,2,3)个候选类别都属于混淆类,则激活第 2 阶段的混淆类分类器.本实验

将比较分析扩展后的激活条件对分类性能的影响.从实验 3 可以得出,在 Newsgroup 和 863 评测语料的分类实验中,初始分类器采用基于信息增益的特征选取方法分类性能最佳,因此,在比较实验中采用信息增益作为初始分类器的特征选取方法.在图 6 和图 7 中,top1 表示默认激活条件,top2 表示激活条件考虑前两个候选类别是否属于混淆类,top3 表示激活条件考虑前 3 个候选类别是否属于混淆类.

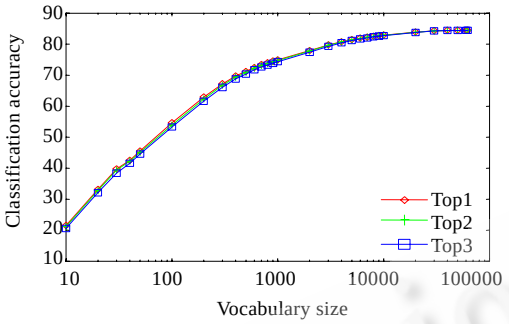

Fig.6 Experiments of classification using different activation conditions on Newsgroup corpus

图 6 在 Newsgroup 语料上的不同激活条件的分类实验

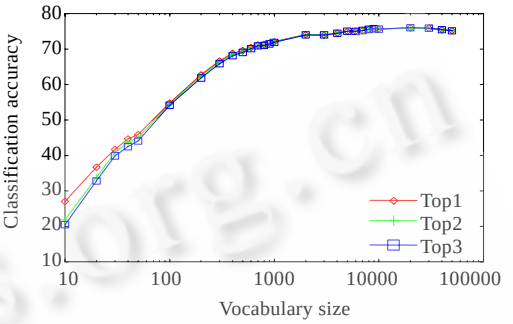

Fig.7 Experiments of classification using different activation conditions on 863 Chinese evaluation corpus

图 7 在 863 中文评测语料上的不同激活条件的分类实验

从图 6 和图 7 的分类实验结果可以看出,默认激活条件 top1 性能略于 top2 和 top3.前文提到,实际上,多类别之间的混淆关系属于单向关系.在 top$n$(n>1)的激活条件中,由于同时考虑前 n 个类别是否属于混淆类,因此很多由于类别混淆关系造成分类错误的测试文本不能被第 2 阶段的混淆类分类器重新分类.本文提出的方法主要针对单标签、多类分类器研究混淆类判别技术.根据混淆类的特性 1),在进行混淆类识别中,只是基于 SMC 体系的分类器的分类错误分布.由于在 SMC 体系中只考虑第 1 个候选类别作为输出,并没有考虑第 n(>1)个类别的分类错误分布.因此,造成 top$n$(n>1)的激活条件在 SMC 体系下的基于两阶段的分类器中效果不如默认激活条件(如果在非 SMC 体系中,则该结论可能有所不同,这将在下一步研究工作中加以验证).

5 结束语

目前,很多研究工作从分类模型选择、特征降维技术和训练语料构建方法等方面来改善分类器的性能,取得了很好的效果.本文主要通过分析文本分类中存在的混淆类现象,深入研究了混淆类的判别技术,改善了文本分类性能.其中,首先分析了混淆类的一些特性,并提出了一种基于分类错误分布的混淆类识别技术,识别预定义类别中的混淆类集合.为了有效地判别混淆类,提出了一种基于判别能力的特征选取技术,通过评价某一特征对类别之间的判别能力来特征选取,实现特征降维目的.最后,通过基于两阶段的分类器设计框架,将初始分类器和混淆类分类器进行集成,组合两个阶段的分类结果作为最后输出.实验结果显示,在 Newsgroup 和 863 中文评测语料上,针对单标签多类分类器体系,本文提出的技术有效地改善了分类性能.实际应用中,单个文档可能属于多个类别,即多标签多类分类器(multi-label and multi-class classifier,简称 MMC).在下一步研究工作中,将针对 MMC 开展混淆类识别和构建两个阶段分类器设计框架的研究.由于多标签的特性,会造成不同混淆类之间存在交集,并且本文提出的混淆类特性 3)和特性 4)也可能需要进行修正,这将是一个研究难点,也是值得进一步探讨的地方.

致谢 在本文的研究工作中,感谢 Prof. Keh-Yih Su 关于基于判别能力的特征选取技术的有价值的讨论,同时感谢实验室的陈晴、王振兴和王安慧同学对混淆类识别算法优化的一些建议.

References:

[1] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002,34(1):1−47.

[2] Lewis D, Schapire R, Callan J, Papka R. Training algorithms for linear text classifiers. In: Proc. of the ACM SIGIR. 1996. 298−306. http://ciir.cs.umass.edu/pubfiles/callansigir96b.ps.gz

[3] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proc. of the Machine Learning: ECML'98, 10th European Conf. on Machine Learning. 1998. 137−142. http://www.cs.cornell.edu/People/tj/publications/joachims_98a.pdf

[4] Lewis D. A comparison of two learning algorithms for text categorization. In: Proc. of Symp. on Document Analysis and IR. 1994. http://www.cs.cmu.edu/~mnr/papers/categ.ps

[5] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: Proc. of the IJCAI '99 Workshop on Machine Learning for Information Filtering. 1999. 61−67. http://www.cs.umass.edu/~mccallum/papers/maxent-ijcaiws99.ps

[6] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In: Proc. of the AAAI '98 Workshop on Learning for Text Categorization. 1998. http://www.scils.rutgers.edu/~muresan/IR/Docs/Articles/aaaiMcCallum1998.ps

[7] Jain AK, Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(2):153−158.

[8] Zhu MH, Zhu JB, Chen WL. Effect analysis of dimension reduction on support vector machines. In: Proc. of the IEEE Int'l Conf. on Natural Language Processing and Knowledge Engineering. 2005. http://www.nlplab.cn/chinese/lunwen.htm

[9] Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. In: Proc. of the 14th Int'l Conf. on Machine Learning (ICML'97). 1997. 412−420. http://www.hpl.hp.com/personal/Carl_Staelin/cs236601/yang1997.ps.gz

[10] Jain AK, Duin RPW, Mao JC. Statistical pattern recognition: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(1):4−37.

[11] Joachims T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Proc. of the ICML'97. 1997. http://www.cs.cornell.edu/People/tj/publications/joachims_97a.pdf

[12] Aggarwal CC, Gates SC, Yu PS. On using partial supervision for text categorization. IEEE Trans. on Knowledge and Data Engineering, 2004,16(2):245−255.

[13] Su KY, Lee CH. Speech recognition using weighted HMM and subspace projection approach. IEEE Trans. on Speech and Audio Processing, 1994,2(1):69−79.

[14] Bressan M, Vitria J. On the selection and classification of independent features. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(10):1312−1317.

[15] Tol JT, Gonzalez RC. Pattern Recognition Principles. Addison-Wesley Publishing Company, 1974.

[16] Chen WL. Research on text feature learning for text categorization [Ph.D. Thesis]. Shenyang: Northeastern University, 2005 (in Chinese with English abstract).

[17] McCallum A, Kachites A. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. 1996. http://www.cs.cmu.edu/~mccallum/bow

附中文参考文献:

[16] 陈文亮.面向文本分类的文本特征学习技术研究[博士学位论文].沈阳:东北大学,2005.

朱靖波(1973－),男,浙江金华人,博士,教授,CCF 高级会员,主要研究领域为自然语言处理.

张希娟(1984－),女,硕士生,主要研究领域为自然语言处理.

王会珍(1980－),女,博士生,助教,CCF 学生会员,主要研究领域为自然语言处理.