基于朴素贝叶斯和遗传算法的两类文本分类方法*)

万狄飞　樊兴华　王国胤
(重庆邮电大学计算机科学与技术研究所　重庆 400065)

摘　要　本文提出了一种基于朴素贝叶斯和遗传算法的两类文本分类方法，该方法将朴素贝叶斯分类器变换为在二维空间中的一条分割线，在分割线临近的文本分类不可靠区间内，利用遗传算法搜索最优文本分割线，从而使分类器达到最佳性能。在由12600篇文本构成的中文语料数据集上的实验表明，该方法具有较高的分类性能和效率，查准率、查全率和F1值分别达到97.98%，91.05%和94.39%。

关键词　文本分类，遗传算法，最优分割线，文本二维空间，朴素贝叶斯分类器

Two-class Text Categorization Method Based on Naive Bayes and GA

WAN Di-Fei　FAN Xing-Huan　WANG Guo-Yin
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065)

Abstract　A two-class text categorization method based on Naive Bayes and GA is proposed. It transforms a Naive Bayesian classifier into a problem of search for a division line that fits the text data set distribution in a constructed two-dimensional space. A genetic algorithm is used to search an optimal division line on unreliable text area. The experiment results on a chinese text data set consisting of 12600 texts show that our method has good performance and efficiency. The precision, recall and F1 are 97.98%, 91.05% and 94.39% respectively.

Keywords　Text classification, Genetic algorithm, Optimal dividing line, Text two-dimensional space, Naive Bayesian classifier.

1　引言

自动文本分类就是在给定的分类体系下让计算机根据文本的内容确定与它相关联的类别，是进行基于内容的自动信息管理的核心技术。国内外在自动文本分类以及相关的信息检索、信息抽取等领域进行了较为深入的研究。文本分类的方法很多[1]，典型的文本分类方法有朴素贝叶斯分类器、基于向量空间模型的分类器、基于实例的分类器和用支持向量机建立的分类器等。

通常人们认为朴素贝叶斯分类器具有较好的分类效率，而支持向量机具有最好的分类性能。本文的目标就是设计一种两类文本分类方法，使之兼顾分类效率和性能两方面。在文[2]中，樊等提出了采用两步分类策略的高性能文本分类方法，其基本思想是直接利用分类器的输出构造一个二维空间，将分类器变换为在此空间中的一条分割线，根据错误分类文本的分布特点(即大多数错误分类的文本聚集在分割直线附近的两侧)确定一个模糊区域，对那些落在模糊区域易于分错的文本进行二次分类。

本文对此二维空间中错误分类文本的分布进行了重新观察，借用支持向量机中最优分割面的思想，发现由分类器直接变换来的分割直线不是一条最优的分割线。据此提出了一种基于朴素贝叶斯和遗传算法的两类文本分类方法，其基本思想为：在此二维空间中不可靠文本区域，利用遗传算法对分割线进行优化获得一条新的具有较高分类能力的分割直线，对文本进行单步分类以保证改造后的分类器既具有朴素贝叶斯分类器的高分类效率，同时又具有近似支持向量机的高分类性能。

2　基于朴素贝叶斯和遗传算法的分类方法

2.1　基于朴素贝叶斯模型的二维空间构造

给定二值文本向量 $d=(W_1,W_2,\cdots,W_M)$，$W_i=0$ 或者1，如果第 i 个特征出现在文本中，$W_i=1$，否则 $W_i=0$。令 $p_{ki}=P\{W_k=1|c_i\}$，$Pr\{\cdot\}$ 表示求事件 $\{\cdot\}$ 发生的概率。两类朴素贝叶斯分类器的判别函数可表示为：

$$f(d)=\log\frac{P\{c_1|d\}}{P\{c_2|d\}}=\log\frac{P\{c_1\}}{P\{c_2\}}+\sum_{k=1}^{M}\frac{\log(1-p_{k1})}{\log(1-p_{k2})}+\sum_{k=1}^{M}W_k\log\frac{p_{k1}}{1-P_{k1}}-\sum_{k=1}^{M}W_k\log\frac{p_{k2}}{1-p_{k2}} \quad (1)$$

当 $f(d)\geqslant 0$ 时，文本 d 属于类型 c_1，否则属于类型 c_2。令

$$Con=\log\frac{P\{c_1\}}{P\{c_2\}}+\sum_{k=1}^{M}\frac{\log(1-p_{k1})}{\log(1-p_{k2})} \quad (2)$$

$$X=\sum_{k=1}^{M}W_k\log\frac{p_{k1}}{1-p_{k1}} \quad (3)$$

$$Y=\sum_{k=1}^{M}W_k\log\frac{p_{k2}}{1-p_{k2}} \quad (4)$$

Con 只与所采用的训练样本集有关，不随文本 d 的变化而变化，为常数；X 表示根据特征估算出来的文本 d 属于类型 c_1 的测度；Y 表示根据特征估算出来的文本 d 属于类型 c_2 的测度，则公式(1)可改写为

$$f(d)=X-Y+Con \quad (5)$$

*)本文受重庆市自然科学基金(2005BA2003，2006BB2374)项目资助。**万狄飞**　硕士研究生，主要研究方向为自然语言处理、机器学习、数据挖掘；**樊兴华**　博士，教授，主要研究领域为人工智能、自然语言处理，信息检索；**王国胤**　博士，教授，博导，主要研究领域包括Rough集理论、神经网络、机器学习、数据挖掘等。

式(5)表示两类朴素贝叶斯分类器可看作是在由 X 和 Y 构成的二维空间中寻求一条分割直线 $f(d)=0$。这样，利用式(3)和(4)，可将文本表示为二维空间中的一个点(x,y)，该点到分割直线 $f(d)=0$ 的距离 $dist$ 为：

$$dist=\frac{1}{\sqrt{2}}(x-y+Con) \tag{6}$$

如图 1 所示，当 $dist\geqslant 0$ 时，表示文本 d 属于类型 c_1；当 $dist<0$ 时，表示文本 d 属于类型 c_2。

我们将公式(1)改写为公式(5)，再演变为公式(6)的目的是：1)利用公式(6)可以在由 X 和 Y 构成的二维空间中方便地考察、分析文本分类错误，探讨在给定分类方法和文本特征集的条件下，距离 $dist$ 与分类错误的关系；2) 利用公式(6)可以根据距离 $dist$ 的大小方便地评估分类的可靠程度。

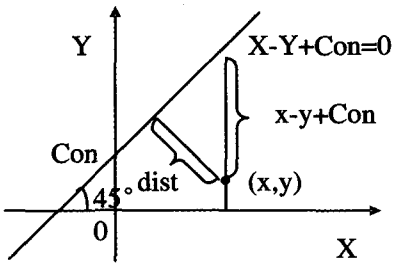

图 1　文体点到分割直线的距离计算

2.2　错误分类的文本观察

以 3 节实验中所使用的语料为样本，以 X 为横坐标、Y 为纵坐标，我们统计出了文本点在二维空间中的分布情况如图 2 所示。图中可以看出，在二维空间中两类文本以条带形状分布在分割直线的两边；被错误分类的文本到分割直线的距离很近。文[2]中给出如下结论。

结论：文本分类器的性能与由公式(6)计算得到的文本到分割直线的距离 $dist$ 有关，大多数的错误发生在一个距离很小的狭窄区域内。也就是说，如果将待分类文本集中到分割直线距离很近的文本去掉，那么分类器在由剩余文本构成的新文本集上的测试性能将会提高。

$$\begin{cases} Dist2\leqslant dist\leqslant Dist1, & \text{对文本 } d \text{ 的任何分类决策都是不可靠的} \\ dist>Dist1, & \text{文本 } d \text{ 属于类型 } c_1 \text{ 且分类结果可靠} \\ dist<Dist2, & \text{文本 } d \text{ 属于类型 } c_2 \text{ 且分类结果可靠} \end{cases} \tag{7}$$

根据结论，可将由 X 和 Y 构成的二维平面分成可靠和不可靠两个区域。根据式(7)进行分类判别。在(7)式中，$Dist1$ 和 $Dist2$ 是由实验确定的两个分界常数，$Dist1$ 为正实数，$Dist2$ 为负实数。

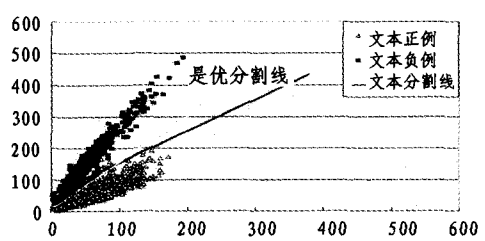

图 2　由 X 和 Y 构成的二维空间中的文本点分布

由图 2 中可以看出，原有的分割线(黑)不是一条最优分割线，如果在不可靠区域内将其调整为如图所示的最优分割线(白)，则分类器的分类能力将提高。

2.3　最优朴素贝叶斯分类模型

最优分割线可以通过在二维文本空间中不可靠区域内对原有分割直线的平移和旋转来实现，即在改写的判定函数公式(5)中，加入了 α 和 β 两个参数，将其改写为公式(8)。

$$f(d)=\alpha * X-Y+\beta * Con=0 \tag{8}$$

公式(8)中，β 用于对原有分割直线的平移，而 α 用于对原有分割直线的旋转。利用公式(8)，我们可以通过遗传算法在文本分布的不可靠区域对参数 β 和 α 的搜索来确定最优分割直线。

2.4　利用遗传算法寻求最优分割直线

2.4.1　遗传算法基本原理

遗传算法(GA)是一种基于自然选择和遗传变异等生物进化机制的全局性概率搜索算法。与基于导数的解析方法和其他启发式搜索方法(如爬山方法，模拟退火方法，Monte Carlo 方法)一样，进化算法在形式上也是一种迭代方法。由于 GA 在问题空间搜索最优值所表现的优良特性，我们考虑将 GA 引入到基于最优朴素贝叶斯分类模型中以确定阈值 α 和 β。

2.4.2　染色体表示

利用遗传算法在二维文本空间不可靠的区域对阈值 α 和 β 进行选择。阈值 α 和 β 的取值范围与二维文本空间中不可靠的区域的范围有关，

若 $Con>0$，则

$$\beta\in\left(1-\sqrt{2}*\frac{|Dist2|}{Con},1+\sqrt{2}*\frac{|Dist1|}{Con}\right); \tag{9}$$

若 $Con<0$，则

$$\beta\in\left(1+\sqrt{2}*\frac{|Dist1|}{Con},1-\sqrt{2}*\frac{|Dist2|}{Con}\right); \tag{10}$$

若 $Con=0$，则 $\beta=0$。 (11)

在二维文本空间不可靠的区域内，文本分割线与 X 轴夹角的范围理论上可取 0°～90°，这里我们取阈值 α 经验值范围为：$\alpha=(0.36,2.75)$，即约为 20°～70°之间。

阈值 α 和 β 是取值在一定范围之内的实数，可以视为遗传算法的表现型形式，从表现型到基因型的映射称为编码。我们采用二进制编码形式，将 α 和 β 变量值代表的个体表示为一个{0,1}二进制串，串长取决于求解的精度。

2.4.3　适应度函数

文本分类中有三个主要的性能、效率评估指标：查准率、查全率和 F-measure。用遗传算法对阈值 α 和 β 的搜索时用 F1 作为适应度函数，F1 值越大证明该分类器的分类性能越好。

查准率(Precision，P)

$$P=\frac{\text{正确分为某类的文本数}}{\text{数据集中分为该类型的文本总数}}\times 100\%$$

查全率(Recall，R)

$$R=\frac{\text{正确分为某类的文本数}}{\text{数据集中属于该类型的文本总数}}\times 100\%$$

F-measure

$$F_1=\frac{2\times P\times R}{P+R}$$

3　实验

为了评估本文提出的方法，分别对文[2]中的中文数据集和朴素贝叶斯分类器以及支持向量机分类器进行了对比实验。

3.1　实验数据集

本文用于实验的中文数据集收集文本共 12600 篇，其中

(下转第 173 页)

"或积"两种运算，较好地实现了上述目的，尤其在不同关系性质条件下，粗糙集理论中的上、下近似计算和对象关系类的矩阵表示定理，为基于粗糙集理论的知识表示与知识获取提供了一种能行与可计算的思路与方法。矩阵论方法在粗糙集理论中应用的意义在于它具有一般性，如：通过定义特征向量的"模"，就可以对粗糙集的精度进行表示与计算，此外，矩阵论方法还可推广到与模糊粗糙集和粗糙模糊集有的基本概念与基本运算的表示，并为之提供能行有效的算法。

感谢 高尚博士为本文提供资料和成文过程中的积极建议。

参 考 文 献

1 Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341～356
2 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001
3 雷晓蔚. 粗集理论的矩阵方法[J]. 计算机工程与应用，2006，42(17)：73～75
4 李龙星，运士伟，杨炳儒. 粗糙集概念与运算的布尔矩阵表示[J]. 计算机工程，2005，31(14)：16～17
5 仁艳玲，朱明放. 基于粗糙集的属性约简的矩阵方法[J]. 陕西理工学院学报，2006，22(3)：76～80
6 张桂芸，黄国兴，杨炳儒. 基于分辨相似矩阵的相似粗糙集的属性约简算法[J]. 计算机工程，2006，32(10)：43～44
7 高学军，丁军. 基于简化差别矩阵的属性约简算法[J]. 系统工程理论与实践，2006，20(6)：101～107

(上接第 161 页)

宣扬、传播对国家安全有害内容的文本为 1800 篇，它们构成属于类型 c_1 的文本集；揭露、批判这种对国家安全有害内容的文本为 3716 篇，它们构成属于类型 c_2^1 的文本集；内容与那些对国家安全有害的内容完全不同，但它们使用的词语中有相当部分是相同的文本为 828 篇，它们构成属于类型 c_2^2 的文本集；其他文本为 6256 篇，它们构成属于类型 c_2^3 的文本集；文本集 c_2^1、c_2^2 和 c_2^3 共同构成属于类型 c_2 的文本集，共 10800 篇。为了模拟现实环境中两类文本出现的实际情况，属于类型 c_1 和属于类型 c_2 的文本数目比例为 1∶6。将属于类型 c_1 和属于类型 c_2 的文本集随机地平均分为四份，以其中的一份构成测试集，另外的三份构成训练集，按四栏进行交叉验证，以四栏实验的平均值作为最终的性能指标。对文本分类的效率用分类所耗时间来进行评估。实验所用 PC 配置如下：CPU(Intel Pentium4 3.0)、内存(DDRII533 1G)。

3.2 特征的选择

在实验中，以词语作为中文语料的特征，文中选用了清华大学开发的 CsegTag3.0 对中文进行分词。文中采用了改进的互信息公式(12)[3]进行特征选择。

$$MI_1(t_k,c_i)=\sum_{i=1}^{n}P\{t_k,c_i\}\log\frac{P\{t_k,c_j\}}{P\{t_k\}P\{c_i\}} \qquad (12)$$

3.3 实验结果

由于测试集中，属于类型 c_1 和属于类型 c_2 的文档比例为 1∶6，如果将所有文本都标记为 c_2，类型 c_2 的分类精度也能达到 85.7%，因此类型 c_2 的分类性能对所选择的分类方法不敏感。为了节约篇幅，在下面的实验中只给出类型 c_1 的分类性能。

实验 1 两个分界常数 *Dist*1 和 *Dist*2 的确定。

错误率和区域百分比两个评估指标，定义如下：

错误率(Error Rate)：

$$ER=\frac{\text{在某区域内错误分类的文本总数}}{\text{数据集中错误分类的文本总数}}\times 100\%$$

区域百分比(Region Per)：

$$RP=\frac{\text{在某区域内的文本总数}}{\text{数据集中的文本总数}}\times 100\%$$

以词语为特征，以改进互信息公式(12)进行特征选择，采用朴素贝叶斯分类器进行分类，确定文本不可靠区间，如表 1 所示。

表 1 给定区域错误率和区域百分比

区域	$14>dist>0$	$0\geqslant dist\geqslant -28$	$14\geqslant dist\geqslant -28$
错误率	28.56%	66.85%	95.21%
百分比	2.08%	34.79%	36.87%

从表 1 可以看出，不可靠区域为 $-28\leqslant dist\leqslant 14$，95.21% 的错误出现在该区域内，而该区域文本总数仅占所有文本的 36.87%。

实验 2 三种分类模型的性能比较

确定阈值 α 和 β 时，遗传算法初始参数设定参考文[4]中给出的经验值，这些经验在一定程度上具有一定的代表性。上一个实验已经求得二维文本空间中不可靠区域为：$-28\leqslant dist\leqslant 14$，*Dist*1 和 *Dist*2 分别为 14 和 -28，在该不可靠区域内利用遗传算法获取最优 α 和 β，三种分类方法性能比较如表 3 所示。

表 3 在中文语料中，三种分类方法性能的比较

Classifier	SVM	Bayesian	OP-Bayesian
α		1	1.16
β		1	0.93
Precision	99.37%	93.35%	97.98%
Recall	88.94%	88.78%	91.05%
F1	93.85%	91.00%	94.39%
Term Num	1000	500	500
time	23154ms	5241ms	6823ms

结论 本文提出了一种基于朴素贝叶斯和遗传算法的两类文本分类方法，该方法利用文本特征估算文本属于两种类型的测度 X 和 Y，以 X 为横坐标、Y 为纵坐标构造二维文本空间，将文本映射为二维空间中的点，将分类器变换为二维空间中的分割直线；利用遗传算法在此二维空间不可靠区域中寻求一条符合语料集分布的最优分割直线，从而使分类器达到最佳性能。在由 12600 篇文本构成的中文语料数据集上的实验表明，该方法具有较高的分类性能和效率。

参 考 文 献

1 Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1): 1～47
2 樊兴华，孙茂松. 一种高性能的两类中文文本分类方法. 计算机学报，2006，29(1)：124～131
3 Sahami M, Dumais S, Hecherman D, Horvitz E. A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization: Papers from the AAAI Workshop, Madison Wisconsin: [AAAI Technical Report WS-98-05]. 1998. 55～62
4 王小平，曹立明. 遗传算法-理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002. 189～200