基于最小类差异的无关信息预处理算法

陈治平, 林亚平, 彭 雅, 王 雷, 童调生
(湖南大学计算机与通信学院, 湖南长沙 410082)

摘 要: 为了降低无关信息对文本分类精度的影响, 提出了基于最小类差异的预处理算法. 算法通过分析文本特征在类中的分布情况, 将特征划分为三种类型, 按照特征在各类间的分布差异, 保留对分类有作用的单类特征与多类特征, 而将类分布差异较小的一般特征进行过滤. 实验结果表明, 采用新算法进行分类预处理所得到的分类精度明显优于信息增益、互信息量等预处理算法.

关键词: 信息增益; 互信息量; 朴素贝叶斯
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2003) 11-1750-04

An Irrelevant Information Preprocess Based on the Minimal Class Difference

CHEN Zhi-ping, LIN Ya-ping, PENG Ya, WANG Lei, TONG Tiao-sheng
(Coll. of Computer and Communication, Hunan University, Changsha, Hunan 410082, China)

Abstract: An irrelevant feature preprocess based on the minimal class difference is proposed. It computes the class distribution difference of features according to their distribution, then divides the features into three types. The new preprocess keeps the features including single-class features and multi-class features which make for classification, and filters the general features with little use for classification. The experimental results show that better performance can be obtained using the new algorithm than using those algorithms such as information gain, mutual information, and cross entropy.

Key words: information gain; mutual information; naive Bayesian

1 引言

随着因特网的普及, 网上文本信息呈指数增长, 如迅速增长的 WWW 信息服务数据、远程教学、远程医疗和数字图书馆中的海量信息, 这种日益增加的海量数据的处理成为重要的研究课题. 而对这些信息进行处理时, 自动文本分类问题是首先需要研究的问题.

文本分类通常采用向量模式或词频矩阵方式表示文本信息[1], 由于所涉及到的文本向量和词频矩阵非常大, 一般需要对文本信息进行预处理, 常用的预处理方法包括特征选取、禁用词表[2](stop list) 和奇异值分解(singular value decomposition, SVD) 等方法. 特征选取主要基于 PCA (Principal Component Analysis)、特征出现频度、文本频度、互信息[3](mutual information) 和信息增益[4]等方法选择能够满足分类性能要求的极少量的特征, 由于过滤的其他特征数目比较大, 容易造成某些重要信息的丢失, 导致分类精度降低. 禁用词表方法是通过人工判别, 将所有文本中都可能使用到的常用词(如英文中的冠词、介词等)放入禁用词表中, 在处理文本时过滤掉这些特征词, 从而减少计算量和提高文本分类的准确性. 这种方法的缺点是要求人工选择禁用词, 同时不同应用领域所对应的无关属性也不一样; 奇异值分解方法对向量或矩阵进行压缩, 但矩阵变换不仅复杂, 运算量大, 而且压缩后的特征库同样会造成重要信息的丢失. 同时这些方法处理后的信息中仍然存在大量的无关信息与弱相关信息, 这些噪声信息的存在也将影响分类结果精度.

为了降低无关信息对文本分类精度的影响, 本文提出一种适合文本分类的最小类差异预处理算法, 通过计算文本特征在各类间的分布差异, 将类分布差异较小的特征进行过滤. 实验结果表明, 在文本分类中采用新算法, 明显优于信息增益、互信息量等预处理算法.

2 过滤模型基础知识

一个样本空间的文本由相互联系的特征词所构成, 为了简化文本处理的复杂度, 一般假设特征词之间完全独立, 这样, 文本就简单地由相互独立的特征所构成, 从而可以使用传统的向量空间模型来表示. 对于文本分类而言, 隶属于同一类的不同文本具有较大的相似性, 表现在由不同特征所构成的高维特征空间中, 属于同一分类的样本具有较多的相似程度比较高的一维子空间.

从单个特征所构成的特征子空间来看, 其在类空间中的

收稿日期: 2002-12-17; 修回日期: 2003-07-22
基金项目: 国家自然科学基金(No. 60272051); 湖南省自然科学基金(No. 01jjy1007)

分布可以分为三种类型: 第一种特征大量地出现在某一类中, 而在其他类中的出现次数非常少; 第二种特征出现在有限的几个文本分类中, 而在其他的分类中出现的情况相对比较少; 第三种特征在各文本分类中表现比较平淡, 如英文中的 of、a、the 等, 在各种类中出现的概率相差不大, 分布比较均匀.

由于单类特征可以看成是多类特征的一种特例, 因此我们主要讨论多类特征与一般特征的情形. 为了区分这两种类型的特征, 首先引入特征所表现出来的类差异性的定义:

定义 1　某特征 A 在类 i 与类 j 之间所表示出来的类差异性为特征A 在两类中的分布概率的相互比值之和.

其类差异性 C(i, j| A) 的计算公式为:

$$C(i,j|A) = \frac{P(C_i|A)}{P(C_i)} / \frac{P(C_j|A)}{P(C_j)} + \frac{P(C_j|A)}{P(C_j)} / \frac{P(C_i|A)}{P(C_i)}$$
$$= \frac{P(C_i|A)}{P(C_j|A)} / \frac{P(C_j)}{P(C_i)} + \frac{P(C_j|A)}{P(C_i|A)} / \frac{P(C_i)}{P(C_j)}$$
$$= \frac{P(C_i,A)}{P(C_j,A)} / \frac{(C_j)}{(C_i)} + \frac{P(C_j,A)}{P(C_i,A)} / \frac{(C_i)}{(C_j)} \qquad (1)$$

其中| C_i| 为第 i 类的文本数目. 利用类差异性 C(i, j| A) 的计算公式(1), 给出多类特征与一般特征的形式化的定义:

定义 2　特征 A 为多类特征: v 8 < {x| x= 1, , , N}, P i, j, 且 j| 8, kI 8, 当 i| 8 时 C(i, j| A) [, 而 C(k, j| A)> .

由于多类特征的分布主要集中于某些类中, 在一定程度上反映了这些类的某一种共性, 保留这些特征对于分类可以起到相应的作用.

定义 3　特征 A 为一般特征: P i, j, C(i, j| A) [

由于一般特征在各类中的分布相差不大, 从特征上不能区分任意一个类. 因此这种类型的特征的存在将导致特征空间维数增加, 使计算过程更加复杂化, 导致分类精度降低. 因此在预处理过程中要求能够尽可能地过滤掉这些特征.

对于某一特征来说, 判断该特征是否是一般特征, 可以利用任意两类之间的差异性 C(i, j| A) 是否都非常小来进行判别:

P i, j, i, j= 1, , , N, C(i, j| A) [, 为某个给定的常量 Z 特征 A 为一般特征.

将上式进行变化可得: $\max_{i,j} C(i,j|A)$ [Z 特征 A 为一般特征.

由于一般特征表现出来对各类的相似性, 或者说, 这一类特征在各类中的分布相差不大, 不会提高分类的精度, 相反会降低分类精度. 因此在分类的预处理过程中如何有效地过滤一般特征, 保留单类特征与多类特征是分类预处理必须要解决的问题. 根据这种思想, 我们提出了基于最小类差异的过滤模型.

3　最小类差异过滤模型

设数据样本集合中有 m 个不同类, C_i 表示第 i 个类, (i = 1, , , m), | C_i| 表示 C_i 中的样本数. 在对文本信息进行预处理过程中, 设 A 为待选择的特征. 定义 A 在 C_i 类中的类影响程度因子为:

$$E_i(A) = \left(\sum_{j=1}^{|C_i|} d_{ij} + 1 \right) \backslash |C_i|, \quad i = 1, , , m \qquad (2)$$

其中 d_{ij}I {0, 1}, 当 C_i 类中的第 j 个样本包含特征 A 时 d_{ij}= 1, 否则 d_{ij}= 0.

令:
$$E_{min}(A) = \min(E_i(A)), \quad i = 1, , , m$$
$$E_{max}(A) = \max(E_i(A)), \quad i = 1, , , m \qquad (3)$$

定义特征 A 对类的区分程度因子 R(A) 为:

$$R(A) = (E_{max}(A) / E_{min}(A)) @ (m / N(A)) \qquad (4)$$

其中 N(A) 为文本中包含特征 A 的类的数目, 0[N(A) [m.

从式(4)容易得知, 由于 $E_{min}(A)$ [$E_{max}(A)$, N(A) [m, 因此 R(A) \ 1.

考虑特征词是常用冠词和介词的情况, 由于这些特征词一般在每一个样本中都会出现, 因此, 在| C_i| 相同的情况下, 一般有 $E_{min}(A) = E_{max}(A)$, N(A)= m, 由公式(4)可知 R(A) 取最小值 1. 实际上这些特征词对于后续文本分类作用不大, 因此这些词可以进行过滤, 其作用类似于禁用词表方法, 不同的是过滤的特征词不必人工选择, 而是通过计算自动选取. 这种方式对于基于领域性的文本进行分类时效果特别明显, 如基于计算机科学领域的论文集, 文本中一般都包含 computer 特征词, 且出现在各类中的频度大致相等, 由此得到的 R(A) 也接近于 1. 实际上这些词不能体现计算机科学的某一类的特性, 对分类作用不大. 因此, 使用这种方法可以有效地降低分类计算中的复杂度, 从而提高分类的准确度.

4　实验及分析

我们采用 RainBow 文本分类系统[5] 作为实验测试平台. 系统自带 2000 篇新闻组短文章, 预先已分为 20 类, 每类 100 篇文章, 系统通过扫描文本集合所得到的特征词表包含 37344 个特征词, 系统自带的禁用词表包含 524 个特征词. 同时系统还提供了用 perl 脚本编写的 rainbow2stats 程序用于评价算法好坏的性能指标: 平均精度. 其平均精度的计算是在设定训练样本数以及重复测试若干次实验的基础上计算其平均的正确率.

在实验过程中, 我们从每类中选取 90 篇作为训练样本, 另外 10 篇作为测试样本, 同时设定系统重复 10 次, 并分别利用朴素贝叶斯[6] (Naïve Bayesian, NB)、最大期望值[7, 8] (Expected Maximum, EM)、K 近邻[9] (k2nearest neighbor, KNN)、支持向量机[10] (Supported Vector Machine, SVM) 算法进行分类, 得到相应分类结果.

首先, 我们利用 RainBow 系统在不使用禁用词表与给定系统自带禁用词表的情况下进行测试, 得到相应的分类结果如表 1 所示.

表 1　不使用预处理算法所得到的分类结果

方式	过滤特征词数	剩余特征词数	平均精度			
			NB	EM	KNN	SVM
无禁用表	0	37344	72.00	72.75	50.21	61.37
固定词表	481	36863	81.00	81.00	58.29	61.87

从表 1 中可以看出, 对于没有采用禁用词表的分类处理, NB 方法与 EM 得到的平均精度只有 72.%, 而 KNN 方法只有 50.21% 的平均精度; 利用系统自带的禁用词表 NB、EM、KNN 与 SVM 三种方法所得到的平均精度分别达到了 81%、81%、58.29%、61.87%. 通过分析, 这是由于系统自带禁用词表中的噪声信息的过滤使得在进行 NB、EM 的计算过程中, 提高了有用特征词的概率计算, 因此分类的平均精度均有所提高. 因此, 无关特征的存在可以影响分类的精度, 因此尽可能地过滤无关特征, 将会提高分类的精度.

利用本文提出的预处理算法, 并设定门槛值 R= 1 进行实验, 得到如表 2 所示的禁用词表(包含有 12 个特征词)

表 2 门槛值 R= 1 时的禁用词表

Path cantaloupe srv cs cmu edu from newsgroups subject message id date

通过分析发现, 由于训练样本集由新闻组成, 这些新闻都包含有路径、主题等信息(如表 3 所示), 因此, 这些特征词出现在所有的新闻中, 但没有出现在系统自带的固定禁用词表中. 使用新的预处理算法可以有效地将这些无关特征进行过滤.

表 3 新闻样例

Newsgroups: comp. graphics
Path: cantaloupe. srv. cs. cmu. edu! crabapple. srv. cs. cmu. edu! bb3. andrew. cmu. edu! news. sei. cmu. edu! cis. ohio2state. edu! zaphod. mps. ohio2state. edu
! cs. utexas. edu! uunet! pdn! parsec! rjs002c
From: rjs002c@parsec. paradyne. com(Robert Synoski)
Subject: 24bit Graphics cards
Message2ID: 3 1993Apr14. 215934. 17733@pdn. paradyne. com4
Sender: usenet@pdn. paradyne. com(News Subsystem)
Nntp2Posting2Host: parsec
Reply2To: rjs002c@parsec. paradyne. com
Organization: AT&T Paradyne, Largo Florida
Date: Wed, 14 Apr 1993 21: 59: 34 GMT
Lines: 10

I am looking for EISA or VESA local bus graphic cards that support at least
1024@786 @24 resolution. I know Matrox has one, but it is very expensive. All theother cards I know of, that support that resoul2tion, are striaght ISA.

Also are there any X servers for a unix PC that support 24 bits?
Thanks

调整门槛值的大小, 并结合 NB、EM、KNN 以及 SVM 等算法进行分类, 得到相应的分类结果与过滤特征词的情况分别如图 1、2 所示.

从图 1 的结果可以得出, 随着门槛值的提升, NB、EM、KNN 三种分类算法的平均精度都有明显的提高, 其中 NB 算法与 EM 算法都由原来采用常用的禁用词表的 81% 提高到 92%, 而 KNN 算法则由 50% 提高到 81.22%. 相应地, 从图 2 可知, 随着门槛值的提升, 系统过滤的特征词量也不断增加, 并且增加的现象在门槛值为 50~ 60 变化非常大, 过滤的特征词量由 12698(R= 50) 到 32568(R= 60), 在这个阶段中所排除掉的特征词量达到近 20000 个, 而系统剩余的特征词仅为 4776 个. 在门槛值 60 以后特征词的变化量基本上处于一个比较平稳的状态. 分析认为, 由于类区分程度因子小于 50 的特征属性主要是一般特征, 如/ of0 √ a0等, 而在 50~ 60 之间则是大量的多类特征, 这些特征词具有一定的信息内容但对分类精度影响不太大; 而 60 以上则主要表现为对类的区分程度比较大而数量又比较少的单类特征.

结合图 1 的结果和表 1 的数据, 可以得出, 新算法明显优于固定禁用词表方法.

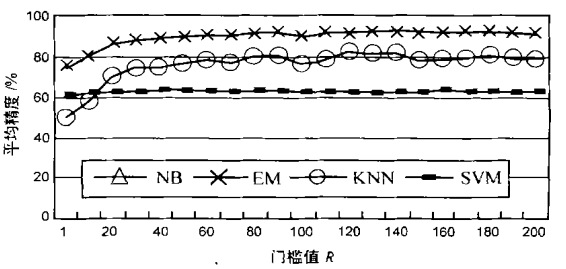

图 1 使用预处理算法所得到的分类结果

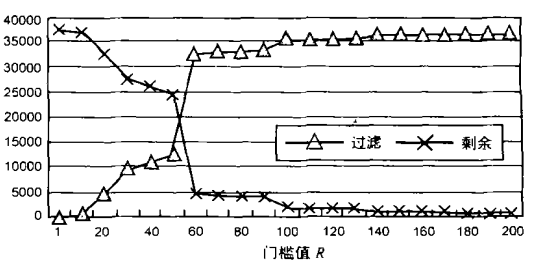

图 2 不同门槛值特征词过滤情况

在实验中, 模型对 SVM 分类算法的精度提高不太明显. 结合对 SVM 算法的研究表明, 由于 SVM 算法是利用最佳超平面将其中一类与其他类进行分离, 其思想与新算法在模式上有共同之处, 即可以简单地将一个特征词作为一个超平面进行分割, 从而导致过滤算法对 SVM 等算法的分类结果影响不大.

为了与其他特征预处理方法进行比较, 我们进行了另外的实验. 在相同的实验环境下利用系统自带的信息增益算法、互信息量、交叉熵[7] (Cross entropy) 与最小类差异的过滤算法进行了比较实验. 首先, 利用新算法并设定门槛值为 100(剩余特征词量接近 2000 个) 进行了测试; 其次, 利用信息增益算法、互信息量算法与交叉熵, 并挑选 2000 个特征词进行重复实验, 得到表 4 所示的结果.

从表 4 的实验数据可以得知, 基于最小类差异的算法所得到的分类精度明显高于使用信息增益与交叉熵所得到的分类精度. 这主要是因为在基于最小类差异预处理中将大量的一般特征进行过滤, 从而有效地降低无关特征对分类算法的影响, 使分类精度得到提高.

实验中基于互信息量的特征预处理方法所得到的精度比

较低, 这主要是由于在特征选取算法中, 互信息表示的是某个特征词与某个类的相互关系, 其公式为:

$$MI(t) = \sum_i I(t, c_i) = \sum_i \log(p(t/c_i)/p(t))$$

其中 $p(t)$ 表示某特征词的出现概率, $p(t/c_i)$ 表示该特征词在某个类 c_i 中出现的概率, 当 $p(t)$ 越大, 而 $p(t/c_i)$ 越小, 甚至远小于 $p(t)$, 表明该特征在这一类中出现的频度非常低, 不足以代表该类的特征, 计算出的值可能出现负数, 但这种特征词对于其他类来说相应的信息量比较大, 可以得到比较大的正数, 在评价该特征时, 由于负数的加入, 降低了该特征词所应具有的信息量, 从而导致互信息算法用作特征选取进行分类得到的效果非常低劣.

表 4　不同预处理方法的分类结果

使用方法与测试精度			1 次	2 次	3 次	4 次	5 次	平均值
方法	剩余特征词量	分类算法						
最小类差异算法 (R= 100)	1876	NB	90.3	91.00	90.75	88.25	91.25	90.31
		KNN	77.04	79.38	74.67	78.06	76.15	77.06
		EM	90.35	89.62	90.35	91.00	90.50	90.36
信息增益	2000	NB	83.5	82.5	84.5	82.00	83.0	83.1
		KNN	57.5	61	53	57.5	55	56.8
		EM	87.0	79.5	83.0	86.0	85.0	84.1
互信息量	2000	NB	27.5	28.5	24	29	30.5	27.9
		KNN	22.5	24.5	20	20.5	18	21.1
		EM	27	28.5	28.5	32.5	33	29.9
交叉熵	2000	NB	83	79	83.5	86.5	85	83.4
		KNN	66	60.5	56	62.5	59	60.8
		EM	88.5	83.5	85.0	82.5	85	84.9

5　结束语

利用最小类差异的方法可以有效地将大量的与分类无关的信息或弱相关信息进行过滤, 由于这些信息的过滤, 使得文本信息在进行分类的过程中不需要再对这些信息进行比较, 从而在一定程度上加快了分类的速度, 并且由于分类过程中无关信息的有效去除, 使得分类过程更加准确.

本文研究的文本对象是英文文本, 对于中文文本的处理以及多元组的特征选取也可以利用本文提出的算法实现无关信息的过滤.

参考文献:

[1] 陈治平, 林亚平, 等. 基于 N 层向量空间模型的信息检索算法[J]. 计算机研究与发展, 2002, 39(10): 1233- 1237.

[2] Christopher Fox. A stop list for general text[J]. SIGIR Forum, 1990, 24(1): 19- 35.

[3] Georgios C A, et al. Category regions as new geometrical concepts in Fuzzy2ART and Fuzzy2ARTMAP[J]. Neural Networks, 2002, 15(10): 1205- 1221.

[4] S Ruggieri. Efficient C4. 5[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(2): 438- 444.

[5] Rainbow Text Classification System [EB/OL]. http://www.cs.cmu.edu/mcallum/bow/.

[6] 范焱, 等. 用 Naïve Bayes 方法协调分类 Web 网页[J]. 软件学报, 2001, 12(9): 1386- 1392.

[7] Dempster A, et al. Maximum likelihood form incomplete data via EM algorithm[J]. Journal of Royal Statistical Society, Series. B, 1977, 39: 1- 58.

[8] Tom M Mitchell. 机器学习[M]. 北京: 机械工业出版社, 2003.

[9] Thomas G A, et al. Information filtering in TREC29 and TDT23: A comparative analysis[J]. Information Retrieval, 2002, 5(223): 159- 187.

[10] 陶卿, 等. 一种新的机器学习算法: Support vector machines[J]. 模式识别与人工智能, 2000, 3(13): 285- 290.

作者简介:

陈治平　男, 1971 年生于湖南省安化县, 讲师, 主要研究方向为机器学习.

林亚平　男, 1955 年生于湖南省邵阳市, 先后获湖南大学学士学位、国防科技大学硕士学位、湖南大学博士学位, 现为湖南大学计算机与通信学院教授、博士生导师, 主要研究方向为计算机通信网络和机器学习.