文章编号:1001 - 9081(2005)11 - 2658 - 03

基于类别的特征选择算法的文本分类系统

蒋伟贞[1],陶宏才[2]
(1. 暨南大学 信息科学技术学院,广东 广州 510084;
2. 西南交通大学 信息科学与技术学院,四川 成都 610031)
(vivianjiang00@sina.com)

摘 要:目前的索引词选择算法大多是基于词频的,没有利用训练样本中的类别信息,为此提出了一种新的基于类别的特征选择算法。该算法根据某个词是否存在于文档中导致该类文档相似度的区别,来确定该词区分不同文档的分辨力,以此分辨力作为选取关键词的重要度。以该算法为基础,设计了一个英文文本自动分类系统,并对该系统进行了测试和结果分析。

关键词:文本自动分类;特征选择;向量空间模型;朴素贝叶斯;分辨力
中图分类号: TP181 **文献标识码**:A

An automatic text classifier of class-based feature selection algorithm

JIANG Wei-zhen[1], TAO Hong-cai[2]
(1. *School of Information Science and Technology, Jinan University, Guagnzhou Guangdong* 510084, *China*;
2. *School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan* 610031, *China*)

Abstract: Current feature selection algorithms are all based on term frequency, and ignore the class information in the training sample set. A new feature selection algorithm based on class information was put forward. The principle of the algorithm is as follows: according to the similarity difference caused by whether or not a word existed in a document, the discriminative power with that this word distinguished different documents could be determined. And then, the discriminative power was taken as the importance for keyword selection. Based on this algorithm, an automatic English text classifier was designed, and the system test and result analysis were made.

Key words: automatic text classification; feature selection; VSM model; naive Bayes; discriminative power

文本分类中的特征选择一直是文本分类的关键技术和瓶颈技术。作者对当前文本分类中各种常用特征选择算法的性能以及优缺点进行了分析后,发现目前的索引词选择算法都是基于词频的[1],没有利用训练样本中的类别信息。为此,提出了一种新的基于类别的特征选择方法,并以此为基础设计了一个英文文本自动分类系统。

1 基于类别的特征选择算法

在向量空间模型中,文档被表示为一个 N 维空间,空间的每一维表示一个特征项,特征项的权重对应这一维的坐标值。两个文档之间的相似度可以用以下公式计算:

$$S_{jk} = \frac{\sum_{i=1}^{N}(w_{ij} \cdot w_{ik})}{\sqrt{\sum_{i=1}^{N} w_{ij}^2 \cdot \sum_{i=1}^{N} w_{ik}^2}} \quad (1)$$

其中,w_{ij} 和 w_{ik} 表示文档 D_j 和文档 D_k 的索引词权重,S_{jk} 表示文档 D_j 和文档 D_k 的相似度。

G. Salton 提出[2],在文本分类中,某一类文档的密度,可用该类的每对文档间的相似性总和来表示:

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{k=1,k\neq i}^{N} Sim(D_i, D_k) \quad (2)$$

其中,$Sim(D_i, D_k)$ 表示文档 D_i 和文档 D_k 的相似度。Q 表示该文档类的密度。

那么,如果要测试一类中某一个索引词区分文档的能力,可用该索引词对该类文档的相似度的贡献来代表,即:用 Q 来表示索引词存在于该文档前的类的总体相似度,而用 Q_j 表示索引词存在于该文档后的类的总体相似度,计算 Q 和 Q_j 的差值,可得到一个索引词的分辨力 dv_j(Discriminative value,以后简称为 DISC)。

$$dv_j = Q - Q_j \quad (3)$$

在得到一个类的质心后,两个类之间的相似度可以用一个类中所有文档与另一类的质心的相似度总和来计算。那么,一类中某一个索引词区分两类文档的能力,可用该索引词对该两类文档的相似度的贡献来代表,用(3)式计算索引词存在于该文档前的类的两个类间相似度总和 Q 和索引词存在于该文档后的两个类间相似度总和 Q_j 的差值,同样可得到一个索引词的区分两类文档的分辨力 dv_j。

综上所述,可以提出一种新的索引词选择方法:基于训练集中的初始分类,计算出其中的每个词对其所属类的分辨力 dv_{j1},以及该索引词区分其所属类与其他类的分辨力 dv_{j2},将分辨力作为选取索引词的重要度计算基础,用于新文档分类。这一新的特征抽取方法利用了训练集中的类别信息,从一个新的角度出发考虑特征提取,目的是以此为基础开发出高精度的分类算法。

收稿日期:2005 - 05 - 24;**修订日期**:2005 - 08 - 08

作者简介:蒋伟贞(1975 -),女,广西玉林人,助教,主要研究方向:网络数据库; 陶宏才(1964 -),男,湖北武汉人,副教授,主要研究方向:计算机网络与信息系统、数据库、网络安全.

算法描述如下:

在去除停用词以及词干化处理后,从训练集中的所有文本中获得 N 个词 W。对这 N 个词 $W_j \in W(0 \leqslant j \leqslant N)$,进行以下计算:

1) 计算每个类的整体密度 Q。计算每个类的质心 C。对每个词 W_j,计算该词对其所属类的分辨力 dv_{j1},以及该索引词区分其所属类与其他类的分辨力 dv_{j2};

2) 按照一定的次序循环计算所有词的分辨力;

3) 基于 dv_{j1} 和 dv_{j2},按照一定的标准决定是否保留这个词;

4) 重复 1),2),3),最后获得关键词。

2　文本分类系统设计

2.1　系统体系结构

根据以上算法,采用朴素贝叶斯分类方法,设计了一个采用基于类别的特征选择方法的英文文本自动分类系统,其体系结构图 1 所示。

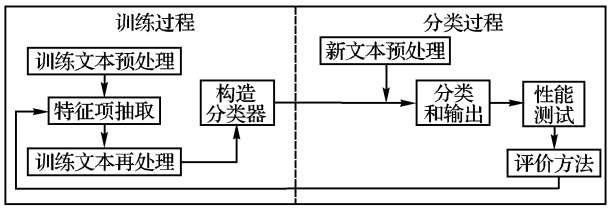

图 1　文本分类系统体系结构

系统由两大部分组成,分别是训练部分和分类部分。其中,训练部分包括训练文本预处理、特征项抽取、训练文本再处理、构造分类器等模块;分类部分包括新文本预处理、分类和输出、评价方法以及性能检测等模块。

2.2　训练过程

在训练过程部分,要进行文本预处理,计算关键词的分辨值,并根据关键词的分辨值训练分类器,该过程得到的是分类器的参数值。

训练过程描述如下:

1) 输入训练文本,获得文本中所有单词,计算词频矩阵。去除高频低频词,去除停用词列表中的词,利用 porter stemmer 算法[3] 对词取词干(去除前后缀)。

2) 根据词频矩阵计算预处理文本的相似度和类密度,根据公式(2),(3)计算上一步得到的所有词的分辨值,根据分辨值保存符合标准的关键词。

3) 再次输入训练文本,对上一步得到的关键词,根据贝叶斯算法[4] 计算关键词属于各类的概率。将分类器特征项属性表输出保存。

2.3　分类过程

在这个过程中,根据训练过程得到的分类器,对新文本分类,并用评价方法检测分类性能。这个过程最终得到的是分类器的性能。

分类过程描述如下:1) 输入新文本,进行文本预处理;2) 根据特征项属性表获得文本中含有的关键词并计算其词频;3) 计算文本所属类别,输出类别;4) 比较文本的分类类别和实际类别,计算分类精度;5) 使用评价函数对分类精度进行评价,反馈给系统。

3　测试及结果分析

特征抽取作为分类的前处理过程,其有效性可以通过分类的效果来测试。为评价分类效果,我们采用最通用的性能评价方法:召回率 R (Recall)、准确率 P (Precision) 和 $F1$ 度量。

为客观评价本分类系统的分类性能,采用了标准文档集 20_newsgroups[5] 进行测试。

3.1　特征算法有效性测试

为了排除未知因素对结果的影响,在获取关键词和训练分类器时使用同样的训练文档。先使用 20_newsgroups 数据集的 20 类,每类 300 篇,共 6 000 篇英文文本对分类系统进行训练。在训练文档集中,经过文本预处理,去除停止词,词干化处理后,得到一个包含 23 379 个单词的词汇表。以这 23 379 个单词为基础,从训练文档集中根据词的分辨力初步获取关键词。

在获取关键词时,采用不同的分辨值阈值进行特征提取。分辨力阈值的确定,是一个比较关键,也是一个比较困难的问题。理论上,没有很好的解决方法,一般采用预定初始值,然后给出测试文本,使用分类器进行分类,再根据分类的准确程度调整初始值的方法。

首先,使用传统的特征提取方式(文档频率 DF)和公认为是目前对英文文本的分类精度最好的 SVM 算法对同样的训练文档集进行分类。提取词频位于 25% ~75% 范围内的词作为关键词,使用网上下载的 libsvm 分类系统[6] 进行分类。分类的结果是 80.3% 的准确率和 78.3% 的召回率。计算此时提取的关键词的分辨值,记录这些词的分辨值的分布范围。

然后,使用本文提出的特征提取方法,在相同的训练集下用 libsvm 分类系统进行分类。分别设置分辨值阈值在以上所记录的分辨值范围内浮动,经过数次实验,选择分类结果最好的值作为阈值。得到的最好的分类结果为:82.1% 的准确率和 76.5% 的召回率。

以上测试及结果表明,本文提出的特征提取算法的分类准确率提高了,算法是有效的。

3.2　不同测试条件下的分类结果及性能分析

1)封闭测试与开放测试比较

利用每类均为 200 篇,10 类共 2 000 篇训练集文档和每类均为 100 篇,10 类共 1 000 篇测试集文档,对系统分别进行封闭性测试和开放性测试,表 1 为测试结果。

从表中可以看出,封闭性测试结果的平均查全率和平均精度分别达到了 82.2 % 和 85.7%,说明特征提取算法较好地抽取出了训练集的模式与特征。在对测试集文档的开放性测试结果中,平均查全率和平均精度也达到了 77.9% 和 80.8 %,与封闭性测试的结果较为相近,说明所抽取出的文档类的模式和特征具有普遍性和有效性。

2)不同分辨值阈值比较

本实验将设置不同分辨值阈值,从分类结果对分辨值阈值进行分析。一个好的索引词,其分辨力应该提高同类文档间的相似性,而降低不同类文档之间的相似性。实验时,将所有不同类别的文档集混合在一起进行计算,对全部类别同时来抽取特征词。如果一个索引词使这些不同类文档集的整体相似性变小,这个词就应该保留,否则应该剔除。如前所述,使用公式(1) ~(3)计算词的分辨值 $DISC$ 。

在训练文本集为每类 200 篇,10 类总共 2 000 篇文本的情况下,根据经验我们选择最大分辨值为 $DISC$ = 50 000, 150 000,500 000,1 000 000 进行对比实验。实验的分类结果如表 2 和图 2 所示,值为宏平均值。

表 1 2000 篇文档的封闭测试与开放测试结果

类别名称	正确识别文档数		误判文档数		R(%)		P(%)	
	封闭	开放	封闭	开放	封闭	开放	封闭	开放
rec. sport. baseball	71	67	17	15	71	67	81	80
alt. atheism	77	68	8	12	77	68	90	85
sci. crypt	79	69	15	15	79	69	84	80
comp. os. ms-windows. misc	74	68	16	14	74	68	82	82
rec. autos	94	92	14	18	94	92	87	84
rec. motorcycles	89	77	12	22	89	77	88	78
comp. graphics	88	79	17	18	88	79	84	80
comp. windows. x	95	91	12	16	95	91	89	85
comp. sys. ibm. pc. hardware	69	64	13	19	69	64	84	77
rec. sport. hockey	86	77	12	21	86	77	88	77
宏平均	/	/	/	/	82.2	77.9	85.7	80.8

表 2 不同分辨值阈值下的分类结果

分辨值 *DISC*	关键词个数	R(%)	P(%)	F1(%)
50000	552	73.39	74.11	73.54
150000	803	79.78	79.42	79.60
500000	1458	61.08	70.25	65.34
1000000	2401	61.17	60.29	60.73

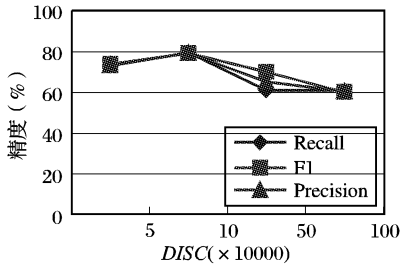

图 2 不同分辨值阈值下的分类结果

分类结果显示，*DISC* = 150000 的 *F*1 值以及召回率、正确率最好。分类结果显示 *DISC* 值小些的时候，结果好一些。但如果选择的分辨值绝对值过小，如 *DISC* = 50000，那么词的选择范围将太小，一些对分类很有意义的关键词被剔除掉了，导致很多文本不能识别出来，召回率变差。在 *DISC* = 500000 和 *DISC* = 1000000 时，分类结果比较差，而且后者更差一些。图 2 精度曲线呈下降趋势。这说明分辨值阈值比较大的时候，保留了过多的对分类无用的关键词，反而对分类造成了干扰。所以通过实验，我们选择了最大分辨值 *DISC* = 150000 作为阈值。当然，针对不同的训练样本数，应该设置不同的阈值。

3)不同分类方法的比较

获得了较优化的分辨值后，在训练文本集为每类 200 篇，10 类总共 2000 篇文本，测试集为每类 100 篇，10 类总共1000 篇文本的情况下，在本分类系统与 libsvm 分类系统比较了传统的特征选择方法与本特征选择方法的分类效果。

对 libsvm 分类系统和本分类系统，在同样的训练集和测试集上，分别使用 DF 方法和特征选择方法进行了训练和测试。DF 方法选择提取词频位于 25% ~75% 范围内的词作为关键词。

表 3 列出了在这些不同条件下的总的分类结果。从表中可以看出，不管是正确率还是召回率，不管是使用本文设计的分类系统还是 SVM 分类系统，本特征选择方法总是优于 DF 方法，这说明本特征选择方法确实优于 DF 选择方法，能够提取出比 DF 方法更有效的关键词来。而且，不管是正确率还是召回率，本分类系统的性能总是高于朴素贝叶斯(Naïve Bayes, NB)分类器的性能。

表 3 在不同的分类系统下与传统特征选择方法的性能宏平均比较

特征选择方法	本分类系统			libsvm 分类系统		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
DF	73.39	63.43	68.41	80.21	77.97	79.18
本特征选择方法	81.78	70.45	75.69	83.03	79.17	81.05

4)不同训练文本数目的比较

为考查算法的稳定性和提高分类效果的能力，在不同的训练样本数目下进行了实验。表 4 是训练文本数分别为每类取相同的文本数(如每类 100 个，总共 1000 个)的 10 类文本的分类结果。

从表 4 可以看出，在条件 6000 个文本下，原始特征空间的维数是 13732，提取的关键词是 2558 个时，测试的结果最好，达到宏平均准确率 85.25%，召回率为 76.08%，F1 值为 82.05%。

通过实验看到，随着训练样本数目的增大，分类的效果变好。这是因为当训练语料库没有达到一定规模的时候，特征空间中必然存在相当数量的出现频数很低(比如低于三次)的特征。因为它们较低的出现频数，决定这些特征只属于少数的类别。而本特征选择方法是基于类别信息的统计方法，必定认为这些低频词具有强烈的类别信息，从而对它们有不同程度的倚重。但是经过仔细观察发现，这些低频词中只有很少的词确实带有较强的类别信息，大多数的词都是噪音词，不应该成为特征[7]。所以，训练样本数规模应该达到一定的规模才能提取合适的特征，获得好的分类效果。虽然如此，在本实验中，每类 200 个，仅 2000 个文本作为训练样本仍然获得了 63.42% 的精度，所以本特征选择方法的稳定性较好。

表 4 不同训练样本文中数目下的分类结果

训练样本数目	原始特征空间维数	关键词个数	R(%)	P(%)	F1(%)
1000	4094	803	53.39	62.16	57.64
2000	6304	1232	54.78	63.42	58.78
4000	9342	1821	70.54	73.72	73.09
6000	13732	2558	76.08	85.25	82.05
8000	17459	2808	78.11	78.97	77.95
9000	19520	3468	71.32	77.65	74.35

3.3 分类效率分析

(下转第 2678 页)

2 模糊关联规则的具体应用

某石油天然气公司控制关键设备的参数点有3000多个，后台采用实时数据库(infoplus)系统采集现场数据，每半分钟取一次数据，一年的数据以数百GB记，为数据挖掘提供了丰富的数据资源。针对于本文主题，可采集影响干气质量的参数进行处理以后，存入数据库中，训练相应参数形成模糊关联规则模型。

2.1 实际的生产设备参数

举一个例子说明算法的实现过程，采集8个不同时刻的干气丙烷含量、环境温度、原料丙烷含量和SHPCGP_CRY:Z2TI268的区间值，得到表1，按Fuzzy_ClustApriori算法步骤1进行处理，得到表2。

表1 干气丙烷含量和各参数的区间值

事务号	环境温度(℃)	原料丙烷含量(%)	SHPCGP_CRY:Z2TI269(℃)	干气丙烷含量(%)
1	[2.1,2.4]	[3.6,3.7]	[-77.6,-77.8]	[0.3,0.4]
2	[21.2,21.6]	[3.7,3.7]	[-68.4,-66.7]	[1.2,1.3]
3	[28.8,29.1]	[4.0,4.1]	[-80.1,-80.3]	[0.2,0.3]
4	[16.5,16.6]	[3.2,3.3]	[-77.8,-77.9]	[0.4,0.5]
5	[34.2,34.4]	[3.3,3.4]	[-60.5,-60.7]	[1.4,1.6]
6	[26.4,26.5]	[4.2,4.3]	[-64.6,-64.8]	[1.3,1.4]
7	[8.4,8.9]	[3.3,3.5]	[-81.4,-81.6]	[0.1,0.2]
8	[22.5,22.6]	[3.5,3.6]	[-79.4,-79.5]	[0.5,0.6]

表2 生产参数模糊集合表

事务号	环境温度			原料丙烷含量			SHPCGP_CRY:Z2TI269			丙烷收率	
	L	M	H	L	M	H	L	M	H	G	B
1	0.92	0.07	0.01	0.24	0.54	0.22	0.64	0.22	0.14	0.76	0.24
2	0.34	0.44	0.22	0.2	0.56	0.24	0.07	0.12	0.81	0.23	0.77
3	0.16	0.32	0.52	0.07	0.10	0.83	0.83	0.11	0.06	0.87	0.13
4	0.56	0.30	0.14	0.91	0.05	0.04	0.66	0.21	0.13	0.58	0.42
5	0.02	0.13	0.85	0.83	0.1	0.07	0.02	0.05	0.93	0.06	0.94
6	0.10	0.37	0.53	0.02	0.06	0.92	0.05	0.08	0.87	0.17	0.83
7	0.84	0.10	0.06	0.44	0.43	0.13	0.92	0.06	0.02	0.96	0.04
8	0.30	0.46	0.24	0.32	0.52	0.16	0.76	0.16	0.08	0.51	0.49
$w\_t_p^j$	3.24	2.19	2.57	3.03	2.36	2.61	3.95	1.01	3.04	4.14	3.86

2.2 具体实现

设最小支持度minsup为0.31，最小信任度为0.9，按Fuzzy_ClustApriori算法一步步做下去，根据频繁集得到模糊规则：如果环境温度低，则干气质量优，其支持度为0.37，可信度为0.92；如果SHPCGP_CRY：Z2TI269出口温度低，则干气质量优，其支持度为0.45，可信度为0.92；如果SHPCGP_CRY：Z2TI269出口温度高，则干气质量差，其支持度为0.37，可信度为0.97；如果环境温度低和SHPCGP_CRY：Z2TI269出口温度低，则干气质量优，其支持度为0.33，可信度为0.98。

3 结语

本文采用基于模糊聚类的模糊关联规则对某流程企业的大量历史数据进行分析，通过使用RFCM算法把生产参数聚成c个类，并给出模糊数，然后用模糊理论和Apriori算法相结合形成的Fuzzy_ClustApriori算法对模糊化的生产参数进行分析，求取了模糊频繁集，并得到了符合人类思维的模糊规则，为企业生产优化提供了理论依据。

参考文献：

[1] HATHAWAY RJ, DAVENPORT JW, BEZDEK JC. Relational Dual of the C2Means Algorithms[J]. Pattern Recognition, 1989, 22(2):205-212.

[2] HONG TP, KUO C-S. A fuzzy AprioriTid mining algorithm with reduced computational time[J]. Applied Soft Computing, 2004, 5:1-10.

[3] KUOK CM, FU A, WONG MH. Mining Fuzzy Association Rules in Database[A]. Proceedings of the ACM Sixth International Conference on Information and Knowledge Management[C], 1997. 10-14.

[4] 欧阳为民，郑诚，蔡庆生. 数据库中加权关联规则的发现[J]. 软件学报，2001,12(4):612-619.

[5] CHEN GQ, WEI Q. Fuzzy Association Rules and the Extended Mining Algorithm[J]. Information Sciences, 2002, 147:201-228.

[6] BRIN S, RASTOGI R. Mining Optimized Gain Rules for Numeric Attributes[J]. IEEE Transactions on Knowledge and Data Engineering, 2003,15(2):324-338.

[7] HUA Y-C, CHENA R-S, TZENG G-H. Discovering fuzzy association rules using fuzzy partition methods[J]. Knowledge-Based Systems, 2003,16: 137-147.

(上接第2660页)

以上的实验均获得了较好的分类精度，证明了本文提出的特征选择方法比传统的DF特征选择方法效果好。但从算法的计算规模考虑，假设训练文本集的文本数为m个，抽取的原始特征项为n维，则本文提出的特征选择算法的计算复杂度为$O(m^2n)$，而DF算法的计算复杂度则为$O(mn)$。

在训练和分类的时间上，我们将使用基于类别的特征选择方法的本分类系统与使用DF提取方法的Naïve Bayes分类器进行了比较。在训练样本为每类200篇，20类总共4000篇文本的大规模文本的情况下，使用Naïve Bayes分类器，训练时间是10min，而使用本分类系统，共需要18min的训练时间。但在分类时，本分类系统和Naïve Bayes分类器一样，4000篇文本仅需要3 min，就可以得到结果。也就是说，本分类系统完成一篇长度为几KB文档的分类只需0.05s。

由此可见，虽然本分类系统在训练时间上比较长，但在分类时速度是很快的。这与我们选用了朴素贝叶斯分类算法有关，而训练阶段可以是非在线进行的。所以，本特征选择方法仍然可以应用到在线文本分类器中或者一些非实时性的应用中，其分类效率也可以达到要求。

参考文献：

[1] 蒋伟贞，陶宏才. 文本分类中特征提取方法综述及一种新方法的提出[J]. 计算机应用，2003,23(增刊):104-105.

[2] SALTON G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer[M]. Addison Wesley Publishing, 1989.

[3] PORTER MF. An algorithm for suffix stripping[J]. Program, 1980,14(3):130-137.

[4] JOACHIMS T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorazation[A]. FISHER DH, ed. Proceedings of the 14th International Conference on Machine Learning[C]. San Francisco: Morgan Kaufmann Publishers, 1997. 143-151.

[5] http://www.ai.mit.edu/~jrennie/20_newsgroups/[EB/OL],2005-05.

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm[EB/OL], 2005-05.

[7] 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,2003,18(1).