一种基于朴素贝叶斯分类的特征选择方法*

余 芳[1]，姜云飞[2]
（1. 暨南大学计算机科学系，广东 广州 510632；
2. 中山大学软件研究所，广东 广州 510275）

摘 要：由于朴素贝叶斯文本分类中的独立假设前提，使得在特征选择步骤能否准确有效地选出能代表文本的特征显得尤为重要，而特征选择标准中的 MI 标准与 TFIDF 标准其优缺正好互补，因此在用朴素贝叶斯文本分类方法中的多项式模型实现了一个 web 页面分类系统——WEBCAT 的基础上，提出将 MI 标准与 TFIDF 标准结合进行特征选择。实验显示：用改进的方法可以更准确地选出能代表文本的特征，文本分类结果也比单独使用 TFIDF 标准或单独使用 MI 标准进行特征选择的分类结果更加精确。

关键词：朴素贝叶斯分类；特征选择；MI 标准；TFIDF 标准
中图分类号：TP391.1 文献标识码：A 文章编号：0529-6579（2004）05-0118-03

文本分类是文本处理领域的重要研究内容之一，其任务是：预先定义若干类，系统在学习各个类的训练文档的基础上，预测未知类文档的类别。对文本分类的研究能为文本提取、文本过滤及文档聚类等其它文本处理领域的研究提供参考依据[1]。

文本分类的经典方法有 k-NN 方法、神经网络方法、朴素贝叶斯方法、决策树方法以及较新的支持矢量机方法等。朴素贝叶斯方法应用统计理论进行文本分类，其实际分类效果丝毫不比其它方法逊色[2]。本论文的实验采用朴素贝叶斯方法中的多项式模型对 web 页面进行分类。

朴素贝叶斯方法中有一个独立假设前提——在给定目标值时属性值之间相互条件独立[3]。这个前提忽略了特征之间存在的条件依赖关系，因此使得特征选择步骤显得尤为重要。因为如果特征选择步骤不能找到那些能够标识一个类别的特征，则应用朴素贝叶斯的分类效率必然低下。因此，对于应用朴素贝叶斯方法进行的分类，有必要重点研究其特征选择步骤。本文在一个应用朴素贝叶斯方法进行分类的 web 页面分类系统——WEBCAT 中，将 MI 标准与 TFIDF 标准结合进行特征选择，比仅采用 TFIDF 标准或仅采用 MI 标准进行特征选择的朴素贝叶斯分类获得了更好的分类结果。

1 特征选择的标准

一个 web 页面经去除标签、分词、去除停用词、取词干等步骤预处理后形成一个特征集，参与特征选择。此时构成一个页面的特征集仍旧是很大的，在我们的实验中，一篇 Yahoo！新闻经预处理后形成一个具有 300～600 个特征的特征集，考虑到训练页面的数目，对于主要使用统计方法来进行分类的分类器来说，计算量是非常大的。为了减少计算量，但同时又尽量少丢失有用的信息，应该找出那些能最好地代表一个文档或一个类别的特征参与分类计算。实际上在特征的预处理过程中，“去除停用词”步骤就是这个原则的另一体现，所不同的是它找的是那些不能很好区分各类别的特征。特征选择步骤同样也遵循该原则。特征选择的常用标准有 DF、IG、MI、CHI 及最早的 TFIDF 等[1,4]，以下详细介绍 TFIDF 标准及 MI 标准。

1.1 TFIDF 标准 TFIDF 是使用最普遍的标准，它用特征的 TF×IDF 值来评估一个特征。TF 因子表示特征频度，如 TF（f_i，p_j）定义为特征 f_i 在页面 p_j 中出现的次数（$1\leqslant i\leqslant |V|$，$1\leqslant j\leqslant N$，$V$ 为所有训练页面中的特征构成的集合，N 为训练页面总数）。TF 值反映了特征相对于某个页面的重要程度，默认为出现的次数越多越重要。此外，有一些特征几乎出现在所有页面中，例如 Yahoo 新闻网页中‘Yahoo’几乎在所有页面中都出现。这类特征对分类显然没有帮助，应该从特征集中除去。由此引入 IDF 因子，表示逆文档频度，如 IDF（f_i）定义为 $\log\left\{\frac{N}{\mathrm{DF}(f_i)}\right\}$，DF（$f_i$）为含有特征 f_i 的页面数。IDF 因子的含义是：包含某特征的页面越多，则该特征越不重要。将 TF（f_i，p_j）与 IDF（f_i）的乘积标准化

* 收稿日期：2003－12－18
基金项目：国家自然科学基金资助项目（60173039）；暨南大学自然科学基金资助项目
作者简介：余芳（1976 年生），女，助教；E-mail：yufangh4@tom.com

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

后得到TFIDF的完整定义[5]：

$$\mathrm{TFIDF}(f_i, p_j) = \left(0.5 + 0.5\frac{\mathrm{TF}(f_i, p_j)}{\mathrm{MAX}(_{v=1}^{|V|}\mathrm{TF}(f_v, p_j))}\right)\log\left(\frac{N}{\mathrm{DF}(f_i)}\right) \quad (1)$$

1.2　互信息标准　由于TFIDF标准仅考虑了特征与文档（即页面）之间的关系，特征与类的关系并没有反映出来，因此一些新的标准，如IG、MI等用来反映类与特征之间的关系[4]。计算MI的方法如下：

$$\mathrm{MI}(f_i, c_k) = \log\frac{\Pr(c_k \wedge f_i)}{\Pr(f_i)\times\Pr(c_k)} \quad (2)$$

其中，$1 \leqslant k \leqslant C$，$C$ 为预先定义的类别的数量；$\Pr(c_k)$ 为类别 c_k 的概率，$\Pr(f_i)$ 为含有特征 f_i 的训练页面的概率，$\Pr(c_k \wedge f_i)$ 为属于类别 c_k 且含有特征 f_i 的训练页面的概率。由于给定训练页面集后，$\Pr(c_k)$ 的值固定，因此含有特征 f_i 的训练页面中属于类别 c_k 的页面越多，$\mathrm{MI}(f_i, c_k)$ 的值就越大（把式(2)稍加变形即可知道），这说明 f_i 越具备代表类 c_k 的能力。

2　使用2种标准结合进行特征选择

由于TFIDF标准中的TF因子及IDF因子都仅反映特征与页面间的关系，而MI标准则反映特征与类别间的关系，因此将这2个标准结合进行特征选择将肯定使得分类结果更加精确。在本论文的实验中采取经过2个步骤的特征选择。

首先是MI选择，此步骤将许多与类关系不是很紧密的特征从特征集中剔除。观察公式(2)可知，如果 f_i、c_k 相互独立的话，MI的值为0，这种情况说明特征 f_i 在各类别训练页面中的分布与类别无关。这样，我们可以想到，那些MI值超出“平均数”的特征就是与类关系更加紧密的特征。基于这种想法，在MI选择步骤将选中那些MI值大于0的特征。在第一步选择的基础上再进行TFIDF选择，该步骤的阈值选取一般基于经验。

3　实验与评测

3.1　本实验选取的数据集　本文选取了不同时期的Yahoo！Reuters新闻页面作为训练数据集。该数据集包括Business、Entertainment、Health、Science、Sports五大新闻类别，每一类取120个页面，共600条页面数据，按照2∶1的原则，各类的页面分为80个训练页面及40个测试页面。

本文采用查准率(Precision)及查全率(Recall)对实验结果进行评估，相关定义见文[6]。

3.2　朴素贝叶斯分类器　给定一个页面 p_j $(f_1, f_2, \cdots f_n)$，$f_1, f_2 \cdots f_n$ 为进行预处理后构成该页面的所有特征。我们要计算该页面被分到某个类别 c_k 的概率 $\Pr(c_k \mid f_1, f_2 \cdots f_n)$，应用一般概率公式得：

$$\Pr(c_k \mid f_1, f_2, \cdots f_n) = \Pr(c_k)\prod_{i=1}^{n}\frac{\Pr(f_i \mid c_k, f_1, f_2, \cdots f_{i-1})}{\Pr(f_i \mid f_2, \cdots f_{i-1})}$$

应用朴素贝叶斯分类的独立假设，该式简化为：

$$\Pr(c_k)\prod_{i=1}^{n}\frac{\Pr(f_i \mid c_k)}{\Pr(f_i)} \quad (3)$$

$\Pr(c_k)$、$\Pr(f_i)$ 如2.2节中的定义；$\Pr(f_i \mid c_k)$ 为类别 c_k 的训练页面中含有特征 f_i 的页面概率。(3)式的进一步计算在文[6, 7]中有详细的说明，本文采用其中的多项式模型进行分类。

3.3　实验结果及结论　以上3个分类实验中，各类别最终参与分类的特征数目都在1 500左右，为初始特征数的1/4。由表1—3可以看出，采用2种标准结合选择特征的分类结果其Precision及Recall值最好，仅采用MI标准的分类结果次之，仅采用TFIDF标准的分类结果最差。由于TFIDF标准不包含类别的信息，因此区分各类别页面的能力较差，仅使用它进行特征选择得到这样的分类结果是显而易见的。而对于仅采用MI标准进行分类的结果，可看到，表1的结果比表3的结果提高不是很大，但确实有提高。

表1　2种标准结合选择特征的分类结果

Tab.1　Categorizing result using the new method

	本属此类	分为此类	正确分类	查准率%	查全率%	错误分类详情
Business	40	44	40	90.9	100	H-2、Sc-2
Entertainment	40	40	40	100	100	
Health	40	40	37	92.5	92.5	Sc-3
Science	40	36	35	97.2	87.5	H-1
Sports	40	40	40	100	100	

表2　仅应用TFIDF标准选择的分类结果

Tab.2　Categorizing result using TFIDF only

	本属此类	分为此类	正确分类	查准率%	查全率%	错误分类详情
Business	40	46	36	78.3	90	E-2、H-6、Sc-2
Entertainment	40	39	36	92.3	90	Sc-1、H-2
Health	40	36	23	63.9	57.5	B-3、Sc-10
Science	40	36	26	72.2	65	B-1、E-1、H-8
Sports	40	43	40	93	100	E-1、H-1、Sc-1

分析实验数据得知：被结合标准选中却不被仅采用MI标准选中的特征数：Business为109，Entertainment为79，Health为108，Science为74，Sports为80。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 3 仅应用 MI 标准选择的分类结果

Tab. 3 Categorizing result using MI only

	本属此类	分为此类	正确分类	查准率 %	查全率 %	错误分类详情
Business	40	47	40	85.1	100	H-4, Sc-3
Entertainment	40	41	40	97.6	100	Sp-1
Health	40	39	36	92.3	90	Sc-3
Science	40	34	34	100	85	
Sports	40	39	39	100	97.5	

相对于各类最终参与分类的特征数目而言，这些特征只占小部分，因此提高确实不会像相对于仅采用 TFIDF 标准的实验那样大。但是对各类特征进行分析，以 Business 为例，有‘firm’、‘billion’、‘percentag’、‘bond’、‘earn’、‘incom’、‘shar’、‘market’、‘company’、‘func’等一些 business 类文章典型的特征；再如‘canc’（‘cancer’经预处理去掉后缀‘er’）、‘diseas’、‘medicar’、‘smok’、‘blood’、‘brain’、‘stress’、‘immun’、‘drug’等 Health 类文章典型的特征也未被仅采用 MI 标准的实验选中，但却被使用结合标准的实验选中。考察这些特征的 MI 值，因为它们仅在相应类的一部分页面中出现，因此 MI 值不高不低，不容易被仅采用 MI 标准的实验选中；但因为它们在页面中可能会被重复多次（TF 值较高，如‘diseas’在某个页面中的 TF 值为 20，‘billion’为 21），且专门性较强，集中出现于一部分页面当中（IDF 值较高），因此容易被 TFIDF 标准选中。这说明，2 个标准结合确实能筛选出更多的对分类有帮助的特征。综合以上数据及分析，两者结合进行特征选择是有意义的。

从表 1—3 中均可以看到，Science 和 Health 类别的页面分错的现象较多。这是因为 Science 类的页面经常会讨论有关医药方面的内容，因此它们的标识特征重叠较多。例如，经第一步的 MI 选择后，Science 类的特征与 Health 类的特征重叠的数量为 724 个，占 Science 类特征总数的 1/5；而与 Business、Entertainment 及 Sports 类的重复特征数分别为 373、252 及 238 个，仅占各类特征的 1/15 ~ 1/20。因为朴素贝叶斯分类是基于统计方法的，标识特征重叠越多，则类别之间相互区分的能力就越弱。因此，Science 类的页面被较多地错分为 Health 类，而相对于 Entertainment 及 Sports 类别，被错分为 Business 类别的情况又更多一些，这在数据中也得到了反映。

4 结 语

通过以上实验，证明了在基于朴素贝叶斯方法的文本分类中，用 MI 标准和 TFIDF 标准结合进行特征选择可以获得更积极的分类结果。下一步，作者将考虑把研究结果应用到中文页面分类、多类别分类中，并在停用词表的半自动化生成及不同特征类型的特征值加权等方面作更加深入的研究。

参考文献：

[1] LEWIS D D. Representation and learning in information retrieval[D]. Maassachusetts: Graduate School of the University of Maassachusetts, 1992.

[2] LEWIS D D, RINGUETTE M. A comparison of two learning algorithms for text categorization[M]. Proceedings of SDAIR—94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994: 81—93.

[3] MITCHELL T M. 机器学习[M]. 曾华军, 张银奎, 译. 北京: 机械工业出版社, 2003: 127.

[4] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[M]. Proceedings of ICML—97, 14th International Conference on Machine Learning, 1998.

[5] SALTON G, BUCKLEY C. Weighting approaches in automatic text retrieval [J]. Information Processing and Management, 1988, 24(5): 513—523.

[6] McCALLUM A, NIGAM K. A comparison of event models for Naive Bayes text classification[M]. Proceedings of AAAI—98 Workshop on Learning for Text Categorization, 1998.

[7] CRAVEN M, DiPASQUO D, FREITAG D, et al. Learning to extract symbolic knowledge from the World Wide Web[M]. Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98), 1998: 509—516.

A Feature Selection Method for NB-based Classifier

YU Fang[1], *JIANG Yun-fei*[2]

(1. Department of Computer Science, Jinan University, Guangzhou 510632, China;
2. Software Institute, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: An effective feature selection is very important for an NB-based classifier which uses the conditional independence assumption. MI and TFIDF are two general feature selection methods in text categorization. Considering their each deficiency in representing documents or categories, they are combined into a two-stage selection method and applied to WEBCAT——a web page NB-based classifier using the multinomial model. Experiments show that this new method works more effectively than that only using MI or TFIDF in selecting those representative features and in categorizing.

Key words: Naive Bayes classifier; feature selection; MI; TFIDF

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net