面向不平衡文本的特征选择方法

廖一星[1,2]，潘雪增[1]

(1. 浙江大学计算机科学与技术学院 杭州 310012; 2. 浙江财经学院东方学院 杭州 310018)

【摘要】在分析了传统特征选择方法构造的4项基本信息元素的基础上提出一种强类别信息的度量标准，并在此基础上，提出一种适用于不平衡文本的特征选择方法。该方法综合考虑了类别信息因子、词频因子，分别用于提高少数类和多数类类别分类精度。该方法在reuter-21578数据集上进行了实验，实验结果表明，该特征选择方法比IG、CHI方法都更好，不但微平均指标有一定程度的提高，而且宏平均指标也有一定程度的提高。

关 键 词 特征选择方法; 不平衡数据集; 强类别相关; 文本分类

中图分类号 TP181 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.04.022

Feature Selection Method on Imbalanced Text

LIAO Yi-xing[1,2] and PAN Xue-zeng

(1. College of Computer Science and Technology, Zhejiang University Hangzhou 310012;
2. Department of Information, Zhejiang University of Finance & Economics Hangzhou 310018)

Abstract After analyzing the four basic information elements of traditional feature selection methods, a new measurement of strong class information is introduced and a new feature selection method is proposed for imbalanced text classification. The strong class information and the frequency of terms are used to improve the classification performance of minority classes and majority classes respectively. The experiments on reuter-21578 dataset show that the proposed method is better than IG and CHI. Both Micro F_1 and Macro F_1 are improved to some degree.

Key words feature selection; imbalanced dataset; strong class-related; text classification

随着机器学习、信息检索从发展到成熟，数据集的不平衡问题成为了一个新的重要挑战。文本分类是信息检索与数据挖掘领域的研究热点与核心技术，也面临着数据集分布偏斜的问题，即类别间样本的数量可能存在数量级的差距。这种偏斜的样本往往无法准确反映整个空间的数据分布，使分类器容易被大类淹没而忽略小类，导致了分类效果不太理想。

文本自动分类算法的一个很大的挑战是高维的特征，特征通常高达数万维或几十万维[1]，因此，筛选出有效的维数较低的特征子集对于提高分类精度和效率具有重要的意义。常用的特征选择方法有信息增益、互信息和X2统计等。然而在处理不平衡数据集时，这些常用的特征选择方法通常会倾向于选择出对多数类有利的特征子集，从而使在少数类类别上的分类效果很差。

文献[2]提出了一种面向不平衡数据集的特征权重方法。本文借鉴了文献[2]的部分思想，在分析了传统特征选择方法构造的4项基本信息元素的基础上，提出一种强类别信息的度量标准，由此提出一种适用于不平衡文本的特征选择方法。该方法综合考虑了类别信息因子、词频因子，分别用于提高少数类和多数类类别分类精度，并将该方法应用在retuter-21578数据集上。实验结果表明，该特征选择方法可以保持大类类别分类性能基本保持不变甚至提高的情况下，还能继续提高少数类类别的分类性能。

1 相关工作

不平衡样本分类器容易误判为大类的原因为传统的特征选择方法选择出的特征子集容易偏向于大类。目前采用特征选择方法改进不平衡文本的相关研究相对较少。文献[3]针对极不均衡的数据集问题，

收稿日期：2010 – 10 – 15; 修回日期：2011 – 06 – 29
基金项目：浙江省科技计划项目(2010C31003)
作者简介：廖一星(1977 –)，女，在职博士生，主要从事数据挖掘方面的研究.

分析和比较了信息增益、期望交叉熵、文本证据权及优势率等方法，结合贝叶斯分类器，发现二元优势率是最好的选择方法，而倾向高频的IG效果相对较差。文献[4]提出了CTD方法，结合文档频率信息和ICF类别信息进行特征选择。文献[5]提出SCTW方法，选择带有强类别信息的词条，这两种方法都得到了较高的分类性能。文献[6]以基于事例学习的框架为基础，提出了一种与测试样本相关的动态特征加权方法。文献[7]提出一个多策略的方法，多个学习方法并行，每一个学习方法使用不同的进化技术得到自己的特征选择，最后进行综合处理。文献[8]针对文本分类中存在的不平衡分类问题，按照一个经验性的样本比例，挑选正负两个样本集，分别从中选择最能表示该类样本的特征集，然后将这些特征集合并作为最后挑选的特征。对不同规模的特征集进行特征挑选的仿真实验表明，该特征挑选方法能有效地提高文本分类的F_1测度。文献[9]对反例进行了分析，通过实验发现将反例从特征中去掉会降低分类的性能，所以反例在高性能分类中也是必要的。文献[10]提出一种形如DFICF的特征选择方法，最后在reuter数据集上做了实验，结果表明DFICF方法比IG方法的效果更好。

2　文本向量空间模型

本文采用文献[12]的向量空间模型VSM，基本思想是词袋法，每个特征词对应特征空间的一维，将文本表示成欧氏空间的一个向量。如文本d_j表示为$V(d_j)=(w_{j1},w_{j2},\cdots,w_{jn})$，其中，$w_{jk}$表示文档$d_j$的权重，一般采用TF-IDF公式。TF-IDF公式主要考虑词频和反文档频率两个因素，并最后进行归一化，则有：

$$w_{jk}=\frac{\mathrm{tf}(T_{jk})\mathrm{idf}(T_k)}{\sqrt{\sum_{k=1}^{t}\mathrm{tf}(T_{jk})^2\mathrm{idf}(T_k)^2}} \tag{1}$$

$$\mathrm{idf}(T_k)=\ln(N/n_k+0.1) \tag{2}$$

式中，T_{jk}表示第j个文档的第k个特征；N表示训练集的总文档数；n_k表示包含第k个特征的文档数。

3　4项基本信息元素

通过对传统的特征选择方法进行分析，可以发现传统的特征选择方法通常是由4项基本信息元素组成的，如表1所示。

表1中，A表示c_i中出现t_k的文档数；B表示不属于c_i类别中出现t_k的文档数；C表示c_i中不出现t_k的文档数，D表示不属于c_i类别中不出现t_k的文档数；N表示总的文档数，即$A+B+C+D$。

表1　文本分类特征选择方法的4项基本信息元素

	c_i	$\overline{c_i}$
t_k	A	B
$\overline{t_k}$	C	D

特征选择方法中参数的含义如下：$p(t_k)$表示特征词t_k在文本集中出现的概率；$p(\overline{t_k})$表示特征词t_k在文本集中不出现的概率；$p(c_i)$表示类别c_i的概率；$p(c_i,t_k)$表示类别c_i出现特征词t_k的概率；$p(c_i,\overline{t_k})$表示类别c_i不出现特征词t_k的概率；$p(\overline{t_k},\overline{c_i})$表示非$c_i$类别中不出现特征词$t_k$的概率；$p(t_k,\overline{c_i})$表示非$c_i$类别中出现特征词$t_k$的概率。

下面对信息增益、X2统计、互信息进行分析，并将这些特征选择方法用表1的4种基本信息元素进行表示。

3.1　信息增益

信息增益的基本思想是：根据特征在文本中出现与否两种情况下计算条件概率，以确定该特征提供信息量的大小进行选取特征值，信息量大的特征选取。

$$\mathrm{IG}(t_k,c_i)=p(t_k,c_i)\ln\frac{p(t_k,c_i)}{p(t_k)p(c_i)}+$$

$$p(\overline{t_k},c_i)\ln\frac{p(\overline{t_k},c_i)}{p(\overline{t_k})p(c_i)}=-$$

$$\frac{A+C}{N}\ln\frac{A+C}{N}+\frac{A}{N}\ln\frac{A}{A+B}+\frac{C}{N}\ln\frac{C}{C+D} \tag{3a}$$

$$\mathrm{IG}(t_k)=\sum_i\mathrm{IG}(t_k,c_i) \tag{3b}$$

3.2　X2统计

X2方法认为单词t_k与文本类别c_t之间的非独立关系类似于具有一维自由度的X2分布。词条t_k对于类别c_t的X2统计量为：

$$\mathrm{CHI}(t_k,c_i)=\frac{N[p(t_k,c_i)p(\overline{t_k},\overline{c_i})-p(t_k,\overline{c_i})p(\overline{t_k},c_i)]^2}{p(t_k)p(\overline{t_k})p(c_i)p(\overline{c_i})}=$$

$$\frac{(AD-CB)^2\times N}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{4a}$$

$$\mathrm{CHI}(t_k)=\sum_i p(c_i)\mathrm{CHI}(t_k,c_i) \tag{4b}$$

X2统计量的值越高，说明词条与类别之间的独立性就越小；值越低，说明词条与类别之间的独立

性越大。

3.3 互信息

互信息(mutual information，MI)广泛应用于统计语言模型，对于类别 c_t 和词条 t_k，它们之间的互信息定义为：

$$\mathrm{MI}(t_k,c_i)=\ln\frac{p(t_k,c_i)}{p(t_k)p(c_i)}=\ln\frac{AN}{(A+B)\times(A+C)} \tag{5a}$$

$$\mathrm{MI}(t_k)=\sum_i p(c_i)\mathrm{MI}(t_k,c_i) \tag{5b}$$

从式(3)～式(5)可以看出，传统的特征选择方法都可以由表1的4项基本信息元素表示。在后面提出的特征选择方法中，将采用这4项基本信息元素来提取强类别相关的特征。

4 新的特征选择方法

对于不平衡文档，传统的特征选择方法倾向于挑选出有利于多数类的特征，有利于少数类的特征较少甚至为零，从而使少数类的分类精度不高。为此，应该在保证多数类特征的基础上尽量挑选出有利于少数类特征。

由于少数类包含的文档数少，从而特征的DF值也较小，如果特征选择方法倾向于选择高频特征，就不能选择出少数类的特征。而强类别特征选择方法与特征的DF值无关，仅与类别强相关，故不仅可以挑选出多数类的强类别相关特征，也可以挑选出较多有利于少数类的特征。这样，类别之间的区分更加明显，越有利于类别的划分。文献[10-11]指出选择有较强类别信息的特征是提高稀有类别分类性能的关键。因此，本文提出一种强类别相关的特征选择方法，用于提取有利于少数类的特征。首先，采用最直接的方法，即利用表1的4项基本信息元素来表示某个特征 t_k 与某个类别 c_i 的相关程度，及特征 t_k 与非类别 c_i 的相关程度。特征 t_k 与某个类别 c_i 的相关程度越大，与非类别 c_i 的相关程度越小，那么特征对类别 c_i 的区分能力较好；反之，区分能力较差。下面是几个式子的含义：

1) A/B：A/B越大，表示类别 c_i 中出现特征 t_k 的文档数越多，非类别 c_i 中出现特征 t_k 的文档数越少，特征 t_k 与类别 c_i 的相关程度越大；反之相反。

2) A/C：A/C越大，表示类别 c_i 中出现特征 t_k 的文档数越多，不出现特征 t_k 的文档数越少，特征 t_k 与类别 c_i 的相关程度越大；反之相反。

3) B/A：B/A越大，表示非类别 c_i 中出现特征 t_k 的文档数越多，类别 c_i 中出现特征 t_k 的文档数越少，特征 t_k 与非类别 c_i 的相关程度越大；反之相反。

4) B/D：B/D越大，表示非类别 c_i 中出现特征 t_k 的文档数越多，不出现特征 t_k 的文档数越少，特征 t_k 与非类别 c_i 的相关程度越大；反之相反。

因此，特征 t_k 与某个类别 c_i 的相关程度可以用A/B和A/C表示，特征 t_k 与非类别 c_i 的相关程度可以由B/A和B/D表示。特征 t_k 对类别 c_i 的区分能力为：

$$\mathrm{dis}(t_k,c_i)=A^2/(BC)-B^2/(AD)=\frac{p(c_i,t_k)^2}{p(\overline{c_i},t_k)p(c_i,\overline{t_k})}-\frac{p(\overline{c_i},t_k)^2}{p(c_i,t_k)p(\overline{c_i},\overline{t_k})} \tag{6}$$

为了找出最强类别相关特征，特征t_k的类别区分能力为：

$$\mathrm{distinction}(t_k)=\max(\mathrm{dis}(t_k,c_i)) \tag{7}$$

另外，考虑到特征t_k的词频对大类类别的分类精度的影响，因此特征t_k的衡量标准为：

$$\mathrm{def}(t_k)=p(t_k)\mathrm{distinction}(t_k) \tag{8}$$

5 实验及结果

本文采用reuter-21578数据集，并从中挑选10个类别：acq，corn，crude，earn，grain，interest，money-fx，ship，trade，wheat。训练数据共7 165篇，测试数据共2 779篇。该数据集的分布是非均匀的，训练样本集分布如图1所示。

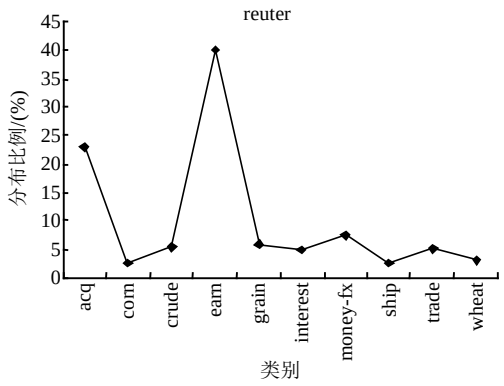

图1 分布比例

由于微平均F_1将所有文档一块儿计算，然后平均，指标倾向于大类；而宏平均对每个类求值，然后平均，指标倾向于小类。因此本文对文本分类评价指标采用微平均F_1和宏平均F_1，既可以看到该特征选择方法对大类的影响，也可以看到对小类的影响。

支持向量机是性能较好的分类器，也是文本分类中常用的分类器，因此分类器采用支持向量机，支持向量机软件采用libsvm[15]软件。

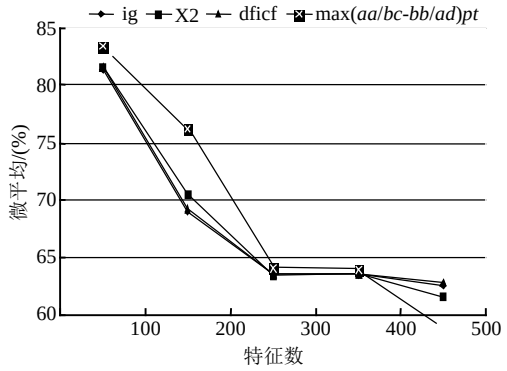

图2　微平均F_1比较

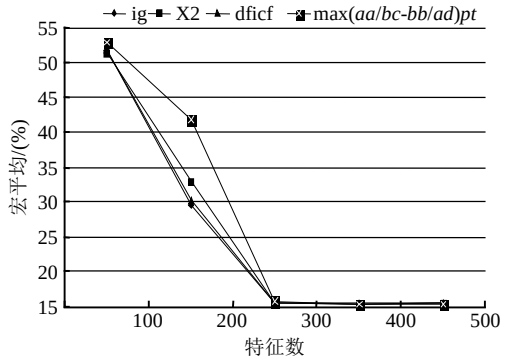

图3　宏平均F_1比较

表1　100个特征时各种方法性能比较

F_1	IG/(%)	CHI/(%)	DFICF/(%)	新方法/(%)
微平均F_1	81.36	81.58	81.58	83.48
宏平均F_1	51.71	51.38	51.78	52.93

图2和图3分别显示了微平均F_1和宏平均F_1采用不同特征选择方法，并在不同特征数目下的性能比较。由图可以看出，无论是宏平均F_1还是微平均F_1指标，都比IG、X2和DFICF有较大程度的提高；在特征数为200时，提高的幅度最大；在特征数为100时分类效果最好，此时新方法与其他特征选择方法的性能比较见表1。可以看出新的特征选择方法的微平均F_1分别比IG、X2和DFICF方法提高2.12%、1.91%和1.91%，宏平均F_1分别比IG、X2和DFICF方法提高1.21%、1.55%和1.14%。说明新的特征选择方法不仅可以提高大类的分类性能，也可以提高小类的分类性能，即新的特征选择方法对不平衡数据集具有较好的处理能力。

6　结　论

本文提出了一种面向不平衡文本分类的特征选择方法。实验结果表明，该方法可以在较大程度上提高文本分类效果，无论是少数类还是多数类的分类性能都得到了一定程度的提高。但是从图中可以看到，宏平均指标总体上还是较低，说明少数类类别的分类性能还不是很好。因此，下一步的研究方向是如何进一步提高少数类类别样本的分类精度，可以考虑采用特征权重方法进一步加强少数类特征的权重。

参　考　文　献

[1] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proceedings of ICML. San Francisco, USA: Morgan Kaufmann, 1997.

[2] LIU Ying, HAN TONG LOH, AIXIN SUN. Imbalanced text classification: A term weighting approach[J]. Expert Systems with Application, 2009, 36(1): 690-701.

[3] MLADENIC D, GROBELNK M. Feature selection for unbalanced class distribution and naïve bayes[C]//Proc of the 16th International Conf Machine Learning. San Francisco, USA: Morgan Kaufmann, 1999.

[4] BONG C H, NARAYANAN K. An empirical study of feature selection for text categorization based on term weight[C]//IEEE International Conference on Web Intelligence. Washington D C, USA: IEEE Computer Society, 2004.

[5] LI Shou-shan, ZONG Cheng-qing. A new approach to feature selection for text categorization[C]//IEEE International Conference on Natural Language Processing and Knowledge Engineering(NLP-KE). Washington D C, USA: IEEE Computer Society, 2005.

[6] CARDIE C, NOWE N. Improving minority class predicting using case-specific feature weights[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaumann, 1997.

[7] CASTILLO M D D, SERRANO J I. A multi-strategy approach for digital text categorization from imbalanced documents.[J] SIGKDD Explorations, Newsletter, 2004, 6(1): 70-79.

[8] ZHENG Z H. Optimally combining positive and negative features for text categorization [C]// ICML2003. Washington, D C, USA: AAAI Press, 2005.

[9] FORMAN G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research, 2003, 3(1): 1289-1305.

[10] 徐燕, 李锦涛, 王斌, 等. 不均衡数据集上文本分类的特征选择研究[J]. 计算机研究与发展, 2007, z2: 58-62.
XU yan, LI Jin-tao, WANG Bin, et al. A study of feature selection for text categorization on imbalanced data[J]. Journal of Computer Research and Development, 2007, z2: 58-62.

[11] 靖红芳, 王斌, 杨雅辉, 等. 基于类别分布的特征选择框架[J]. 计算机研究与发展, 2009, 46(9): 1586-1593.
JING Hong-fang, WANG bin, YANG Ya-hui, et al. Category distribution-based feature selection framework[J]. Journal of Computer Research and Development, 2009, 46(9): 1586-1593.

[12] ROCCHIO J. The smart retrieval system-experiments in automatic, document processing[M]. Englewood Clifts NJ, USA: Prentice-Hall Inc, 1971.

编　辑　黄　莘