基于信息增益的文本特征选择方法

任永功 杨荣杰 尹明飞 马名威

(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 在类和特征分布不均时,传统信息增益算法的分类性能急剧下降。针对此不足,提出一种基于信息增益的文本特征选择方法(TDpIG)。首先对数据集按类进行特征选择,以减少数据集不平衡性对特征选取的影响。其次运用特征出现概率计算信息增益权值,以降低低频词对特征选择的干扰。最后使用离散度分析特征在每类中的信息增益值,过滤掉高频词中的相对冗余特征,并对选取的特征应用信息增益差值做进一步细化,获取均匀精确的特征子集。通过对比实验表明,选取的特征具有更好的分类性能。

关键词 特征选择,文本分类,信息增益值,冗余特征,不平衡数据集

中图法分类号 TP301.6 **文献标识码** A

Information-gain-based Text Feature Selection Method

REN Yong-gong YANG Rong-jie YIN Ming-fei MA Ming-wei

(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract Due to the maldistribution of class and feature, the classification performance of traditional information gain algorithm will decrease sharply. Considering that, a text feature selection method TDpIG based on the information gain was proposed. First of all, selected feature in dataset based on the class, which can reduce the effect of dataset imbalance on feature selection. Secondly, calculated information gain weight by using feature occurrence probability to decrease the interference of low frequency words to feature selection. At last, analysed the increasing information of each class by use of dispersion, filtering out the relative redundant features of high frequency words, further refining the selected feature applied increasing information, and getting the uniform and accurate subsets. The comparison experiment shows that the method has better classification performance.

Keywords Feature selection, Text classification, Information gain, Redundant feature, Imbalanced dataset

1 引言

随着 Web 的迅猛发展,网络信息迅速增加,文本分类成为处理和组织大量文档数据的关键技术,但其高维特征空间不仅增加了分类的时间复杂度和空间复杂度,还影响分类精度。特征选择通过降低特征空间维度以及去除噪音特征来提高分类效率及精度。常见特征选择方法有交互信息(MutualInformation, MI)、信息增益(Information Gain, IG)、X2 统计量(Chi-square, CHI)、特征权(Term Strength, TS)、期望交叉熵(ExpectedCrossEntropy, ECE)、文本证据权(WeightofEvidence, WE)、几率比(Odds Ratio, OR)等。这些方法从不同的角度度量特征对分类的重要性。

特征选择 IG 是一种有效的特征选择方法,如文献[1]提出 IG 是最好的测度之一;文献[2]比较了 DF (Document Frequency,文档频率)、MI、IG、CHI 及 TS 5 种特征选择方法,其中以 CHI 效果最好,DF、IG 和 CHI 之间存在很大的相关性;文献[3]比较了 IG、DF、odds ratio 3 种特征选择算法,表明 IG 能提取更优的特征子集。IG 在不降低文本分类性能的前提下移走高达 98%的"无用"单词[1],但是此算法考虑到全局变量,在处理不均衡数据时性能急剧下降,并缺少对选取特征的进一步筛选。因此为提高 IG 算法的性能,本文不仅考虑不平衡数据集以及低频词对特征选择的影响,还去除高频冗余特征来降低特征维度,选择区分类别强的特征子集,使分类效率和精度得到明显提高。

2 信息增益简介

2.1 信息增益

定义 1(信息增益, Information Gain, IG) 是某一特征在文本中出现前后的信息熵之差。信息增益考虑特征出现与不出现时,特征对文本类别的信息表示量。以信息量的多少作为特征的权值,进而筛选特征。

$$IG(W)=-\sum_{t}P(C_t)\log P(C_t)+P(W)\sum_{t}P(C_t/W)\log P(C_t/W)+P(\overline{W})\sum_{t}P(C_t/\overline{W})\log P(C_t/\overline{W})$$

定义 1 中,t 表示类别总数,$P(W)$表示特征 W 在文本中出现的概率,$P(C_t)$表示 C_t 类文本在文本集中出现的概率,

到稿日期:2011-01-13 返修日期:2012-06-22 本文受国家自然科学基金项目(60603047),教育部留学回国人员科研启动基金资助项目,辽宁省科技计划项目(2008216014),辽宁省教育厅高等学校科研基金(L2010229),大连市优秀青年科技人才基金(2008J23JH026)资助。

任永功(1972—),男,博士,教授,主要研究方向为数据挖掘、图像处理技术等,E-mail:renyg@dl.cn;**杨荣杰**(1983—),女,硕士生,主要研究方向为数据挖掘;**尹明飞**(1987—),女,硕士生,主要研究方向为数据挖掘;**马名威**(1986—),男,硕士生,主要研究方向为数据挖掘与并行计算。

$P(C_i/W)$表示文本包含特征 W 时属于 C_i 类的条件概率，$P(\overline{W})$表示文本中不包含特征 W 的文本的概率，$P(C_i/\overline{W})$表示文本不包含特征 W 时属于 C_i 类的条件概率。

根据 Yiming Yang 教授[1]对英文文本分类中特征选择算法的深入研究，可得 IG 特征选择公式。

$$IG(W)=P(W)\sum_i P(C_i/W)\log\frac{P(C_i/W)}{P(C_i)}+P(\overline{W})\sum_i P(C_i/\overline{W})\log\frac{P(C_i/\overline{W})}{P(C_i)}$$

在信息增益中，分类系统负载信息量的大小是衡量特征重要性的标准。负载量越大，特征越重要。因此选择 IG 值大的特征构成分类特征子集来提高系统的分类效率。

2.2 信息增益的不足

虽然 IG 算法是有效的全局特征选择算法，但针对不均衡数据集，IG 算法对小样本集抽取概率降低，减少了小样本集的特征提取概率。它考虑特征出现与不出现两种情况，对于小数据集，不出现的特征权值将产生主导作用，因此很难提取小样本集特征。其次 IG 算法倾向于提取高频特征，忽略了提取特征间的相关性，缺少对特征子集的进一步筛选。

鉴于现实生活中不均衡数据集的普遍存在，为广泛应用 IG 算法，需进一步优化 IG 算法。针对 IG 算法只适用于全局变量的缺陷，文献[4-6]分别采用不同方法对 IG 等特征选择算法进行改进，提出适用于不平衡数据集的新算法。但是文献[4]中当类的重叠度较小且存在类不平衡时，此算法精度会下降，且适用范围小，没有普遍性；文献[5]通过 IG、CHI、CC 以及 OR 4 种特征选择的组合，得到适应不平衡数据集的特征选择方法，但是它的限制条件很多，适应性差；文献[6]用 $P(\overline{W}$替换来降低不平衡数据集中低频词对选取特征性能的影响。文献[6]虽然降低了数据集不平衡时低频词对选取特征的影响，但是没从根本上降低不平衡数据集的影响。文献[7]利用词频信息提高 IG、DF 及 IM 的性能，但对提取特征仍需进一步处理进行降维，以提高分类的性能。

基于传统 IG 算法以及文献[4-7]算法的不足，本文从三方面对 IG 算法进行改进，提出一种改进的信息增益特征选择方法。一方面对数据集进行分类特征选择，减少不平衡数据集对特征选取的影响。另一方面使用特征出现概率计算信息增益权值，以降低低频词对特征选择的干扰。最后利用离散度分析在每类中特征的信息增益值，对提取特征做进一步细化，过滤掉高频词中的冗余特征，获取均匀精确的特征子集。

3 基于信息增益的改进算法

3.1 改善不均衡数据集的影响

信息增益只考察特征对整个系统的贡献，而不能具体到某个类，因此它只适用于做“全局”的特征选择，而无法做“本地”的特征选择。例如有 C_1，C_2，C_3，C_4 4 类，前 3 类都含有 5 个样本，C_4 含有 1 个样本。有 3 个特征项：w_1，w_2，w_3。1 表示特征出现，0 表示特征不出现，如表 1 所列。

表 1 特征文档分布情况

特征项	C_1					C_2					C_3					C_4
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1
w_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
w_2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
w_3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1

由定义 1 得：w_1 的值为 0.187，w_2 的值为 0.524，w_3 的值为 0.206。从结果可知数据集平衡程度严重影响了特征的提取，w_1 相对于 w_3 具有更强的类区分性，但根据权重值筛选特征，w_1 更容易被过滤掉。

定义 2(IG_t 信息增益选择函数) 分别计算每类中特征的信息增益值，以避免数据集不平衡时信息增益选取特征覆盖不全的缺点。

$$IG_t(W)=P(W)P(C_t/W)\log\frac{P(C_t/W)}{P(C_t)}+P(\overline{W})P(C_t/\overline{W})\log\frac{P(C_t/\overline{W})}{P(C_t)}$$

定义 2 中 $P(W)$表示特征 W 在 C_t 类中出现的概率，$P(\overline{W})$表示特征 W 在 C_t 类中不出现的概率。由定义 2 得，在 C_1 类中 w_2 的值为 1.678，w_3 的值为 0.251。在 C_4 类中 w_1 的值为 4，w_3 的值为 0.805。因为分类选取贡献率大的特征，所以 C_4 中的 w_1、C_1 中的 w_2 都将被提取，以减少不平衡数据集对选择特征的影响。本文从每类中提取权重大的前 200 个特征组成分类特征子集，使其覆盖所有类别。

3.2 减少低频特征的影响

虽然降低了数据集不均衡对特征选择带来的影响，但是信息增益算法本身还存在不足。

信息增益特征选择公式考虑了特征出现和不出现两种情况，因而在去除“无用词”时，效果显著。但特征不出现时，对文本分类的贡献远小于对分类的干扰，特别在类和特征分布高度不均的情况下，低频词不出现的概率远大于出现的概率，即 $P(\overline{W})\gg P(W)$。此时信息增益值是由公式后半部分决定的，其大大降低了 IG 提取特征的效果。

假设有 C_1，C_2，C_3，C_4 4 类，C_1 含有 10 个样本，而后 3 类都只含有 1 个样本。有两个特征项：w_1，w_2。1 表示特征出现，0 表示特征不出现，如表 2 所列。

表 2 特征文档分布情况

特征项	C_1										C_2			C_3	C_4
	1	2	3	4	5	6	7	8	9	10	1	1	1	1	1
w_1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1
w_2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

由定义 1 可得：w_1 的值为 0.3，w_2 的值为 0.013。根据 C_1 中特征权值，w_1 比 w_2 具有更佳的分类效果。但由表 2 可知，w_2 对 C_1 的区分性能更好。因此 IG 公式应筛选出本类中出现次数多而在其它类中不出现的特征。

定义 3 改进信息增益 pIG 特征选择函数。pIG 考虑特征出现时，特征对文本类别的信息表示量。以信息量的多少作为特征的权值，进而筛选特征。

$$pIG(W)=P(W)\left(\sum_i P(C_i/W)\log\frac{P(C_i/W)}{P(C_i)}+\sum_i P(C_i/\overline{W})\log\frac{P(C_i/\overline{W})}{P(C_i)}\right)$$

由定义 3 得，w_1 的值为 -0.032，w_2 的值为 0.034 时，能对特征进行更好的筛选。不仅选取出对本类信息量贡献大的特征，而且在特征与类分布高度不均的情况下，减少了低频特征对特征权重的影响。

pIG 函数比传统 IG 提取特征的性能得到一定的提高，减少了低频词的影响，但还需要对提取的特征进行冗余处理，选

择更优的特征子集来进一步提高分类性能。

3.3 去除高频特征的冗余

在分类选取特征过程中，由于分别计算每个类中特征的pIG值，使得特征集合中出现重复特征，因此需先删除重复特征，然后对特征集合进行冗余处理。

信息增益值的波动大小可以衡量特征对于分类存在冗余的程度。信息增益权值波动越大，表明特征在每个类中的文本表示值差异越大，则特征值对应到每个类的中心值距离就越远，对文本分类就越精确。反之，则很难对文本进行分类。例如训练集中的特征“促进”在教育、政治、哲学和历史4个类中的pIG值分别为0.933、0.611、0.797、0.825。通过观察信息增益值发现，以平均值0.792为中心，每个类的信息增益值的波动很小。特征在4个类中的向量值集中在以平均值为中心的区域内，特征“促进”很难对测试集中的样本进行类别区分。因此高频特征“促进”属于冗余特征，需从特征子集中删除。

定义4（基于pIG的信息增益特征选择函数DpIG）DpIG使用离散度来计算特征在每类中pIG值的波动大小。离散度是衡量一组数据波动大小的重要量，通过数据方差计算。

$$DpIG_i=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}(w_{ij}-\overline{w_i})^2}{t}$$

定义4中n表示选取的特征总数，w_{ij}表示第i个特征在第j类的pIG取值，$\overline{w_i}$表示第i个特征在所有类中pIG值的平均值，t表示文本类别数。

特征选择最终筛选的是本类中出现概率大且其它类中不出现的特征。但特征离散度衡量的是特征在所有类中的整体波动性，并不能精确表示具体类与类之间的信息增益值变化。

定义5（基于pIG的信息增益特征选择函数TDpIG）TDpIG利用特征在类间的最大信息增益值与第二信息增益值的差值对DpIG选取的特征做进一步筛选，从而选择更精确的分类特征子集。

$$TDpIG_i=\mathrm{MAX}(pIG_i)-SEC(pIG_i)$$

定义5中$TDpIG_i$表示第i个特征的pIG的差值，MAX(pIG_i)表示第i个特征在类中的最大pIG信息增益值，SEC(pIG_i)表示第i个特征在类中的第二大pIG信息增益值。

$TDpIG_i$值越大，表明特征越集中出现在一类中，特征对此类的区分度越强，对该类的贡献率就越大，因此增加$TDpIG_i$限制函数对特征做进一步筛选。特征选择函数DpIG与TDpIG是对特征冗余的深层分析，能达到去除高频冗余特征、降低特征维度、提高分类性能的效果。

3.4 TDpIG算法描述

首先利用pIG算法对训练数据集按类进行特征权值计算，然后根据DpIG选择信息增益值波动大的特征，最后使用TDpIG对特征进行筛选。算法描述如下。

1）T从1到t循环：

①初始化每个特征的权重即$W[w]=0$；

②根据定义3，对类中所有特征$W_i(i=1,2,\cdots,w)$分别计算权值并更新权值：

$$pIG(W)=P(W)\left(P(C_t/W)\log\frac{P(C_t/W)}{P(C_t)}+P(C_t/\overline{W})\log\frac{P(C_t/\overline{W})}{P(C_t)}\right)$$

③在每类中选取权值最大的前200个特征。

循环结束。

2）将选取的所有特征删除重复特征后，放入特征子集S_{mid1}。

3）I从1到n循环

①根据定义4，对选取特征$W_i(i=1,2,3,\cdots,n)$计算离散度：

$$DpIG_i=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}(w_{ij}-\overline{w_i})^2}{t}$$

②将S_{mid1}特征子集中DpIG值>0.1的特征存入S_{mid}特征子集。

循环结束。

4）根据定义5，计算S_{mid}特征子集中特征的TDpIG值：

$$TDpIG_i=\mathrm{MAX}(pIG_i)-SEC(pIG_i)$$

5）将S_{mid}特征子集中TDpIG值>0.2的特征存入S_{goal}特征子集。

6）输出S_{goal}特征子集。

为降低不均衡数据集对特征选择的影响，分类计算特征权重，并通过DpIG和TDpIG对特征进行筛选，设置DpIG的阈值为0.1，TDpIG的阈值为0.2。不同的训练数据计算出的权重值不同，则设置的阈值也不同。阈值太小，对特征筛选没有意义；阈值太大将过分筛选，删除对文本分类重要的特征。本文经过多次对比实验，选择使实验结果最佳的阈值对特征进行筛选。最后使用选取的特征进行分类实验，以验证本文改进的特征选择算法的有效性。

4 实验结果及分析

4.1 性能评测

1）文本表示采用变形的Okapi[8]公式，如式(1)所示：

$$w(t,\vec{d})=\frac{tf(w,\vec{d})}{0.5+1.5*\frac{len(\vec{d})}{avg\_len}+tf(w,\vec{d}}*\log\left(\frac{0.5+N-n(w)}{0.5+n(w)}\right)\tag{1}$$

式中，$tf(w,\vec{d})$表示文档d中特征w的词频，$len(\vec{d})$表示文档d中的词数，avg_len表示平均每个文档中的词数，N表示集合文档中总的词数，$n(w)$表示出现过特征w的文档数。

2）经变形后的文本采用KNN分类器进行分类，KNN分类器中文本间距离采用Mandistance距离进行测量。文本分类的评测指标采用文本分类评测标准中的平均准确率、平均召回率和F1值。各评价参数定义如下：

1）平均准确率

分类的准确率=分类正确文本/分类的实际文本数

$$MacroP=\frac{1}{n}\sum_{j=1}^{n}P_j\tag{2}$$

式中，n为总的分类数，P_j为第j类的准确率。

2）平均召回率

分类的召回率=分类正确文本/分类应有的文本数

$$MacroR=\frac{1}{n}\sum_{j=1}^{n}R_j\tag{3}$$

式中，n 为总的分类数，R_j 为第 j 类的召回率。

3)平均 F1 值

$$MacroF1=\frac{MacroP\times MacroR\times 2}{MacroP+MacroR} \quad (4)$$

式中，$MacroP$ 是平均准确率，$MacroR$ 是平均召回率。

4.2 实验分析

实验数据选取了复旦大学中文语料库中的 2940 篇文本，包括文学、艺术、历史、政治、哲学、教育 6 个类别，其中 70%用来训练，30%用来测试。首先使用 ictclas50 分词包对所选文本进行分词处理；其次提取分词处理后文本中的名词、动词、形容词和量词，并去除其中的停用词和无用词；再次使用特征选择算法选取特征；最后，使用 Okapi 公式将文本表示成向量的形式，使用 KNN 文本分类器对测试样本进行分类。

训练集中对每类出现频率高的前 1000 个特征分别进行 IG、pIG、TDpIG 3 种特征权重计算，然后在每类中选取权重值高的前 200 个特征，使用 KNN[9] 分类器对测试集进行分类，对比 3 种情况下各自的分类精度，以验证本文特征提取算法的有效性。IG 特征选择算法与 pIG 特征选择算法的性能比较如图 1 所示。

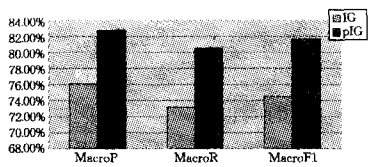

图 1 IG 特征选择算法与 pIG 特征选择算法的性能比较

由图 1 可知，pIG 特征选择算法相对于传统 IG 特征选择算法在 $MacroP$，$MacroR$，$MacroF1$ 都有很大的提高。这是由于 pIG 特征选择算法首先分类进行特征选择，改善了不平衡数据集对特征选择的影响，另外使用 $P(W)$ 代替 $P(\overline{W})$，减少了低频词对特征选择的影响，选取每个特征中出现的高频词。相对于传统的 IG 特征选择算法，其提取的特征对文本分类有更大的贡献率。

由图 2 可知，TDpIG 特征选择算法相对于 pIG 特征选择算法在 $MacroP$，$MacroR$，$MacroF1$ 都有一定的提高。这是因为 TDpIG 特征选择算法增加了对高频词的冗余处理，提取出本类中出现频率高、其余类出现频率低的特征，提取的特征对区分类有更高的贡献率，因此得到了更为理想的分类效果。

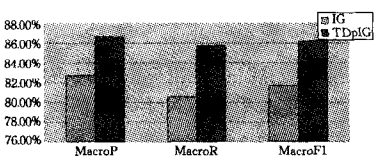

图 2 pIG 特征选择算法与 TDpIG 特征选择算法的性能比较

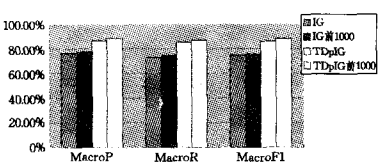

图 3 4 种方法的性能对比

本文使用 IG、TDpIG 两种算法对测试集概率高的前 1000 特征进行了特征值计算。图 3 是 IG 特征选择算法、IG 测试集前 1000 个特征、TDpIG 特征选择算法以及 TDpIG 测试集前 1000 个特征 4 种方法的性能对比。

从图 3 可知，对测试集的前 1000 个特征提取的分类性能略优于对测试集全部特征提取的性能。这是因为在本文实验中训练集只是选取了频率高的前 1000 个特征进行特征选取计算。使用测试集的前 1000 个特征进行特征数值计算，不仅可以减少特征数值计算的时间，还能提高一定的分类精度。

结束语 本文研究了传统信息增益算法，并针对其不足进行了改进。通过分类选择特征，降低不平衡数据集对特征选择的影响；用 $P(W)$ 代替 $P(\overline{W})$，减少了低频特征对分类的影响；使用 DpIG 以及 TDpIG 做判断条件，去除高频特征的冗余。实验结果表明，在类和特征分布不均时，经改进后的信息增益算法的分类性能有很大的提高。最后经对测试集全特征以及前 1000 个特征进行分类的对比实验，可以得到针对前 1000 个特征的分类性能比全局的略好，提高了 KNN 分类效率。

参 考 文 献

[1] Yang Yi-ming, Pedersen J O. A Comparative Study on feature selection in text categorization [C]// Proceedings of the 14th International Conference on Machine Learning(ICML'97). Nashvillr: Morgan Kaufmann Publishers, 1997: 412-420

[2] Ng H, Goh W, Low K. Feature selection, perceptron learning and a usability case study for text categorization [C]// Proceedings of the 20th ACM International Conference onResearch and Development in InformationRetrieval(SIGIR-97). 1997: 67-73

[3] Wang Bin, Jones G J F, Pan Wen-feng. Using online linear classifiers to filter spam emails[J]. Pattern Analysis & Applications, 2006, 9(4): 339-351

[4] 杨玉珍，刘培玉，朱振方，等. 应用特征项分布信息的信息增益改进方法研究[J]. 山东大学学报：理学版，2009(11)：48-51

[5] Zheng Zhao-hui, Wu Xiao-yun, Srihari R. Feature Selection for Text Categorization on Imbalanced Data[J]. ACM SIGKDD Explorations Newsletter, 2004(6): 80-89

[6] 单丽莉，刘秉权，孙承杰，等. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报，2011(3)：319-324

[7] Xu Yan, Chen Lin. Term-frequency Based Feature Selection Methods for Text Categorization[C]// Proceedings of the 2010 Fourth International Conference on Genetic and Evolutionary Computing. Dec. 2010: 280-283

[8] Robertson S E, Walker S, Jones S, et al. Okapi at trec-3[C]// Gaithersburg M D. Proceedings of the Third Text Retrieval Conference (TREC-3). USA: the National Inst. of Standards&Technology(NIST) &Defense Advanced Research Projects Agency(DARPA). 1994: 109-126

[9] Hu Qing-hua, Yu Da-ren, Xie Zong-xia. Neighborhood classifiers [Z]. Scienc Edirect. Dec. 2006