DOI:10.13718/j.cnki.xsxb.2019.03.019

一种基于改进信息增益特征选择的最大熵模型文本分类方法①

何 明

重庆工业职业技术学院 建筑工程与艺术设计学院,重庆 401120

摘要:针对传统信息增益(IG)特征选择算法忽略词频分布的缺陷,该文提出一种新的 IG 特征选择算法. 该算法通过引入均衡比和类内词频位置参数,解决了传统 IG 算法忽略词频分布对分类的弱化问题,修正传统类内词频位置参数,提高特征选择算法的文本分类精度,并将该改进 IG 特征选择算法用于最大熵模型(ME)对文本进行分类. 实验结果表明:该文所提方法在进行文本分类时 $F1$ 值高于传统 IG 算法. 该文方法的 ME 分类精度高于 K 最近邻 KNN(K-Nearest Neighbor)算法,说明本文方法是可行的、有效的.

关 键 词:信息增益;均衡比;词频参数;最大熵模型

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2019)03-0113-06

随着信息时代的迅速发展,信息数据呈指数级增长,面对如此巨大的信息资源,如何有效地管理,使人们更加方便快捷地获得目标信息成为研究热点. 文本信息挖掘中的文本分类技术有效解决了这一问题,文本分类是指未知类别的文本根据其内容信息,自动分类为一个或多个预先定义的类别[1]. 文本分类的核心问题是通过一种算法构建分类器,文本分类方法有很多,常见算法有 Rocchio 算法、朴素贝叶斯分类算法、KNN 算法[2]、神经网络算法、SVM 算法[3]等. 近年来,最大熵模型对未知事件尽可能使其分布均匀的特性,使得该算法在文本分类中得到研究与应用.

文本分类系统常采用向量空间模型(VSM)来实现文本信息描述,VSM 的特征项涉及整个文本的词条,导致了 VSM 的高维性[4-5],这对分类训练时间及准确性造成了很大影响. 特征选择实现了对 VSM 文本特征的降维处理,在文本分类时能够保留携带信息量大、对分类贡献大的词,在降低文本特征空间维数的同时提高分类性能[6]. 特征选择常用方法有互信息、开方检验、特征权(TS)、信息增益(IG)等. 互信息(Mutual Information, MI)表示 2 个变量之间的相关性,可以看成是一个随机变量中包含的关于另一个随机变量的信息量. 信息 IG 算法考虑特征词条未发生情况的优点使得其得到更多应用,常用于图像相关性处理和文本分类.

文献[7]针对传统 IG 算法的缺陷提出了一种改进 IG 算法,该算法引入特征分布差异因子、类内和类间加权因子,分类准确度优于传统算法,但是改进算法增大了算法计算复杂度,使得分类时间变长. 文献[8]通过引入比例因子提出了自适应 IG 算法,该算法自动调节比例因子,使得改进 IG 算法适用于平衡数据集和非平衡数据集等不同的语料库,同时解决了信息增益特征选择方法等比例地组合正、负相关特征导致文本分类中分类精度下降的问题. 文献[9]根据特征对 IG 贡献大小及在新文本中出现的次数,设计了 3 种基于 IG 特征权重的分类算法,能够在保证分类精度的同时提高训练速度,算法具有较低的时间复杂度. 文献[10]针对传统 IG 算法的缺点,对 IG 算法进行改进,该算法首先对数据集类别选择特征,然后优化合并不同类别特征,通过合并特征的出现概率来计算 IG 权重,并引入类内和类间分布因子,改进 IG 算法优

① 收稿日期:2018-04-03
基金项目:重庆市社会科学规划项目(2017YBYS108);重庆工业职业技术学院校级重点项目(GZY201709-2B).
作者简介:何 明(1980-),女,硕士,讲师,主要从事数字化艺术设计及信息化视觉传达研究.

于传统算法.

通过对传统IG算法及已有的改进算法进行分析研究，本文提出了一种新的改进IG算法. 该算法通过引入均衡比和类内词频位置参数，解决了传统IG算法忽略词频分布对分类的弱化问题和局部特征选择缺陷，修正传统类内词频位置参数，提高特征选择算法对分类精度的影响，并将该改进IG特征选择算法用于最大熵模型对文本进行分类，性能优于传统IG算法及其他分类算法.

1 信息增益算法

信息增益(IG)是基于熵理论的评估方法，亦是一种很有效的特征选择方法，具有考虑特征词条未发生情况的优点. 对于特征选择，就是将特征的重要程度量化之后再进行选择，而如何量化特征的重要性，就成为各种特征选择算法间最大的不同. 开方检验量化方法的特点是关联性越强的特征越重要，其方法是使用特征与类别间的关联性来量化特征重要性. IG特征选择算法的特点是特征携带信息越多，特征越重要，IG是根据特征能够为分类系统带来多少信息来量化特征重要性.

IG被定义为通过减少变量的不确定性而获得的信息量：$IG(X,Y)=E(X)-E(X\mid Y)$. IG计算的是词条特征，由信息理论可知，一个术语贡献的信息量越大就越重要. 术语f用于确定C类的信息增益值，定义如下

$$\begin{aligned} IG(f)=&E(C)-E(C\mid f)=-\sum_{i=1}^{n}P(C_i)\log_2 P(C_i)+\\ &P(f)\sum_{i=1}^{n}P(C_i\mid f)\log_2 P(C_i\mid f)+\\ &P(\bar{f})\sum_{i=1}^{n}P(C_i\mid \bar{f})\log_2 P(C_i\mid \bar{f}) \end{aligned} \tag{1}$$

式中C代表语料库中的文档集合；$P(C_i)$表示属于文档集中出现类别C_i的概率；$P(f)$表示特征f出现在文件中的概率；$P(\bar{f})$表示文件中没有出现f的概率；$P(C_i\mid f)$表示包含f的文档属于类别C_i的概率；$P(C_i\mid \bar{f})$表示包含f的文档不属于类别C_i的概率.

通过上述分析，可以得知IG算法是一种非常典型的监督特征提取算法，具有很好的降低特征维数的作用，并被广泛应用于文本分类研究. 然而，传统IG算法存在一些问题，导致IG算法在文本分类中精度还有提升空间. 针对IG算法中存在的问题，本文提出了改进IG特征选择算法，并将该算法与最大熵模型结合，使得本文算法在文本分类时性能达到最优.

2 改进IG特征选择的最大熵模型算法

IG算法中存在的问题主要有2个：

1) IG算法只考虑特征对整个系统的贡献，不能具体到一个特定的类别，这使得IG特征选择算法仅适合于进行所谓的“全局”特征选择(所有类别使用相同的特征集)，并且不能进行“本地”功能选择(因为每个类别都有自己的功能集，一些类别中的部分词出现次数很少).

2) 传统IG算法并没有考虑词频对文本分类的影响，2个特征项A和B分布在同一个文档中，即使A词频是B词频的10倍，仍然得到相同的IG值. 当忽略词频分布时，只考虑文档频率可能会导致特征预测能力弱化，不能选择最有效的特征.

针对以上2个问题，本文对IG特征选择算法进行改进，引入了均衡比因子和类内词频位置参数来提高IG算法的性能.

2.1 均衡比因子

均衡比因子是用来解决一个词在文档中出现多次与出现一次的IG值相同这个问题，IG算法中均衡比α表示词频在类别中分布均衡的程度. 对于训练集中类别$C_i(1\leqslant i\leqslant n)$，特征$f_i$在$C_i$类文本中出现的频数为$tf(f_i)$，设$\lambda_i=tf(f_i)/\sum tf(f_i)$，$tf(f_i)$表示在类别$C_i$中词频$f$的频率数，则词频均衡比$\alpha$可以定义为

$$\alpha=\frac{\sqrt{\frac{1}{n-1}\sum(\lambda_i-\bar{\lambda})^2}}{\bar{\lambda}} \tag{2}$$

其中，$\bar{\lambda}=(1/n)\sum_{i=1}^{m}\lambda_i$. 假设有 3 个类别，每个类别包含 5 个文档，有 2 个特征项 $F1$ 和 $F2$，其分布见表 1.

表 1　文档与词频表

文档	特征	1	2	3	4	5
类别 1	$F1$	15	11	0	0	0
	$F2$	3	1	0	0	0
类别 2	$F1$	0	2	0	0	0
	$F2$	0	2	0	0	0
类别 3	$F1$	0	0	0	0	0
	$F2$	0	0	0	0	0

从表 1 可以看出，$F1$ 较 $F2$ 更频繁出现在类别 1，因此判断 $F1$ 具有较强的分类能力. 但是用公式 (1) 计算时，2 个特征值具有相同的 IG 值，如果引入词频平衡因子 α，则可以得到不同的 IG 值，表 1 中 $\alpha_{F_1}>\alpha_{F_2}$，表明词频均衡比因子可以有效地纠正 IG 限制.

2.2　类内词频位置参数

在同一类别中每个文本的特征分布越均匀，特征分类能力就越强. 因此，本文引入类内词频位置参数，该参数由样本方差反映，统计样本方差的本质是反映样本间的分散程度. 假定在类别 $C_i(1\leqslant i\leqslant n)$ 的文本 $d_{ik}(1\leqslant i\leqslant N_i)$ 中，特征 f_i 出现的频数为 $tf(f_i)$，则每个频率之间的样本方差为

$$\beta_j=\sqrt{\frac{1}{N_i-1}\sum_{i=1}^{N_i}\left[tf_{ik}(f_j)-\frac{1}{N_i-1}\sum_{i=1}^{N_i}(tf_{ik}(f_j))\right]^2} \tag{3}$$

类别中每个文本的频率方差变化越小，分类能力越强，两者之间的反比关系越大. 因此，上述参数应归一化为

$$\beta=1-\beta_j/\sqrt{\sum_{j=1}^{m}\beta_j{}^2} \tag{4}$$

类别 C_i 文本的分布越均匀，则 β 越大，分类能力越强.

基于传统的 IG 算法引入了上述均衡比 α 和类内词频位置参数 β，则改进了算法可表示为

$$\begin{aligned}IG_{new}(f)=&\alpha\cdot\{-\sum_{i=1}^{n}P(C_i)\log_2P(C_i)+\beta\cdot[P(f)\sum_{i=1}^{n}P(C_i\mid f)\log_2P(C_i\mid f)\\&+P(\bar{f})\sum_{i=1}^{n}P(C_i\mid\bar{f})\log_2P(C_i\mid\bar{f})]\}\end{aligned} \tag{5}$$

2.3　最大熵模型

最大熵模型是一种概率估计，该模型原理表明预测一个随机事件的概率分布时，应当满足全部已知的约束，而未知的情况下不做任何主观假设，这样使得概率分布最均匀，预测风险最小，能够得到概率分布的最大熵.

在本文中，将需要分类的一个文本作为一个事件，则待分类文本集合可表示为$\{(r_1,s_1),(r_2,s_2)\cdots(r_N,s_N)\}$，其中 $r_i(1\leqslant i\leqslant N)$ 为具体文本，表示文本分类结果. 根据最大熵模型，模型的概率分布必须与获得最大熵的训练集一致，通过拉格朗日乘数算法，最大熵的概率分布为

$$p_w(s\mid r)=\frac{\exp(\sum_i w_if_i(r,s))}{Z_w(r)} \tag{6}$$

其中，f_i 表示特征函数，$f(r,s)\to(0,1)$，满足一定约束条件下取值为 1，其他情况下取值为 0；w_i 表示特征函数 f_i 的权值，反映了特征对模型的重要程度，本文使用改进迭代算法计算权值；$Z_w(r)$ 表示归一化因子，有 $Z_w(r)=\sum_s\exp(\sum_i w_if_i(r,s))$，保证对所有可能的文本 r 则有 $\sum_r p(s\mid r)=1$.

改进IG算法解决了传统IG算法的缺陷，在特征选择将特征项在类间和类内分布的均匀程度考虑进来. 另一方面，ME 模型是比较成熟的模型，对比其他常用分类模型如 SVM 和 KNN 等，其分类性能优于其他分类模型.

3　实验结果与分析

3.1　实验分类评价指标

通常我们评估文本分类的结果主要有 3 个方面：简单性、复杂性和算法的有效性. 本文仅考虑评价指

标的有效性，主要有准确度 P、召回率 R 和 $F1$ 值3个指标(表2).

表2 评测指标相关关系

数据集分类 C_i		数据集真实情况	
		属于 C_i 类	不属于 C_i 类
分类结果	属于 C_i 类	TP	FP
	不属于 C_i 类	FN	TN

注：TP 为真正类(true positive)，FP 为假正类(false Positive)，FN 为假负类(false negative)，TN 为真负类(ture negative).

准确率是 C_i 中样本数与 C_i 类中实际样本数之间的比例，定义为

$$p = \frac{TP}{TP + FP} \tag{7}$$

召回率是正确分类的样本数和样本数应按类别 C_i 正确分类的比例，定义为

$$R = \frac{TP}{TP + FN} \tag{8}$$

准确率和召回率从2个方面反映了分类结果，一个是对完整性的研究，另一个是对正确性的研究. 两者是成反比的，往往提高前者的表现会导致后者的表现下降. 因此，对于两者的价值评估将综合考虑使用 $F1$ 值，定义为

$$F_1 = \frac{2PR}{P + R} \tag{9}$$

3.2 实验结果

本文选取军事，金融，艺术，体育4个大类文件进行实验(表3).

表3 实验数据集选取

类 别	军事	金融	艺术	体育
训练文档	160	200	160	300
测试文档	80	100	80	150

首先采用中国科学院开发的中文分词系统处理单词分割和去除停止词，然后使用改进的特征选择算法进行特征提取，最后采用最大熵模型及 KNN 算法用于分类测试. 本文进行3次实验，第1次实验是验证本文特征选择算法的维数与分类性能的比较；第2次实验是改进 IG 特征与不同分类算法的性能比较；第3次实验是在搜狗实验室文本分类语料库上，对本文方法与传统 IG 方法进行比较实验. 上述实验都是采用经典的十折交叉验证方法，实验文本平均分为10份，其中9份作为训练集，1份作为测试集，取10次分类结果的平均值作为测试结果.

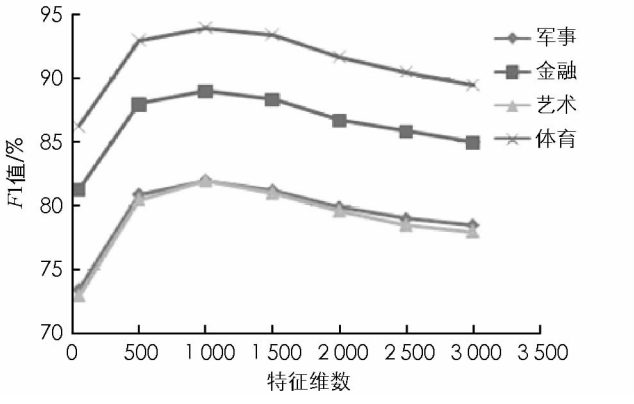

图1 特征向量不同维度下分类 $F1$ 值

分类方法采用最大熵模型算法，改进 IG 算法作为特征选择算法，特征维数对分类性能的影响见图1.

从图1可以看出，当特征维数为1 000时，$F1$ 值最好，本文后2项实验中特征维数都采用1 000.

在本文实验中，采用改进 IG 特征选择算法用于最大熵进行分类，并与 KNN 分类方法进行比较，特征选择的维数为1 000，分类结果见表4.

从上述数据可以看出，基于改进的 IG 特征选择的 ME 分类方法 $F1$ 值比 KNN 分类方法在体育文本的分类提高9.09%，在艺术方面分类提高最多为12.06%，说明本文提出的改进 IG 算法 ME 分类性能优于 KNN 算法. 另外，选择搜狗实验室文本分类语料库作为实验对象，对本文算法进行验证，将本文算法与 SVM 和 KNN 分类算法进行比较，采用经典的十折交叉验证，得到 $F1$ 值见表5.

表 4　不同分类算法的评价指标　%

	ME 算法			KNN 算法		
	P	*R*	*F*1	*P*	*R*	*F*1
军事	76.92	87.72	81.97	66.1	79.59	72.22
金融	84.38	94.17	89.01	73.42	81.69	77.33
艺术	75.76	89.29	81.97	64.29	76.6	69.91
体育	92.54	95.38	93.94	82.35	87.5	84.85

表 5　搜狗语料库下不同分类算法 *F*1 值比较　%

类别	分类算法		
	SVM	KNN	ME
军事	74.51	73.14	82.3
金融	79.32	77.95	90.34
艺术	72.73	71.33	83.01
体育	86.74	85.37	94.57

从表 5 中可以得出，本文算法性能优于 SVM 和 KNN 算法.

改进 IG 算法下 2 种分类的 *F*1 值直观比较见图 2.

分类方法采用最大熵模型算法，将文本方法与传统 IG 算法进行对比实验，本次实验结果采用十折交叉方法，实验结果见表 6.

分析表 6 和图 3 内容，可以得出以下结论：基于改进的 IG 特征选择算法较传统 IG 算法，*F*1 值军事文本分类提高最小，为 6.97%，金融文本分类提高最多，达到 9.27%，说明使用改进的 IG 算法用于 ME 分类方法在准确率、召回率和 *F*1 值 3 个方面都有提高，总体分类结果优于传统 IG 算法.

另外，在搜狗实验室文本分类语料库中将本文算法与传统 IG 算法进行比较，采用经典的十折交叉验证，得到 *F*1 值见图 4.

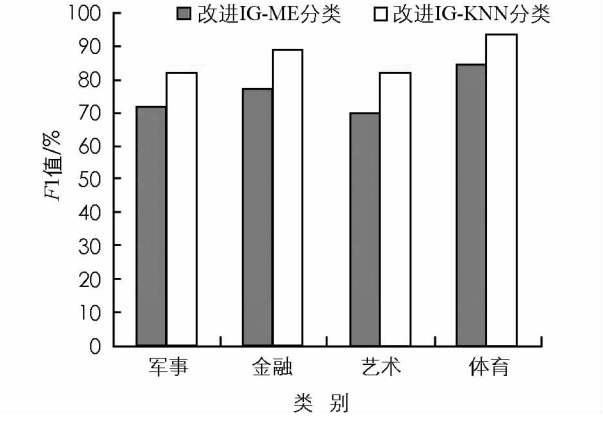

图 2　不同分类方法 *F*1 值对比图

表 6　不同特征选择算法的评价指标　%

	传统 IG			本文方法		
	P	*R*	*F*1	*P*	*R*	*F*1
军事	67.74	84	75	76.92	87.72	81.97
金融	74.39	85.92	79.74	84.38	94.17	89.01
艺术	66.1	81.25	72.9	75.76	89.29	81.97
体育	82.78	89.38	85.95	92.54	95.38	93.94

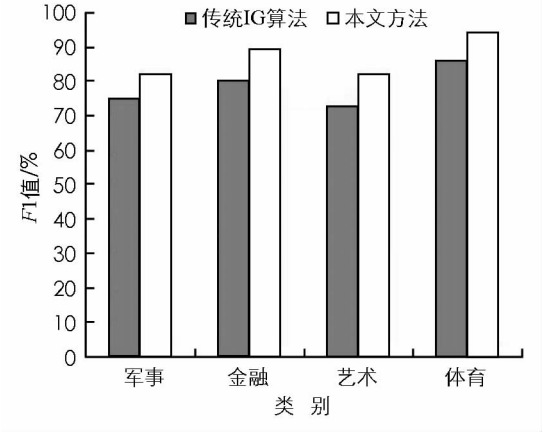

图 3　不同特征选择算法 *F*1 值对比图

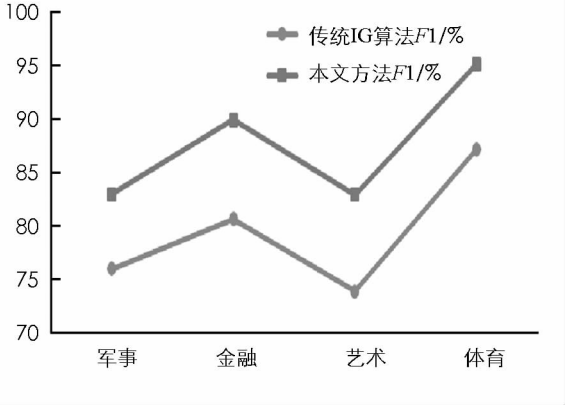

图 4　搜狗语料库中不同特征算法性能

由图4中数据可以看出，本文方法比传统IG算法的$F1$值高，即本文算法在本文分类的性能高于传统IG算法.

4 结 论

作为文本分类的关键技术，特征选择算法对分类性能有直接影响. 本文研究了传统IG特征选择算法，针对传统IG算法的不足，引入均衡比和类内词频位置参数，提出一种新的信息增益(IG)特征选择算法，该算法解决了传统IG算法忽略词频分布对分类的弱化问题，修正传统类内词频位置参数，提高了分类精度，并将该改进IG特征选择算法用于最大熵模型对文本进行分类，实验结果表明本文方法优于传统IG特征选择算法. 把本文方法用于ME分类时性能优于KNN分类算法，说明本文方法可行、有效.

参考文献：

[1] 赖 娟，金 澎，洪艳伟. 文本分类中的主动多域学习[J]. 西南师范大学学报(自然科学版)，2014，39(7)：108-114.
[2] 周庆平，谭长庚，王宏君，等. 基于聚类改进的KNN文本分类算法[J]. 计算机应用研究，2016，33(11)：3374-3377.
[3] 古丽娜孜·艾力木江，孙铁利，乎西旦，等. 一种基于SVM-修正KNN算法的哈萨克语文本分类[J]. 西北师范大学学报(自然科学版)，2014，50(3)：48-53.
[4] JIANG R，KIM S，BANCHS R E，et al. Towards Improving the Performance of Vector Space Model for Chinese Frequently Asked Question Answering [C]//Asian Language Processing (IALP)，2015 International Conference on. Suzhou：IEEE，2015.
[5] SANTOSO I B，DEWA C K，AFIAHAYATI. The Implementation of Vector Space Model for Infectious Disease Diagnosis System Based on Pathophysiology Science [J]. Advanced Science Letters，2018，24(1)：682-685.
[6] 肖 雪，卢建云，余 磊，等. 基于最低词频CHI的特征选择算法研究[J]. 西南大学学报(自然科学版)，2015，37(6)：137-142.
[7] 郭 颂，马 飞. 文本分类中信息增益特征选择算法的改进[J]. 计算机应用与软件，2013，30(8)：139-142.
[8] 董 微，刘 学，倪 宏. 基于信息增益的自适应特征选择方法[J]. 计算机工程与设计，2014(8)：2856-2859.
[9] 李文斌，刘椿年，陈嶷瑛. 基于特征信息增益权重的文本分类算法[J]. 北京工业大学学报，2006，32(5)：456-460.
[10] XU J，JIANG H. An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier [C]//Cyber-Enabled Distributed Computing and Knowledge Discovery(CyberC)，2015 International Conference on. Xi'an：IEEE，2015.

A Maximum Entropy Model Text Classification Method Based on Improved Information Gain Feature Selection

HE Ming

Institute of Construction Engineering and Art Design，Chongqing Industry Polytechnic College，Chongqing 401120，China

Abstract：For the shortcomings of traditional information gain (IG) feature selection algorithm：ignoring word frequency distribution，a new IG feature selection algorithm is proposed in this paper. The algorithm introduces the equalization ratio and word frequency location parameters within class. The new algorithm solves the problem that the traditional IG algorithm ignores the word frequency distribution and modifies the position parameter of word frequency within class to improve the accuracy of text classification. At last，the improved IG feature selection algorithm is applied to maximum entropy model (ME) for text classification. Experimental results show that，Compared with the traditional IG algorithm，the F1 value of the proposed method in this paper is higher than the traditional IG algorithm in text classification. In addition，the ME classification accuracy of this method is higher than the KNN algorithm，which shows that this method is feasible and effective.

Key words：information gain；equalization ratio；word frequency parameter；maximum entropy model

责任编辑 夏 娟