文章编号：2096-1472(2017)-12-19-04

基于词频信息改进的IG特征选择算法在文本分类中的应用研究

牛玉霞
(南通科技职业学院，江苏 南通 226007)

摘　要：IG算法是一种有效的特征选择算法，在文本分类研究领域中得到了广泛应用。本文针对IG算法的不足，提出了一种基于词频信息的改进方法，分别从类内词频信息、类内词频位置分布、类间词频信息等方面进行了改进。通过实验对改进的算法进行了测试，结果表明，改进的算法相对传统算法更有效。

关键词：词频信息；IG算法；特征选择；文本分类
中图分类号：TP391.1　　**文献标识码**：A

Research on the Application of the IG Feature Selection Algorithm Based on Word Frequency Information Improvement in Text Classification

NIU Yuxia
(*Nantong Science and Technology Academy*,*Nantong* 226007,*China*)

Abstract: As an effective feature selection algorithm,the IG algorithm has been widely used in the field of text classification.Aiming at the shortcomings of the IG algorithm,this paper proposes an improved method based on word frequency information,which improves the intra-class frequency information,the intra-class word frequency location distribution and the inter-class word frequency information.Experiments are carried out to test the improved algorithm,and the results show that the improved algorithm is more effective in comparison with the traditional one.

Keywords:word frequency information;IG algorithm;feature selection;text classification

1 引言(Introduction)

随着信息技术的飞速发展，互联网信息资源呈爆炸式增长。面对海量信息，如何合理管理资源，使人们能够快速、准确地获取有效信息，已经成为IT行业的研究热点之一[1]。

文本分类技术是文本信息处理的关键技术之一，能够很好地解决上述问题，在文本分类中，通常用向量空间模型来表示结构化文本，其中，文本特征的高维性和特征权值的稀疏性直接影响文本分类精度。因此，设计合理的特征降维方法可以提高文本自动分类的效率。特征选择模式是常用的文本特征降维方式。该模式计算复杂度低，容易理解。特征选择的主要方法有：文档频度(Document Frequency，DF)、互信息(Mutual Information，MI)、文本证据权(Weight of Evidence，WE)、χ^2统计量(Chi-square，CHI)、期望交叉熵(Expected Cross Entropy，ECE)、信息增益(Information Gain，IG)等。相关研究表明[2,3]，在信息类别分布均衡的情况下，信息增益优势明显，但在类偏斜条件下，信息增益的分类效果就会下降。就信息增益的不足，探索相应的改进方法，提高文本分类的性能，有重要的现实意义。

2 信息增益文本特征选择算法(Information gain text feature selection algorithm)

信息增益(Information Gain，IG)的评估方法是以熵为理论基础的[4]。熵越大，表明体系分布不确定、混乱。设X是随机变量，它可能有n个取值x_1，x_2，…，x_n，每个取值取到的概率分别为$P(x_1)$，$P(x_2)$，…，$P(x_n)$，则X的信息熵为：

$$E(X)=-\sum_{i=1}^{n}P(x_i)\log P(x_i) \quad (1)$$

当Y确定以后，则X的熵为

$$E(X|Y)=\sum_{i=1}^{n}P(x_i)\,E(X|Y=y_i) \quad (2)$$

信息增益是熵的差值，表示在去掉变量的不确定性后得到的信息量，表示为：

$$IG(X,Y)=E(X)-E(X|Y) \quad (3)$$

基金项目：江苏省高等学校大学生创新创业训练计划(项目编号：201712684020T).

IG是针对特征项而言的。设ω为特征项，C为文本类别，用ω在C类中是否出现所带来的信息量来确定ω对C的信息增益值，如式(4)所示。

$$\begin{aligned} IG(\omega) &= E(C) - E(C|\omega) \\ &= \sum_{i=1}^{n} P(C_i)\log_2 P(C_i) + P(\omega)\sum_{i=1}^{n} P(C_i|\omega)\log_2 P(C_i|\omega) \\ &\quad + P(\overline{\omega})\sum_{i=1}^{n} P(C_i|\overline{\omega})\log_2 P(C_i|\overline{\omega}) \end{aligned} \tag{4}$$

其中，n表示总的文档类别数，$P(C_i)$表示在文档集合中属于C_i类的文档出现概率，$P(\omega)$表示含有特征项的文本在文档集合中出现的概率，$P(\overline{\omega})$表示不含特征项在文档集合中出现的概率，$P(C_i|\omega)$表示含ω特征项属于C_i类别的概率，$P(C_i|\overline{\omega})$表示含特征项不属于C_i类别的概率。

3 改进信息增益算法(Improved information gain algorithm)

3.1 基于类内词频信息改进IG算法

传统信息增益算法中计算的概率P均是基于文档数量的，没有考虑特征项ω词频因素[5]。比如，特征项中的ω_x与ω_y在类别C_i中的大部分文本中出现，在其他类别中基本不出现，那么，ω_x和ω_y可能是C_i的特征项。由式(3)计算得到的两个特征项与类别之间的IG值应该基本接近。但是，如果特征项ω_x在类别C_i中出现的次数远远大于特征项ω_y时，即特征项ω_x对C_i的分类能力远远大于特征项ω_y，由式(3)计算得到的两个特征项IG值仍然接近。因此，在评估特征项ω_i对文档类别C_i的分类能力时，传统的信息增益算法考虑了在类别C_i中出现特征项ω_i文档的数量，而没有考虑特征项ω_i在C_i中各个文档中出现的次数。

由上述情况可知，某一个特征项在某一个文档类别中出现的次数越多，则该特征项对文档类别而言分类能力就越强，该特征项的IG应该放大，因此，考虑为特征项增加权重参数，出现频数越大，则分配较大的权重值。记特征项集合$\omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$，文本类别$C_i(1 \le i \le n)$中的文本有$d_{ik}(1 \le k \le N_i)$，其中，$N_i$是类别$C_i$中包含文档的总数。设特征项$\omega_j(1 \le j \le m)$在文档$d_{ik}$(属类别$C_i$)出现的频数为$tf_{ik}(\omega_j)$，那么权重参数为：

$$\alpha_{ij} = \sum_{i=1}^{N_i} \frac{tf_{ik}(\omega_j) - \min_{1\le l\le m}(tf_{ik}(\omega_l))}{\max_{1\le l\le m}(tf_{ik}(\omega_l)) - \min_{1\le l\le m}(tf_{ik}(\omega_l))} \tag{5}$$

不同类别中的文档数量也有所不同，因此，将式(5)进一步做归一化处理。

$$\alpha = \frac{\alpha_{ij}}{\sqrt{\sum_{j=1}^{m}\alpha_{ij}^2}} \tag{6}$$

由改进的模型可以看出，特征项ω_i在文档中出现的频数与其权重值α呈正比关系，即某一特征项在某类别中出现越频繁，则分类能力就越强。

3.2 基于类内词频位置分布信息改进IG算法

相关研究表明，在文本类别中分类能力越强的特征项，不仅出现频数要大，而且在该类别中的分布位置应该均匀[6,7]。比如，在类别C_i中都出现了特征项ω_i和ω_j，特征项ω_i在每个文档中都出现，而且出现频数接近，分布均匀，特征项ω_j只在个别文档中出现，而且出现频数很高，在其他文档中出现频数很少，即特征项ω_j在类C_i中呈偏斜分布。在这种情况下，我们认为特征项ω_i对类别C_i的分类能力更强。但是，公式(3)没有考虑这一因素，计算得到的结论恰恰相反。

因此，基于特征项在类内文本分布信息进行改进，在模型中引入样本方差。样本方差在统计学中用来表示样本之间的离散程度，方差越大，表示样本分布越不均匀，即越偏斜；方差越小，表示样本与其均值之间的偏差越小，分布越均匀。在本文中，表示特征项在同一类别各个文档中频数的分布情况。

记特征项$\omega_j(1 \le j \le m)$在类别C_i的某一文档$d_{ik}(1 \le k \le N_i)$中出现的频数为$tf_{ik}(\omega_j)$，那么每个频数之间的样本方差可表示为

$$\beta_j = \sqrt{\frac{1}{N_i-1}\sum_{k=1}^{N_i}\left[tf_{ik}(\omega_j - \frac{1}{N_i-1}\sum_{k=1}^{N_i} tf_{ik}(\omega_j))\right]} \tag{7}$$

特征项在文档类别中出现频数的方差与其分类能力成反比，即方差越小，分类能力越强。所以，将式(7)表示的方差参数进行进一步修正，如式(8)所示。

$$\beta = 1 - \frac{\beta_j}{\sqrt{\sum_{j=1}^{m}\beta_j^2}} \tag{8}$$

在文档类别C_i中，特征项ω_i每个文本中分布越均匀，β值就越大，相应的分类能力也就越强。

3.3 基于类间词频信息改进IG算法

特征项在不同的文本类别中出现的频数也能反应其相对文本类别的分类能力[8]。如果特征项ω_i在类别C_i中出现频繁，且分布均匀，在其他类别$C_j(j \ne i)$中出现很少，那么ω_i表现出对类别C_i较强的分类能力；相反，如果特征项ω_i在所有文档类别中都频繁出现，那么表现出的分类能力就较差。仍然以特征项在类别中的词频方差作为权重参数，特征项ω_j在不同类别中的词频数方差越大，则分类能力越强。

设特征项$\omega_j(j=1,2,\cdots,m)$在类别$C_i(i=1,2,\cdots,n)$

的所有文档$d_{ik}(1\le k\le N_i)$中出现的频数为$tf_{C_i}(\omega_j)=\sum_{k=1}^{N_i}tf_{ik}(\omega_j)$，则特征项$\omega_j$在每个类别$C_i$中的频数间样本方差可表示为

$$\lambda_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[tf_{C_i}(\omega_j)-\frac{1}{n-1}\sum_{i=1}^{n}tf_{C_i}(\omega_j)\right]^2} \tag{9}$$

在此基础上，将参数做归一化处理，如式(10)所示。

$$\lambda=\frac{\lambda_j}{\sqrt{\sum_{j=1}^{m}\lambda_j^2}} \tag{10}$$

λ参数体现了特征项在不同文本类别中出现频数的分布情况，分布越偏斜，则分类能力更强，反之则弱。

综上所述，通过引入α、β、λ三个权重参数，对传统IG算法进行了优化，得到改进的模型如式(11)所示。

$$\begin{aligned}IG_{imp}(\omega)=&-\sum_{i=1}^{n}P(C_i)\log_2P(C_i)\\&+(P(\omega)\sum_{i=1}^{n}P(C_i|\omega)\log_2P(C_i|\omega)\\&+P(\overline{\omega})\sum_{i=1}^{n}P(C_i|\overline{\omega})\log_2P(C_i|\overline{\omega}))\times\alpha\times\beta\times\lambda\end{aligned} \tag{11}$$

改进的算法综合考虑了类内词频信息、类内词频位值分布信息、类间词频分布信息三个因素的影响，即特征项在类内少数文档类别中出现频数越高，分类能力越强；特征项在类内出现频数高，且分布均匀，分类能力越强；特征项在类别间分布越偏斜，分类能力越强。实验表明，改进的特征选择算法IG_{imp}相对 IG效果更好。

4 实验过程与结果分析(Experiment process and result analysis)

4.1 选取实验文本

本文对改进的模型进行了分本分类实验，实验数据来自复旦大学李荣陆教授提供的语料库，包括教育、经济、环境、计算机、医药、艺术、交通、政治、体育、军事10个主题类别，选取926篇作为测试集，1851篇文本作为训练集，具体分布情况如表1所示。

表1 实验文本详细分布情况

Tab.1 Detailed distribution of experimental text

文档类别	教育	经济	环境	计算机	医药	艺术	交通	政治	体育	军事
测试文档	73	108	66	67	67	79	68	164	150	82
训练文档	146	216	134	135	134	159	136	328	300	164

使用中科院ICTCLAS分词系统进行分词处理，剔除无用词和停用词，使用文中改进的模型进行特征提取，使用KNN分类算法进行测试。

4.2 确定实验K值

KNN分类算法中的K值是不确定的，需要通过实验，选择准确率最高K的取值。用传统IG算法，特征提取维数1000，K分别取5、8、10、12、14、18。从图1中可以看出，当K取12时，分类器准确率达到最高，所以，在对比实验中，K的值取12。

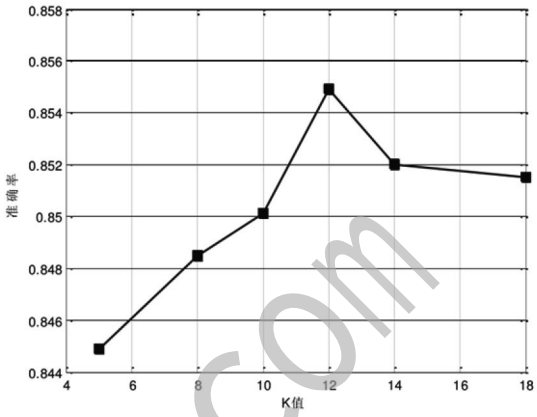

图1 不同K值的准确率情况

Fig.1 Accuracy of different K values

4.3 分析实验结果

本文实验比较了改进算法IG_{imp}与传统IG算法的分类效果，采用KNN分类算法、TF−IDF加权算法，使用查准率P、查全率R和F_1测试值作为分类效果的评估指标。

查准率P=正确分类文本数/实际分类文本数

查全率R=正确分类文本数/类内文本数

$$F_1=2\times P\times R/(P+R)$$

实验统计结果如表2所示，其中，P、R和F_1表示IG算法分类评估值，$P-new$、$R-new$和F_1−new表示改进算法IG_{imp}的分类评估值。

表2 改进算法IG_{imp}与传统IG算法分类效果评估

Tab.2 IG_{imp} and IG classification effect assessment

IG/IG_{imp}	教育	经济	环境	计算机	医药	艺术	交通	政治	体育	军事
查准率P	0.766	0.808	0.877	0.916	0.915	0.910	0.904	0.752	0.849	0.829
查全率R	0.884	0.898	0.606	0.968	0.618	0.850	0.970	0.757	0.943	0.590
F_1	0.821	0.851	0.717	0.941	0.738	0.879	0.936	0.754	0.894	0.689
查准率P−new	0.849	0.846	0.905	0.942	0.957	0.924	0.965	0.773	0.875	0.854
查全率$R-new$	0.879	0.963	0.622	0.984	0.661	0.892	0.981	0.759	0.968	0.621
F_1-new	0.864	0.901	0.737	0.963	0.782	0.908	0.973	0.766	0.919	0.719

为直观比较，将表2中的数据用直方图表示，查准率P直方图如图2所示。从图中可以看出，改进的算法的查准率在实验文本的十个类别中都有所提高，平均提高率为4.27%，尤其是在教育、交通类中，分别提高了10.84%和6.75%。

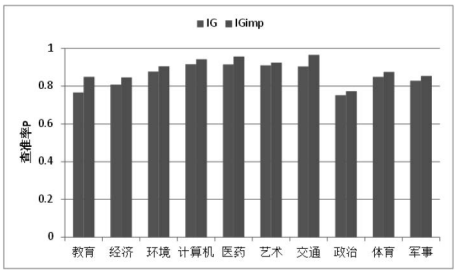

图2 改进算法IG_{imp}与传统IG算法查准率对比图

Fig.2 IG_{imp} and IG comparison diagram of precision ratio

图3为查全率R对比直方图，改进的算法在查全率方面平均提高率为3.04%，经济、医药、艺术类别的查全率高于平均提高率，分别为7.24%、6.96%、4.94%，教育类的查全率稍有下降。

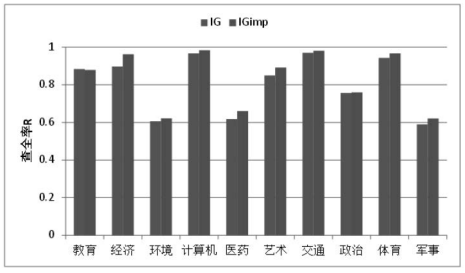

图3 改进算法IG_{imp}与传统IG算法查全率对比图

Fig.3 IG_{imp} and IG comparison diagram of recall rate

图4为F_1评估值对比直方图，改进的算法F_1值在十个种类中都有提高，平均提高率为3.8%，医药、经济、教育类有明显提高，分别提高了5.96%、5.88%、5.24%。

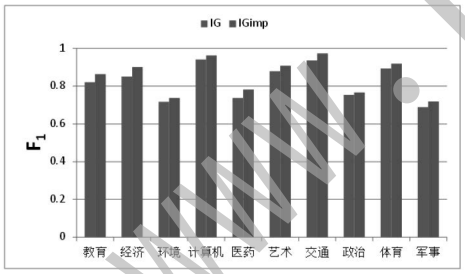

图4 改进算法IG_{imp}与传统IG算法F_1对比图

Fig.4 IG_{imp} and IG comparison diagram of F_1

笔者利用宏平均查准率、宏平均查全率、宏平均F_1三个评估指标，对改进算法IG_{imp}与传统IG算法F_1做了比较，可以从整体上看出两种算法的分类效果。具体数据如表3所示。

表3 改进算法IG_{imp}与传统IG算法宏评价结果

Tab.3 IG_{imp} and IG macro evaluation results

	宏平均查准率	宏平均查全率	宏平均F_1
IG	0.834	0.901	0.847
IG_{imp}	0.860	0.919	0.866

从表3中可以看出，相比较IG算法，IG_{imp}在宏平均查准率、宏平均查全率和宏平均F_1方面分别提高了3.12%、2%、2.24%。

综合分析改进IG_{imp}算法的分类效果，在查准率、查全率和F_1方面比IG算法的效果要好，仅在个别类别的查全率略有下降。从整体上看，改进IG_{imp}算法的文本分类效果优于传统IG算法。

5 结论(Conclusion)

本文针对信息增益算法在特征项频数分布方面的不足进行了改进，引入了三个权重参数，分别从类内词频信息、类内词频分布、类间词频分布三个方面进行了改进，使得优化的信息增益模型IG_{imp}有更强的类别特征选择能力。通过对文本样本分类实验对比，证明了改进的IG_{imp}算法有更强的文本分类能力。

参考文献(References)

[1] Ghosh A K,Chaudhuri P,Murthy C A.Multiscale classi-fication using nearest neighbor density estimates[J].IEEETransactions on Systems,Man,and Cybernetics-part B:Cybernetics,2006,36(5):1139-1148.

[2] Liu L,Ren J Y,Zhou J,et al.Carrier frequency offset and I/Q imbalance compensation for MB-OFDM based UWB system[J].Wireless Personal Communications,2013,71(2):1095-1107.

[3] Sharma R.Lalitha H.Kumar N.Design and development of nono data aided estimation algorithm for carrier frequency-offset and I/Q imbalancing in OFDM-based systems[C].Wireless and Optical Communications Networks,2013:1-4.

[4] 石慧.基于特征选择和特征加权算法的文本分类研究[D].山东师范大学,2015.

[5] 刘海峰,刘守生,宋阿羚.基于词频分布信息的优化IG特征选择方法[J].计算机工程与应用,2017,53(4):113-116;122.

[6] 黄志艳.一种基于信息增益的特征选择方法[J].山东农业大学学报,2013,44(2):252-256.

[7] 任永功,杨荣杰,尹明飞,等.基于信息增益的文本特征选择方法[J].计算机科学,2012,39(11):127-130.

[8] 熊忠阳,黎刚,陈小莉.文本分类中词语权重计算方法的改进与应用[J].计算机工程与应用,2008,44(5):187-189.

作者简介：

牛玉霞(1981-)，女，硕士，讲师.研究领域：计算机应用技术，物联网技术.