基于MRMR的文本分类特征选择方法

李军怀 付静飞 蒋文杰 费 蓉 王怀军
(西安理工大学 西安 710048)

摘 要 特征选择是文本分类技术中重要的处理步骤,特征词选择的优劣直接关系到后续文本分类结果的准确率。使用传统特征选择方法如互信息(MI)、信息增益(IG)、χ^2 统计量(CHI)等提取的特征词仍存在冗余。针对这一问题,通过结合词频-逆文档率(TF_IDF)和最大相关最小冗余标准(MRMR),提出了一种基于MRMR的特征词二次选取方法TFIDF_MRMR。实验结果表明,该方法可以较好地减少特征词之间的冗余,提高文本分类的准确率。

关键词 特征选择,最大相关最小冗余,词频-逆文档率,文本分类
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.043

Feature Selection Method Based on MRMR for Text Classification

LI Jun-huai FU Jing-fei JIANG Wen-jie FEI Rong WANG Huai-jun
(Xi'an University of Technology, Xi'an 710048, China)

Abstract Feature selection is the most important preprocessing step in text classification. The quality of the feature words has a significant impact on the accuracy of the classification results. Using the traditional feature selection method, such as MI, IG, CHI, will still cause the redundanly among the feature words. For this problem, based on the combination of the term frequency-inverse document frequency (TF-IDF) and the maximal relevance minimal redundancy (MRMR), this paper put forward a MRMR based feature words secondary selection method TFIDF_MRMR. The experimental results indicate that this method is able to reduce the redundancy of feature words and improve the accuracy of classification.

Keywords Feature selection, Maximal relevance minimal redundancy (MRMR), Term frequency-inverse document frequency(TF-IDF), Text classification

1 引言

文本分类是指对一系列文档按照预先定义的分类体系进行归类的一门技术。如何在面对海量数据的情况下提取出人们需要的信息?文本分类技术提供了一种有效的解决方法。

文本分类一般包括预处理、特征选择、模型训练、文档分类等步骤。其中,特征选择是从预处理的数据集中选择最优的特征子集,通过特征选择可以对高维特征空间进行降维,得到的低维的特征子集可提高运算效率,缩短训练时间,通常还可以得到更精确的分类结果。因此,特征词选取的优劣直接决定了后续的训练效率和分类效果。目前常用的特征词选取方法有信息增益(Information Gain, IG)、文档词频(Document Frequency, DF)、χ^2 统计量(Chi Square Statistic, CHI)[1]、T检验(Student's t test, t-test)[2]等。这些常用的特征词选取方法考虑了特征词和文档类别之间的关系,没有考虑特征词之间的关系,特征词之间可能还会存在冗余,即需要对特征词进行二次提取,去掉冗余。

在两阶段特征选取方面,Uguz[3]提出了IG-GA和IG-PCA特征选择方法,首先按信息增益(IG)对特征词进行排序,选择满足条件的特征词;之后使用遗传算法(Genetic Algorithm, GA)或主成分分析(Principal Component Analysis, PCA)方法再次提取特征子集。Jiana Meng[4]等人首先使用特征贡献度(Feature Contribution Degree, FCD)进行初次特征选取,之后使用隐式语义索引(Latent Semantic Indexing, LSI)方法考虑特征词间的相关度进行二次选取。Uysal Gunal[5]提出的两阶段特征选取方法采用类似DFS或CHI的过滤器进行初次选取,使用以LSI为导向的GA方法再次提取特征。以上方法在对特征词进行二次选取时均采用某种评判标准选择 $m \ll M$(m:二次特征集,M:初次特征集)的特征子集,导致 m 集中会丢失文档的原始语义。Javed K等人[6]提出的两步骤的特征词选取方法首先使用二元正态分离(Binormal Separation, BNS)方法或信息增益方法对文档进行特征选取,之后采用马尔科夫链过滤(Markov Blanket Filtering, MBF)算法筛选特征子集。该方法虽在一定程度上克服了语

到稿日期:2015-09-21 返修日期:2015-12-16 本文受国家自然科学基金(61172018),陕西教育厅科技计划(15JS077),西安市科技计划(CXY1439(8))资助。

李军怀(1969—),男,博士,教授,CCF高级会员,主要研究领域为网络计算,E-mail:lijunhuai@xaut.edu.cn;付静飞 男,硕士生,主要研究领域为移动计算技术;蒋文杰 男,硕士生,主要研究领域为移动计算;费 蓉 女,副教授,主要研究领域为智能计算;王怀军 男,讲师,主要研究领域为网络计算。

义丢失的问题，但计算相对复杂，分类效率也有待进一步提升。

为了弥补上述方法存在的不足，既能保证特征子集语义的完整，又要求可减小生成子集的计算代价，本文通过结合两种不同的特征选取方法提出了一种 TFIDF_MRMR 两阶段特征优选方法。

2 基于 MRMR 的两阶段特征选择方法

本文使用的分类流程如图 1 所示。其中，在特征选择过程中使用了本文 TFIDF_MRMR 两阶段特征提取方法，首先使用改进的 TF_IDF 方法初步筛选文档特征词，然后使用 MRMR 方法去除特征词之间的冗余，得到最终优选的特征词集。

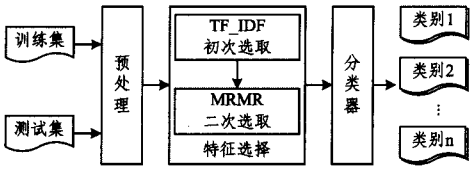


图 1 文本分类过程示意图

2.1 初次选取

TF_IDF(Term Frequency-Inverse Document Frequency)是一种用于资讯检索与文本挖掘的常用加权技术[7,8]。首先，将单词 w 看作是一个随机变量，单词 w 在各个类之间的取值用单词在各个类之间的词频(即单词在各个类中出现的次数)表示。根据方差的含义，若单词 w 在各类之间分布比较均匀，那么 $D(w)$ 就越小，单词 w 对分类的贡献就越小；若单词 w 均匀地分布在各类之间，则 $D(w)$ 为 0，即单词 w 对本次分类没有贡献。因此使用 $D(w)$ 作为修正因子来修正 TF_IDF 公式，TF_IDF 公式转换为式(1)。

$$TF_{IDF(w_i)}=TF(w_i)\times IDF(w_i)\times D_e \tag{1}$$

D_e 表示单词 w 的平均方差，计算公式为：

$$D_e=\frac{1}{n}\sum_{i=1}^{n}[tf_i(w)-\overline{tf(w)}]^2 \tag{2}$$

其中，本文中单词 w 在各类别的取值用词频表示，设共有 n 个类别，$tf_i(w)$ 表示单词 w_i 在类别 c_i 中出现的次数，$\overline{tf(w)}$ 表示单词 w 在各个类别中出现的平均次数。$\overline{tf(w)}$ 计算方法为：

$$\overline{tf(w)}=\frac{1}{n}\sum_{i=1}^{n}tf_i(w) \tag{3}$$

2.2 二次选取

使用改进的 TF_IDF 方法进行特征词选择时，仅考虑了特征词与类别之间的相关性，没有考虑特征词之间的相关性，所以使用 TF_IDF 方法选取的特征词之间会存在冗余，因此要对使用 TF_IDF 提取的特征词进行二次选取。本文采用最大相关最小冗余标准(Minimal Redundancy Maximal Relevance，MRMR)[9-11]消除特征词间的冗余。MRMR 算法是典型的基于空间搜索的过滤方法，使用互信息衡量特征的相关性与冗余度[11]。最大相关表示特征词与文档类别相关度大，即能最大程度反映文档类别信息；最小冗余表示特征词之间相关度最小，即特征冗余度最小。

二次特征选取时使用的语料库由 2.1 节使用 TF_IDF 初次选取获得的 $tfidfDic$(由 TF_IDF 方法生成的单词字典)中的特征词表示。每类文档表示为 $Doc_i=\{w_1,w_2,\cdots,w_n\}$，其中 $tfidfDic$ 含有 n 个单词。进行二次特征选取时，要在 $tfidfDic$ 的 n 个单词中使用 MRMR 标准选取一个合适的特征子集 S，其中 $S\subset tfidfDic$，特征子集 S 用于后续的文本分类。

为了选择与类别 c 相关性大的单词，需要使用最大相关式(4)选择符合条件的特征词，其中 $|S|$ 表示集合 S 中已经存在的单词的个数。

$$\max D(S,c),D=\frac{1}{|S|}\sum_{t_i\in S}I(t_i,c) \tag{4}$$

使用最大相关式(4)选择的特征词之间可能会存在冗余。如果两个单词高度依赖，那么移除一个单词之后，不会影响分类的效果，所以要对特征子集 S 使用最小冗余标准来检查冗余，如式(5)所示。

$$\min R(S),R=\frac{1}{|S|^2}\sum_{w_i,w_j\in S}I(w_i,w_j) \tag{5}$$

结合上述两种约束条件，产生 MRMR，如式(6)。

$$\max \emptyset(D,R),\emptyset=D-R \tag{6}$$

由于单词之间的互信息 $I(w_i,w_j)$ 计算比较耗时，本文使用增量式搜索思想获得合适的特征子集 S。设已经选取了特征子集 S_{m-1}，即特征子集里已经包含 $m-1$ 个特征词，然后从剩余的单词集合 $\{tfidfDic-S_{m-1}\}$ 选取出第 m 个特征单词。选取第 m 个特征词的公式如式(7)所示。

$$\max_{t_j\in tfidfDic-S_{m-1}}[I(w_j,c)-\frac{1}{m-1}\sum_{t_j\in S_{m-1}}I(w_j,w_i)] \tag{7}$$

基于 MRMR 的特征词选择算法如下。

算法 1 MRMR 特征词选择算法

输入：classDicList：类字典集
　　docList：文档集
　　tfidfDic：初次选择后特征集
　　selectNum：需求的特征数
输出：selectList：特征词集合
步骤：

```
1. new wordClassMIDic//存放单词和类别之间的互信息
2. //计算 tfidfDic 单词字典每一个单词和类别之间的互信息，降序排序
3. CalculateMIWordAndClass(tfidfDic,classDicList,setDic)
4. //声明 sum 数组存放单词和最后一个选择的单词 lastSelected-
   Word 的 MIij的累加和。
5. sum=new Dictionary(string,double)(wordClassMIDic. Count-1)
6. sum=[0]
7. 添加 wordClassMIDic 中值最大的单词到特征集合 selectList
8. string tempWord="";
9. while (featureList. Count<=selectFeatureNum)
10.   max=0. 0;
11.   For i=1:size(tfidfDic)
12.     MIij=Wordi 和 selectList 中最后一个入选的单词的互信息
13.     sum(i)=sum(i)+MIij;
14.     temp=wordClassMIDic(i)-(sum(i)/size(S));
15.     if(temp>max)
16.       tempWord=word1;
17.         max=temp;
18.     EndFor
19.   //添加 tempWord 到 featureList;
20.     featureList. Add(tempWord);
21.     //特征词集移出该单词
22.     wordClassMIDic. Remove(tempWord);
23.     //sum 中移出该单词对应的值
```

```
24.     sum. Remove(tempWord);
25.     //更换 lastSelectedWord
26.  EndWhile
```

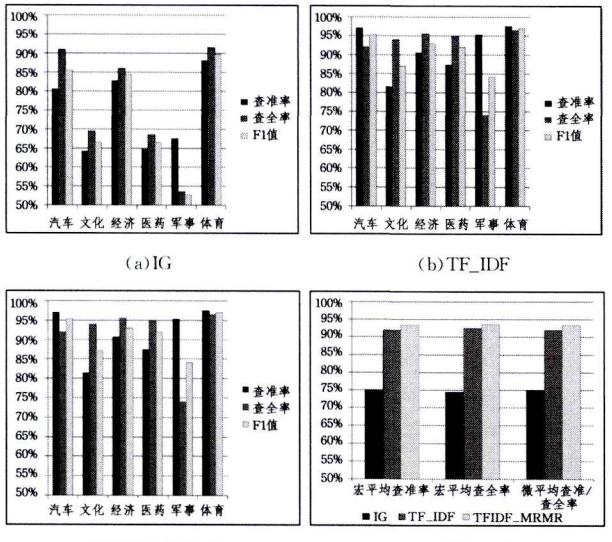
其中，classDicList 表示类字典集合，存放每一个类别的特征单词；docList 表示使用 tfidfDic 字典表示的文档集合；tfidfDic 表示 TFIDF 字典集合；selectNum 表示特征词的选择个数；selectList 表示选择的特征词集合。

3 实验分析

文本分类算法的性能评价指标[12,13]有查全率（也称召回率：Recall）、查准率（Precision）、F1-measure 以及用于评价全局性能的宏平均（Macro-average）和微平均（Micro-average）。本文使用 IG，TF_IDF，TFIDF_MRMR 3 种不同的方法进行特征选择，使用 Naïve Bayes 分类器对 3 类特征进行模型训练并分类，最后使用查全率、查准率和 F1-measure 对分类效果进行分析。

实验使用网易文本分类标准数据集[14]，其包含 6 个新闻类别，每个类别包含 600 篇新闻文本，采用交叉验证的方法进行实验对比。分别将每个类别的前 200 篇、中间 200 篇、后 200 篇作为测试集，各类别剩余的 400 篇作为训练集。

实验 1　选择各类别中编号为 0－199 的文本作为测试集，编号为 200－599 的文本作为训练集。使用 IG，TF_IDF，TFIDF_MRMR 3 种方法提取特征词，最后每个类别对应分类结果的查准率、查全率、F1 值如图 2 所示。



(a)IG　(b)TF_IDF

(c)TFIDF_MRMR　(d)性能对比

图 2　第一组实验分类结果

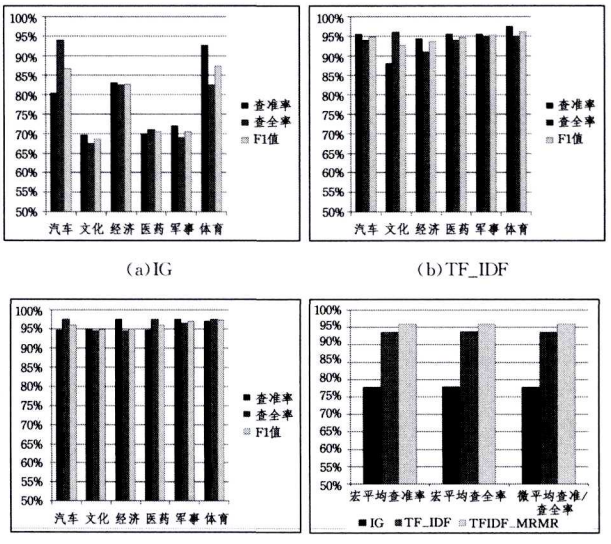
第一组的实验分类结果的评价参数如表 1 所列。

表 1　第一组实验分类评价

方法名称	宏查全率(%)	宏查准率(%)	微查准/查全率(%)
IG	75.00	74.55	75.00
TF_IDF	91.50	92.26	91.50
TFIDF_MRMR	93.49	93.41	93.49

第一组实验结果分析：通过 TF_IDF 方法提取特征词，使用 Bayes 方法进行文本分类的宏查全率、宏查准率、微查全率/微查准率均高于 IG 方法；TFIDF_MRMR 方法的分类评价参数都高于 IG 方法和 TF_IDF 方法，其中查准率提高了 1.15 个百分点，正确分类的文章增加了 14 篇。

实验 2　选择每个类别编号为 200－399 的文本作为测试集，编号为 0－199 和 300－599 的文本作为训练集。分别使用 IG，TF_IDF，TF_IDF_MRMR 3 种方法提取特征词之后进行文本分类，每个类别对应的查准率、查全率、F1 值如图 3 所示。



(a)IG　(b)TF_IDF

(c)TFIDF_MRMR　(d)性能对比

图 3　第二组实验分类结果

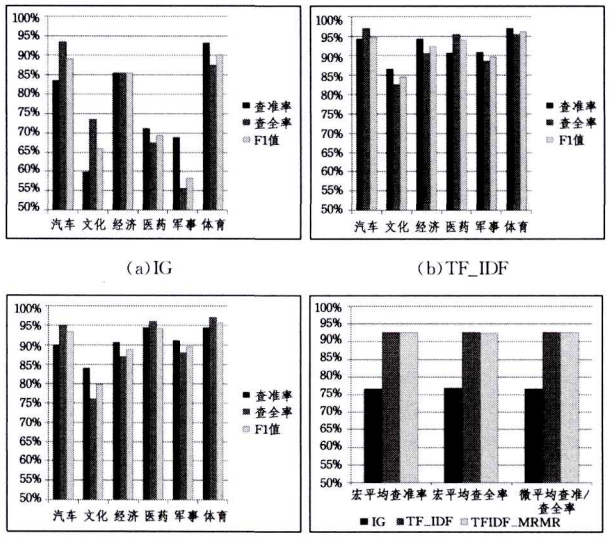
第二组实验的文本分类结果评价参数如表 2 所列。

表 2　第二组实验分类评价

方法名称	宏查全率(%)	宏查准率(%)	微查准/查全率(%)
IG	77.75	77.89	77.75
TF_IDF	93.58	93.80	93.58
TFIDF_MRMR	96.00	96.02	96.00

第二组实验结果分析：通过 TF_IDF 方法提取特征词，使用 Bayes 方法进行文本分类的宏查全率、宏查准率、微查全率/微查准率都高于 IG 方法；使用 TFIDF_MRMR 方法提取特征词之后，分类评价参数都高于 IG 方法和 TF_IDF 方法，其中查准率提高了 2.2 个百分点，正确分类的文章增加了 26 篇。

实验 3　选择各类别编号为 400－599 的文本作为测试集，编号为 0－399 的文本作为训练集。分别使用 IG，TF_IDF，TF_IDF_MRMR 3 种方法提取特征词之后进行文本分类，每个类别对应的查准率、查全率、F1 值如图 4 所示。



(a)IG　(b)TF_IDF

(c)TFIDF_MRMR　(d)性能对比

图 4　第三组实验结果

第三组实验结果分析：使用 TFIDF_MRMR 方法提取特征词时文本分类评价参数与 TF_IDF 方法基本持平，高于 IG 方法。第三组实验的文本分类结果评价参数如表 3 所列。

表 3　第三组实验分类评价表

方法名称	宏查全率(%)	宏查准率(%)	微查准/查全率(%)
IG	76.67	76.92	76.67
TF_IDF	92.50	92.50	92.50
TFIDF_MRMR	92.42	92.33	92.42

通过将上面三组实验的宏查全率、宏查准率等参数求平均值，得到使用 3 种方法提取特征词之后的文本分类评价参数。使用 TFIDF_MRMR 方法提取特征词，与 TF_IDF 方法进行比较，文本分类的评价参数均有所提高，宏查准率平均提高 1 个百分点，对应 12 篇文章；与 IG 提取特征词方法相比，宏查准率提高了 18 个百分点。通过上述分析说明使用 TFIDF_MRMR 方法提取特征词时提高了文本分类的宏查全率、宏查准率、微查全率/微查准率。

表 4　三组实验分类结果平均值表

方法名称	宏查全率(%)	宏查准率(%)	微查准/查全率(%)
IG	76.47	76.45	76.47
TF_IDF	92.52	92.85	92.50
TFIDF_MRMR	93.97	93.83	93.97

使用 IG，TF_IDF，TFIDF_MRMR 进行特征词提取时，文本分类性能对比如图 5 所示。

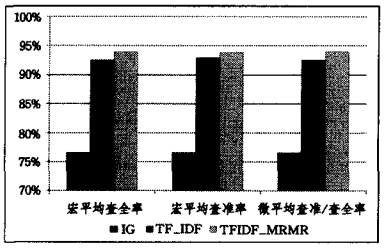
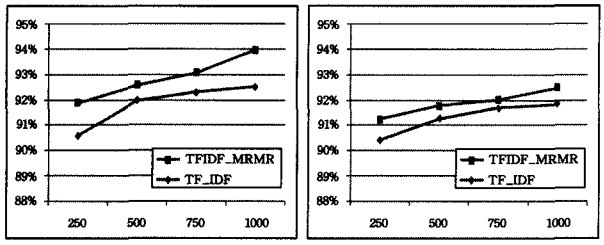


图 5　IG，TF_IDF，TFIDF_MRMR 分类性能对比

图 6 展示了在不同特征词纬度下，使用 TF_IDF 方法和 TFIDF_MRMR 方法进行特征选取，并使用朴素贝叶斯分类器分类结果的查准率和查全率的对比。通过图 6 可以看出，在不同的维度下使用 TFIDF_MRMR 方法进行特征词提取时文本分类的效果均优于 TF_IDF 方法。



(a)查准率对比　　(b)查全率对比

图 6　不同特征维度下查准率/查全率对比

结束语　特征选择是文本分类中一个重要的研究问题。传统的特征选择方法着重关注特征与类别的相关性，忽视了特征词间的冗余。本文通过结合 TF-IDF 和 MRMR 标准，提出了一种两阶段特征选择 TFIDF_MRMR 方法来消除特征词间的冗余，并采用贝叶斯分类方法进行分类实验。通过将 TFIDF_MRMR 特征选择方法和 IG 及 TF-IDF 特征选择方法进行对比，实验结果表明本文提出的特征词提取方法具有较好的分类效果。下一步工作将使用 SVM 和 Centroid-Based 等分类器验证并改进 TFIDF_MRMR 特征选择方法，此外，还将使用英文语料库验证本文方法的有效性。

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). 1997: 412-420

[2] Wang De-qing, Zhang Hui, Liu Rui, et al. t-Test feature selection approach based on term frequency for text categorization[J]. Pattern Recognition Letters, 2014, 45(11): 1-10

[3] Harun U. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm[J]. Knowledge-Based Systems, 2011, 24(7): 1024-1032

[4] Meng Jia-na, Lin Hong-fei, Yu Yu-hai. A two-stage feature selection method for text categorization[C]//2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, 2010: 1492-1496

[5] Kursat U A, Serkan G. A Novel Probabilistic Feature Selection Method For Text Classification[J]. Knowledge-Based Systems, 2012, 36(6): 226-235

[6] Kashif J, Sammen M, Babri Haroon A. A two-stage Markov blanket based feature selection algorithm for text classification [J]. Neurocomputing, 2015, 157: 91-104

[7] Lu Zhong-ning, Zhang Bao-wei. A text categorization method based on improved TF-IDF function[J]. Journal of Henan Normal University (Natural Science Edition), 2012, 40(6): 158-160 (in Chinese)
卢中宁，张保威. 一种基于改进 TF-IDF 函数的文本分类方法[J]. 河南师范大学学报(自然科学版)，2012，40(6)：158-160

[8] Carol F, Rindflesch T C, Milton C, et al. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine[J]. Journal of Biomedical Informatics, 2013, 46(5): 765-773

[9] Saleh S N, El-Sonbaty Y. A feature selection algorithm with redundancy reduction for text classification[J]. Computer and Information Sciences, 2007, 22: 1-6

[10] Lin Yao-jin, Hu Qing-hua, Liu Jing-hua, et al. Multi-label feature selection based on max-dependency and min-redundancy [J]. Neurocomputing, 2015, 168: 92-103

[11] Ding C, Peng Han-chuan. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205

[12] George F, Guyon I, Elisseeff A. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research, 2003, 3: 1289-1305

[13] Feng Guo-he. Review of Performance Evaluation of Text Classification[J]. Journal of Intelligence, 2011, 30(8): 66-70 (in Chinese)
奉国和. 文本分类性能评价研究[J]. 情报杂志，2011，30(8)：66-70

[14] NetEase article classification experiment data set[EB/OL]. http://www.datatang.com/data/11967 (in Chinese)
网易文章分类实验数据集[EB/OL]. http://www.datatang.com/data/11967