一种基于权重的文本特征选择方法

雷军程[1,2] 黄同成[2] 柳小文[2]
(长沙理工大学计算机与通信工程学院 长沙 410076)[1] (邵阳学院信息工程系 邵阳 422000)[2]

摘 要 在分析比较几种常用的特征选择方法的基础上,提出了一种引入文本类区分加权频率的特征选择方法TFIDF_Ci。它将具体类的文档出现频率引入TFIDF函数,提高了特征项所在文档所属类区分其他类的能力。实验中采用KNN分类算法对该方法和其他特征选择方法进行了比较测试。结果表明,TFIDF_Ci方法较其他方法在不同的训练集规模情况下具有更高的分类精度和稳定性。

关键词 特征选择,TFIDF,KNN分类算法
中图法分类号 TP391 **文献标识码** A

Improved Text Feature Selection Method Based on Text Feature Weight

LEI Jun-cheng[1,2] HUANG Tong-cheng[2] LIU Xiao-wen[2]
(Institute of Computer and Communication Engineering, Changsha University of Sciences and Technology, Changsha 410076, China)[1]
(Department of Information Engineering, Shaoyang University, Shaoyang 422000, China)[2]

Abstract This paper compared several feature selection methods in text categorization, and proposed a new feature selection method(TFIDF_Ci) based on weighted frequency of distinction between the text. It improves TFIDF function from weighted frequency and the feature items can increase the ability of text categorization in documents. In the experiment, we tested the effect of this feature selection method and other feature selection methods by using KNN classifiers. The experiments show the new method has good performance and stability under different numbers of training sets.

Keywords Feature selection, TFIDF, KNN classifiers

1 引言

随着计算机网络技术的发展,大量的信息资源在网上涌现,各种形式的文本文档以指数级增长。如何对网络上这些海量信息进行检索、分类已成为一个迫切需要解决的问题[1]。文本自动分类技术是数据挖掘和人工智能领域的一项实用的关键技术,它能够实现文本的自动分类。

文本分类领域最常见的向量表示方法是向量空间模型(VSM)[2],它依据文本内容来确定训练样本和向量维数,在对文本进行预处理(分词等)、特征选择[3]和权重计算之后,形成多维的空间向量,然后对未知文本进行预处理,进而判断它们属于事先定义类别中的哪一个或多个类别。在文本特征选择过程中,不是所有的词条都是有效的,利用特征选择方法可以降低文本特征的维数,从而简化分类计算难度,达到减少训练时间和提高分类精度的目的。因此必须要从大量的特征词条中选择出能够最好地代表文本特征的词条。特征选择是实现电子文本有效自动分类的前提和基础[4-6]。

针对特征提取,近年来国内外很多学者提出了一些方法,其可以总结为基于文档频率DF(Document Frequency)的方法、互信息MI(Mutual Information)方法、信息增益IG(Information Gain)方法、$\chi 2$统计CHI方法、期望交叉熵CE(Expected Cross Entropy)方法、文本证据权WT和优势率ODD等[14]。本文第2节针对几种常见特征选择方法进行对比分析,最后提出了一种基于特征权值函数TFIDF改进的文本类区分加权频率特征选择方法TFIDF_Ci。提出的方法将具体类中的文档频率作为一个新的权值引入到TFIDF函数中,从而提高包含特征词条的文档类的类别区分能力。实验结果表明,该方法在不同的训练语料集情况下具有更高的分类准确度和更好的稳定性。

2 常用的特征选择方法

2.1 文档频率

基于文档频率(Document Frequency,DF)的方法是最常见也最简单的特征选择方法之一,某一文本中词条的文档频率是指训练语料中包含该特征词条的所有文档数。文档频率方法有如下假设前提:设定一个频率阈值T,当DF值低于T时,那该词条就是低频词,它不具有类别区分信息。因此必须将该词条从原始文本特征向量中去除,从而降低特征维数,提高文本分类精度。文档频率方法比较简单,只是用词频作为唯一衡量标准。但是实际情况可能是频率较低的词条也包含了很多有用的分类信息,而频率高的词条却包含较少有用的分类信息[7]。而且在信息抽取技术领域的研究中,一般来说

到稿日期:2011-10-15 返修日期:2012-02-12 本文受湖南省教育厅基金项目(09C890)资助。
雷军程(1977—),男,硕士,讲师,系统分析师,主要研究方向为网络安全、算法;黄同成(1964—),男,博士,教授,主要研究方向为数字图像处理、计算机视觉、数据挖掘;柳小文(1978—),女,硕士,讲师,主要研究方向为信息安全。

词频较小的的词条具有更多的信息量，所以不应完全去除它们。

2.2 信息增益

基于信息增益(Information Gain, IG)的方法是人工智能研究方法中一种常见的特征选择方法，它根据某个词条在某文本文档中出现或者没有出现的次数来判断该文档类别，它的定义如下。

$$IG(t)=-\sum_{i=1}^{m}P(Ci)\log P(Ci)+P(t)\sum_{i=1}^{m}P(Ci|t)\log P(Ci|t)+P(\bar{t})\sum_{i=1}^{m}P(Ci|\bar{t})\log P(Ci|\bar{t}) \quad (1)$$

式中，$P(Ci|t)$表示文本中出现特征 t 时，文本属于 Ci 的概率；$P(Ci|\bar{t})$表示文本中不出现特征 t 时，文本属于 Ci 的概率；$P(Ci)$表示文档类别 Ci 的概率；$P(t)$表示特征项 t 在整个文本训练集中出现的概率。

信息增益方法的核心思想是，词条 t 的信息增益越大，该词条的贡献就越大，就越能提高分类精度。基于信息增益的方法是一种比较好的特征词条选择方法，然而它也有缺点，即没有考虑利用信息增益方法所选特征词条数较少情况时可能出现的数据稀疏问题，尤其是在文档类别分布和特征词条分布很不均匀的时候，绝大多数的词条并不会出现，因此信息增益值就由没有出现的词条项来决定，分类效果就不会提高，反而会很大程度地下降[8]。

2.3 期望交叉熵

基于期望交叉熵 (Expected Cross Entropy, CE)的方法也是一种基于特征统计的方法，这与信息增益方法相似。不同的是，它只统计在文本文档中出现的词条[9]。基于期望交叉熵的特征词条选择方法所获得的分类精度要高于信息增益的方法。期望交叉熵定义如下：

$$CE(t)=P(t)\sum P(Ci|t)\log\frac{P(Ci|t)}{P(Ci)} \quad (2)$$

式中的各项含义与前面信息增益公式中给出的一致。

2.4 互信息

基于互信息(Mutual Information, MI)的方法是一种标准，该标准可广泛用于建立词条特征项的统计模型，它的定义如下：

$$MI(t)=\sum P(Ci)\log\frac{P(t|Ci)}{P(t)} \quad (3)$$

$P(Ci)$表示第 i 类文本在训练文本集合中出现的概率，$P(t)$表示词 t 在训练文本集合中出现的概率，$P(t|Ci)$表示在第 i 类的文本中 t 出现的概率。互信息方法考虑了词条与类别之间的相关度，MI 值越高，说明特征项目与类别的贡献越大。缺点是互信息方法没有考虑文档中词条出现的频率，这就可能出现该方法只选择那些稀有词汇而不是高频词汇来作为特征向量的情况，最终会降低分类精度。

2.5 CHI

基于 CHI 统计量方法考虑了特征词条和类之间的相互依赖关系，它的定义如下：

$$\lambda^2(t,Ci)=\frac{N\cdot(AD-CB)^2}{(A+B)\cdot(B+D)\cdot(A+C)\cdot(C+D)} \quad (4)$$

A 表示词条 t 和类别 Ci 同时出现的次数(类 Ci 中包含特征 t 的文本的数目)，B 是词条 t 出现在类 Ci 以外的其他类中的次数(类 Ci 以外的其他类包含词条 t 的文本的数目)，C 是出现在类 Ci 但不含有词条 t 的次数(类 Ci 中不包含词条 t 的文本的数目)，D 是词条 t 和类 Ci 都不出现的次数(类 Ci 以外其他类中不包含词条 t 的文本的数目)，N 是训练预料中总的文本数。CHI 的值越大，说明词条和类之间的关联度(即依赖性)越大。缺点是该方法没有考虑词条在某一文档中的分布情况[10]。

3 TFIDF 方法

基于 TFIDF 的方法经常用于特征词条权值的计算，它是向量空间模型中经典的特征权值函数，在人工智能领域的应用非常广泛。

TFIDF 方法的核心思想是：如果一个词条在某篇文档中出现的频率很高，而且在其它文档中较少出现，认为该词条具有较好的类别区分度，那么该词条就可以用来分类[11]。它用词频乘以逆文档频率来表示词条权值，即：

$$W_{ik}=TF\cdot IDF=TF\cdot\frac{1}{DF}=f_{ik}\cdot\log\frac{N}{n_k} \quad (5)$$

式中，TF 称为词频(Term Frequency)，指词条 i 在文档 k 中出现的频率 f_{ik}，它用于计算该词描述文档内容的能力；IDF 称为逆文本频率(Inverse Document Frequency)，用于计算该词条区分文档的能力：

$$IDF=\log\frac{N}{n_k} \quad (6)$$

式中，N 为所有类别中的文档总数，n_k表示包含词条 i 的文档数量。

然而 TFIDF 方法也存在较明显的缺点，即它把整个文档语料库作为衡量目标，没有考虑到词条在同一类内和不同类间的分布情况，而文本选择方法的关键是衡量一个词条所具有的类区分能力。反文档频率简单地以文档频率较小的词条为重要词条，文档频率较大的词条不重要，而现实情况则可能是文档频率较大的词条也很重要。文献[12]也从词条权重和特征向量旋转的角度，说明了反文档频率的值不可能很好地反映词条的重要程度。

4 改进的 TFIDF 函数——TFIDF_Ci 函数

虽然 TFIDF 函数本身存在一定的缺点，即仅考虑了特征与文档之间的关系，没有反映特征与类别的关系，但是近年来也有不少人将特征选择函数应用于特征的权值计算，并取得了相当好的效果[13]。因此，本文将特征权值函数应用于特征选择，希望获得较好的效果。

将特征的类别信息引入函数，对特征权值函数进行改造。为了增加单词的类区分能力，在原 TFIDF 函数基础上增加一个新因子——文本类区分加权频率，该加权考虑某词条在某一具体类中的文档出现的频率，即词条 i 所在的文档的所属类与其他类的区分能力。

文本类区分加权频率 Ci 定义如下：

$$Ci=\frac{1}{n_k-m+1} \quad (7)$$

改进的 TFIDF_Ci 函数定义如下：

$$TFIDF\_Ci=TF\cdot IDF\cdot Ci=f_{ik}\cdot\log\frac{N}{n_k}\cdot\frac{1}{n_k-m+1} \quad (8)$$

式中，n_k表示包含词条 i 的文档数量，m 表示包含特征 i 最多的类中的文档数量。

文本类区分加权的内涵是当包含特征 i 的某一类中的文档数 m 大的时候，其他所有类中包含 i 的文档数(n_k-m)就小，那么特征 i 就能很好地代表包含 i 最多文档数的这个类的特征，因而该类加权的结果值就大。这也就意味着特征 i 的特征表达能力与 i 在除了最多类之外的其它类中的特征表达能力成反比。因此，特征 i 在某一类中越重要，文本类区分加权频率值也就越大。由于特征 i 可能完全被某一类内的文档所包含，这时 $m=n_k$，因此分母适当修正为 n_k-m+1。

这样，对特征项进行评估时，不是简单地采用将该特征项的术语频率乘以逆文档频率的方法，还考虑到了特征项所在文档的所属类与其他类的关系。无论文本分布的类再多，类分布再不均匀，TFIDF_Ci 方法都仅考虑类本身的影响因素，不增加其它的计算和分析难度，从而适合大规模文本特征选择，有很强的实用性。

5 实验及其分析

5.1 数据集

在实验中采用了复旦大学计算机信息与技术系国际数据库中心以及搜狗实验室的提供的部分中文语料库，从中选取了 7 个类，其中训练文档 14000 篇、测试文档 14000 篇，每个类分别有 2000 篇训练文档和 2000 篇测试文档。

5.2 分类器

由于本文研究的是文本特征选择方法的改进，因此对文本分类算法不做详细讨论，实验选取 KNN(K nearest neighbor)作为文本分类器。KNN 作为一种传统的模式识别方法，被广泛应用于文本分类研究中，其准确率和召回率表现出众。KNN 在已知类别样本中寻找与待分类样本 X 最相近的 K 个样本，文本样本之间的相似性可以通过文本向量之间的余弦来度量，定义如下：

$$\mathrm{sim}(X,Y)=\frac{(X,Y)}{|X|\cdot|Y|} \tag{9}$$

一种简单的预测规则就是将未知样本的类别预测在这 K 个最近邻样本包含最多实例的类别。实验中选取 K 值为 12，相似度阈值为 0.8，特征空间维数为 500 至 20000 不等。

5.3 实验结果及分析

为评价分类效果，通常采用召回率和准确率来评价。对于某一特定的类别，召回率 R(Recall 亦称查全率)定义为被正确分类的文档数和被测试文档总数的比例，即该类样本被分类器正确识别的概率；准确率 P(Precision，亦称查对率)定义为正确分类的文档数与被分类器识别为该类别的文档数的比率，即分类器做出正确决策的概率。综合考虑分类精度，全面衡量分类性能，实验使用宏 $F1$ 值[7]作为评价指标，宏 $F1$ 值越大，反映出分类效果越好。宏 $F1$ 值定义如下：

$$Macro\_F1=\sum_{i=1}^{m}\frac{N_i}{N}\times F1$$
$$=\sum_{i=1}^{m}\frac{N_i}{N}\times P(C_i|t)\frac{2\times precision_i\times recall_i}{precision_i+recall_i} \tag{10}$$

式中，N_i 为第 i 类的测试文档数，N 为测试文档总数。$precision_i$ 和 $recall_i$ 分别为第 i 类的准确率和召回率，共 m 个类别。

我们把本文改进的方法称为 TFIDF_Ci，与常用的特征选择方法、传统 TFIDF 方法、近年来相关学者提出的多个改进的 TFIDF 方法的分类效果在 500 到 20000 个不同特征维数中进行比较，表 1 为不同选择方法下 KNN 分类器的测试结果比较。图 1 为 TFIDF_Ci 方法和传统 TFIDF 方法分别结合 KNN 算法和遗传算法所得的测试结果比较。

表 1 多种 TFIDF 改进加权方法在 KNN 分类器上宏 F1 值的比较(%)

方法 / Number	DF	CE	MI	IG	TFIDF	基于信息熵改进的 TFIDF 方法	TDF	TF-IIDF-DIC	TFIDF_Ci
500	77.336	76.196	77.645	81.13	87.315	88.950	88.451	90.213	92.254
1000	77.514	78.317	77.887	81.978	88.374	91.259	88.945	91.66	91.763
2000	79.637	79.682	78.699	82.397	89.012	90.969	90.136	92.354	92.854
4000	80.3	80.358	79.124	83.88	90.457	90.739	89.854	92.961	92.350
6000	82.478	82.975	80.339	84.569	88.425	90.569	87.596	91.091	91.892
8000	78.791	80.341	82.947	85.02	89.689	91.303	89.355	90.8	92.806
10000	77.269	78.93	80.412	87.441	88.032	89.905	87.980	90.997	91.935
12000	79.96	77.59	79.698	84.596	86.452	90.124	86.265	90.322	89.982
15000	76.4	76.665	77.478	85.076	87.325	89.476	85.547	91.741	91.961
18000	75.79	76.897	78.012	84.127	84.570	92.875	84.663	91.425	91.742
20000	76.58	77.369	76.06	83.294	87.016	89.034	87.308	90.88	91.818
$\bar{x}$(平均值)	78.369	78.665	78.94	83.955	87.879	90.473	87.827	91.313	91.942
S(标准差)	2.041	2.038	1.869	1.735	1.616	1.15	1.772	0.831	0.760

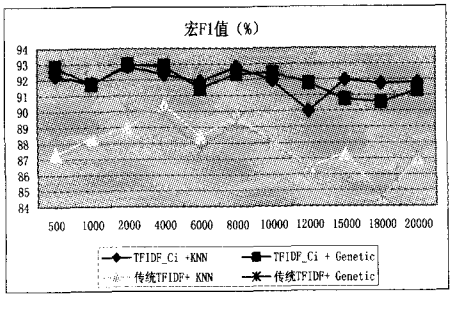

图 1 两种加权方法在不同的分类算法下的宏 F1 值比较

从表 1 可以看出，传统的 TFIDF 方法比常用的特征选择方法(如 DF，CE，MI，IG)在各特征维数上得到的宏 $F1$ 值均要高。TFIDF_Ci 结合 KNN 分类器在各特征维数上的宏 $F1$ 值均高于传统的 TFIDF 方法，平均高出 4.063%。当 TFIDF 在选择 4000 维特征时，获得的宏 $F1$ 值为 90.457，获得最好的分类性能，但其宏 $F1$ 值仍然比 TFIDF_Ci 方法的宏 $F1$ 值低约 2%。

改进的 TFIDF_Ci 方法的分类性能指标的标准差也小于传统 TFIDF 方法。这说明分类精度随着维数的变化能够保持相对平稳，利用 TFIDF_Ci 方法不仅提高了分类精度，而且在某种程度上降低了对特征维数的敏感性，这对于那些对特征维数敏感的分类器尤为有用。

改进的 TFIDF_Ci 方法和其他改进的 TFIDF 特征选择

(下转第 275 页)

ciation[C]// IEEE Workshop on Motion and Video Computing. 2008:1-8

[8] Yu Qian, Medioni G. Multiple-target Tracking by Spatiotemporal Monte Carlo Markov Chain Data Association[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009,31(12):2196-2210

[9] Ma Yun-qian, Yu Qian, Cohen I. Multiple Hypothesis Target Tracking Using Merge and Split of Graph's Nodes[J]. ISVC, 2006(1):783-792

[10] Xue Jian-ru, Zheng Nan-ning. Sequential Sampling Belief Propagation Algorithm in Multi-target Tracking[J]. Science in China, Ser. E, 2005,35(10):1049-1063

[11] Khan Z, Balch T R, Dellaert F. MCMC-based Particle Filtering for Tracking a Variable Number of Interacting Targets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(11):1805-1918

[12] Smith K, Gatica-Perez D, Odobez J M. Using Particles to Track Varying Numbers of Interacting People[C]// Proc. IEEE Conf. Computer Vision and Pattern Recognition. 2005:962-969

[13] Zhao Tao, Nevatia R. Tracking multiple humans in complex situations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004,7:1208-1221

[14] Zhao Tao, Nevatia R, Wu Bo. Segmentation and Tracking of Multiple Humans in Crowded Environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(7):1198-1211

[15] Forsyth D A, Ponce J. Computer Vision: A Modern Approach [M]. Prentice Hall, 2003:58-102

[16] Stauffer C, Grimson E. Learning Patterns of Activity Using Real time Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000,22(8):747-757

[17] Green P J. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination[J]. Biometrika, 1995, 82(4):711-732

[18] OTCBVS. Benchmark Dataset Web [EB/OL]. http://www.cse.ohio-state.edu/otcbvs-bench/

(上接第 252 页)

方法(如周炎涛提出的基于信息熵改进的 TFIDF 方法[7]、柴玉梅提出的基于术语频率和逆文档频率改进的 TDF 方法[15]、台德艺提出基于特征词分布集中度系数改进的 TF-IIDF-DIC 方法[15])比较,其宏 $F1$ 值也略高于其他改进 TFIDF 方法。

从图 1 可以看出,在不同的特征维数上 TFIDF_Ci 方法无论结合 KNN 算法还是遗传算法所得的宏 $F1$ 值都比传统 TFIDF 方法结合 KNN 或遗传算法的宏 $F1$ 值要高。这说明 TFIDF_Ci 方法都比传统 TFIDF 方法有更好的适应性,分类精度也更高。

实验表明,改进的 TFIDF_Ci 方法在训练集规模不同情况下是非常有效的,并且具有高稳定性,性能优于传统分类方法和其他部分改进 TFIDF 分类方法,尤其在类分布不均匀语料集上性能提高显著。同时该方法仅考虑类本身的影响因素,不增加其他的计算和分析难度,适合大规模文本特征选择,有很强的实用性。

结束语 为了解决传统 TFIDF 方法在处理分布类别少、类间区分能力不强的问题,本文通过增加文本类区分加权频率来提升特征词条所在的文档的所属类与其他文档类的区分能力,提出了 TFIDF_Ci 方法。实验表明,TFIDF_Ci 方法性能优于传统的 TFIDF 方法和多个改进的 TFIDF 方法,文本特征选择提升了文本分类精度。

参 考 文 献

[1] Jensen R, Shen Qiang. New Approaches to Fuzzy-Rough Feature Selection[J]. IEEE Transactions on Fuzzy Systems, 2009, 17(4):824-838

[2] Ma Yong-jun, Zhan Lin-qiang. Research on the Evaluation of Feature Selection Based on SVM[J]. Informatics in Control, Automation and Robotics, 2012,133(1):407-414

[3] Fan Wen-tao, Bouguila N, Ziou D. Unsupervised Hybrid Feature Extraction Selection for High-Dimensional Non-Gaussian Data Clustering with Variational Inference[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, PP(99):1

[4] Niu Shen, Hu Le-le, et al. Predicting protein oxidation sites with feature selection and analysis approach[J]. Journal of Biomolecular Structure and Dynamics, 2012,29(6):650-658

[5] Mehdi H A, Nasser G A. Text feature selection using ant colony optimization[J]. Expert Systems with Applications, 2009, 36(3):6843-6853

[6] Chen Hui-ling, Yang Bo, et al. A support vector machine classifier with rough set-based featureselection for breast cancer diagnosis [J]. Expert Systems with Applications, 2011,38(7):9014-9022

[7] 周炎涛,唐剑波,王家琴. 基于信息熵的改进 TFIDF 特征选择算法[J]. 计算机工程与应用,2007,43(35):156-158

[8] 孙德才,孙星明,张伟,等. 基于匹配区域特征的相似字符串匹配过滤算法[J]. 计算机研究与发展,2010,47(4):663-670

[9] 张玉芳,彭时名,吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程,2006,32(19):76-78

[10] 黄华军,谭峻珊,等. 基于高阶统计的网页隐秘信息检测研究[J]. 电子与信息学报,2010,32(5):1136-1140

[11] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,6(29):167-169

[12] 陆玉昌,鲁明羽,李凡,等. 向量空间中单词权重函数的分析和构造[J]. 计算机研究与发展,2002,39(10):1205-1210

[13] 林永民,吕震宇,等. 文本特征加权方法 TF·IDF 的分析与改进[J]. 计算机工程与设计,2008,29(11):2923-2925

[14] 柴玉梅,王宇. 基于 TFIDF 的文本特征选择方法[J]. 微计算机信息,2006,22(3):24-26

[15] 台德艺,王俊. 文本分类特征权重改进算法[J]. 计算机工程,2010,9(36):197-199