

一种改进的文本分类特征选择算法

陈建华,王治和,蒋 芸,许虎寅,樊东辉
(西北师范大学 数学与信息科学学院,甘肃 兰州 730070)

摘 要: 现有的一种特征选择算法 DPM(Discriminating Power Measure)^[1],是通过计算每个特征在某一类别和剩余其他类别中的文档频,比较了特征对一个类别和对其他类别的贡献,提取出具有强类别区分能力的特征词.在研究此特征选择算法的基础上,提出了一种改进的特征选择算法,该算法同时考虑了每个特征的类别频次在计算特征类别区分能力方面的重要性.经实验验证,改进后的特征选择算法能够获得较好的分类效果.

关键词: 特征选择;DPM;文档频;词频

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)12-0180-04

An Improved Feature Selection Algorithm in Text Classification

CHEN Jian-hua, WANG Zhi-he, JIANG Yun, XU Hu-yin, FAN Dong-hui

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

Abstract: By calculating the document frequency of each feature in a category and other categories, an existing feature selection algorithm-DPM(Discriminating Power Measure) compares the contribution of features for one category with other categories, and extracts the features that reveal larger differences among categories. On the basis of the research of DPM, a new feature selection algorithm is proposed, which considers the importance of word frequency of each feature in the calculation of discriminating power of features. Experiments demonstrate that the proposed algorithm has a better performance.

Key words: feature selection; discriminating power measure; document frequency; word frequency

1 引言

文本分类是信息检索领域中一个极为重要的子问题,特别是在半结构化网上文档信息极为丰富的情况下,以文本分类为基础的各种应用,如个人信息代理、搜索引擎、网上信息发布等,已成为有效控制和利用海量信息的重要手段.如今已有诸多分类技术和方法被提出来,例如支持向量机文本分类^[2]、贝叶斯概率分类^[3]、K 最近邻分类^[4]等.

文本分类面临的首要问题就是如何在计算机中合理地表示文本,这种表示方法既要包含足够的信息反映文本的特征,又不至于太过庞大使学习算法

无法处理.最常用的文本特征表示模型是向量空间模型(Vector Space Model, VSM)^[5].在该模型中,文本空间被视为一组正交词条向量所张成的向量空间.向量的维数往往是惊人的,包含噪声,且特征不明显.特征选择作为一种常用的特征降维方法,目的在于选择一个特征子集合,使得在这一特征子集合中可以得到最优的性能^[6].因此,在文本分类系统中,特征选择一直是一项关键技术和瓶颈.现有的特征选择方法主要有文档频率(DF)、词频(WF)、互信息(MI)等^[7].

文献[1]中提出的 DPM(Discriminating Power Measure)特征选择算法能够提取出有强类别区分

收稿日期: 2011-03-29; 修回日期: 2011-05-28

基金项目: 国家自然科学基金(60873196);甘肃省科技计划(甘肃省自然科学基金项目 1010RJZA022);西北师范大学 2010 年第三期知识与创新工程科研骨干项目(nwnu-kjcxgc-03-67)

能力的特征,在一定程度上改善了分类器的性能,但由于该算法忽略了特征在文档中出现的频率这一重要信息,因此本文在 DPM 的基础上提出了一种新的特征选择算法,并通过实验验证了改进后特征选择算法的优越性。

2 DPM 特征选择算法

DPM 算法的目标是提取出具有最强类别区分能力的特征,算法步骤如下所示:

步骤 1:假设共有 m 个类别,对每一个类别 $i(1 \leq i \leq m)$,考虑每个特征在类别 i 里的文档频以及在剩余其他类别里的文档频。

特征在第 i 个类别中的文档频计算公式和在除了第 i 个类别之外其他类别中的文档频计算公式分别为式(1)和式(2):

$$DF_i^{\text{in}} = \frac{n(\text{doc})(\text{term}_k)_{\text{all}}, \text{cat}_i}{n(\text{doc}_{\text{all}}, \text{cat}_i)} \quad (1)$$

$$DF_i^{\text{out}} = \frac{n(\text{doc})(\text{term}_k)_{\text{all}}, \text{collection-cat}_i}{n(\text{doc}_{\text{all}}, \text{collection-cat}_i)} \quad (2)$$

式中, term_k 表示第 k 个特征, cat_i 表示第 i 个类别, collection-cat_i 表示在文档集中除了类别 i 之外剩余的文档, $\text{doc}(\text{term}_k)_{\text{all}}$ 表示包含 term_k 的所有文档, doc_{all} 表示文档集中所有的文档, $n(\text{doc}(\text{term}_k)_{\text{all}}, \text{cat}_i)$ 表示在类别 cat_i 中包含 term_k 的所有文档数, $n(\text{doc}_{\text{all}}, \text{cat}_i)$ 表示类别 cat_i 中的所有文档数。同理, $n(\text{doc}(\text{term}_k)_{\text{all}}, \text{collection-cat}_i)$ 表示除了类别 cat_i 剩余其他类别中所有包含 term_k 的文档数, $n(\text{doc}_{\text{all}}, \text{collection-cat}_i)$ 表示除了类别 cat_i 剩余其他类别中所有的文档数。

步骤 2:计算 term_k 在类别 cat_i 内和类别 cat_i 外文档频的绝对差:

$$d_i = |DF_i^{\text{in}} - DF_i^{\text{out}}| \quad (3)$$

步骤 3:对每一个类别 $i(1 \leq i \leq m)$,计算关于 term_k 的 d_i 值,将所有类别的 d_i 值相加,作为特征 term_k 的 DPM 值为

$$\text{DPM} = \sum_i d_i \quad (4)$$

步骤 4:对特征集里的每一个特征计算相应的 DPM 值,如果某个特征的 DPM 值达到了一个事先给定的阈值,则留下该特征,否则删除。

实验证明,DPM 算法提高了分类器的精度,并且性能优于词频、互信息、Minimum aggregation、Entropy 等特征选择方法^[1]。但该算法只考虑特征发生的文档频而没有考虑单词发生的词频。如果某一特征只在某一类别的少量文档中频繁出现,则通

过 DPM 计算公式计算出来的 DPM 值很低,在特征选择时这种特征就会被排除掉,但这种在少量文档中频繁出现的特征很有可能对分类的贡献很大。并且假设特征 t_1 和 t_2 在所有类别中的文档频相同,那么该算法认为这两个特征的贡献是相同的,而忽略了它们在文档中出现的次数,这显然比较片面。因此,在改进的算法中既考虑了特征的文档频又考虑了特征的类别频次。

3 DPM 算法的改进

3.1 特征的类别频次

特征 term_k 在第 i 个类别中的词频计算公式为

$$WF_i^{\text{in}} = \frac{n(\text{term}_k, \text{cat}_i)}{n(\text{term}_{\text{all}}, \text{cat}_i)} \quad (5)$$

特征 term_k 在除第 i 个类别之外剩余其他类别中的词频计算公式为

$$WF_i^{\text{out}} = \frac{n(\text{term}_k, \text{collection-cat}_i)}{n(\text{term}_{\text{all}}, \text{collection-cat}_i)} \quad (6)$$

式中, $n(\text{term}_k, \text{cat}_i)$ 表示类别 cat_i 中 term_k 出现的次数, $n(\text{term}_{\text{all}}, \text{cat}_i)$ 表示类别 cat_i 中所有特征词个数, $n(\text{term}_k, \text{collection-cat}_i)$ 表示除了类别 cat_i 剩余其他类别中 term_k 出现的次数, $n(\text{term}_{\text{all}}, \text{collection-cat}_i)$ 表示除了类别 cat_i 剩余其他类别中所有的特征词个数。

3.2 新的特征选择算法描述

步骤 1:假设有 m 个类别,对每一个类别 $i(1 \leq i \leq m)$,计算每个特征 term_k 在类别 i 里的文档频、在剩余其他类别里的文档频,在类别 i 里的类别频次以及在剩余其他类别里的类别频次。其中特征文档频的计算用式(1)和式(2),类别频次的计算用式(5)和式(6)来完成。

步骤 2:计算 term_k 在类别 cat_i 内和类别 cat_i 外贡献的差值:

$$d_i = (DF_i^{\text{in}} \times WF_i^{\text{in}} - DF_i^{\text{out}} \times WF_i^{\text{out}})^2 \quad (7)$$

步骤 3:对每一个类别 $i(1 \leq i \leq m)$,计算关于 term_k 的 d_i 值,将最大的 d_i 值作为特征 term_k 的 DPM 值:

$$\text{DPM} = \max_{1 \leq i \leq m} (d_i) \quad (8)$$

步骤 4:对特征集合里的每一个特征计算相应的 DPM 值,如果某个特征的 DPM 值达到了一个事先给定的阈值,则留下该特征,否则删除。

从式(7)可以看出,引入特征的类别频次后,由特征的文档频和类别频次两个因素同时决定该特征对分类的贡献大小,下面通过例子说明改进后的

DPM 算法相对于 DPM 算法的优越性.

3.3 算法实例

假设有 A、B、C、D 四个类别,每个类别均有 10 篇文档,且每个类别的特征总数均为 1 000. 现有 6 个候选特征词 f_1 、 f_2 、 f_3 、 f_4 、 f_5 、 f_6 ,它们在每个类别里的文档频和词频如表 1 所示. 其中 WDP 表示文档频,CP 表示词频,DPM 和 Improved-DPM 分别为根据式(4)和式(8)计算出的 6 个特征词对分类的贡献.

表 1 DPM 算法和改进后 DPM 算法特征选择比较

候选特征词	A		B		C		D		DPM	Improved-DPM
	WDP	CP	WDP	CP	WDP	CP	WDP	CP		
f_1	2	2	1	1	1	2	2	2	0.266 67	5.444×10^{-8}
f_2	2	2	1	1	0	0	0	0	0.400 00	1.512×10^{-7}
f_3	2	60	1	1	0	0	0	0	0.400 00	1.437×10^{-4}
f_4	0	0	0	0	1	20	0	0	0.199 99	4.000×10^{-6}
f_5	0	0	2	60	0	0	0	0	0.400 01	1.440×10^{-4}
f_6	5	10	0	0	0	0	1	1	0.933 34	2.489×10^{-5}

为比较 DPM 算法和优化后的 DPM 算法所选特征词的差异,按两个算法所计算出的贡献从大到小对六个特征词进行排序. DPM 算法选择的特征词的优先顺序为: f_6 、 f_5 、 f_2 和 f_3 、 f_1 、 f_4 . 改进后 DPM 算法选择的特征词的优先顺序为: f_5 、 f_3 、 f_6 、 f_4 、 f_2 、 f_1 . 从表 1 和排序结果可得到如下结论:

- (1) 特征 f_1 几乎没有类别区分能力,因此它的贡献应该是最小的. 特征 f_6 有较强的类别区分能力,因此所得到的它的贡献应该是较大的. 排序结果说明改进后的 DPM 算法能较好地选择有强类别区分能力的特征,也能排除没有类别区分能力的特征.
- (2) f_2 和 f_3 在所有类别中的文档频是相同的,因此用 DPM 算法得出的结果是 f_2 与 f_3 有相同的类别区分能力,但在类别 A 中的类别频次差别很大,很明显 f_3 对分类的贡献大于 f_2 对分类的贡献. 改进后的 DPM 算法得到的结果很好体现了这一点.
- (3) f_4 和 f_5 代表仅在某一类别的少量文档中频繁出现的特征,对分类的贡献是很大的,但是用 DPM 算法进行特征选择时将 f_4 排到了最后,而改进后的 DPM 算法则较好地弥补了这一缺陷.

4 特征选择算法实验验证

4.1 实验设计

实验语料库选用 TanCorpV1.0 中文文本分类语料库^[8],该语料库由谭松波博士收集整理. 实验选取其中的六个类别,依次是财经、教育、科技、人才、体育、艺术,共 2 100 篇文档,每个类别中文档数均为 350 篇,文档集在去除停用词后共 38 897 个不同的特征词.

为了评估特征选择算法的有效性,实验在数据集上使用 Weka 软件中 NaïveBayes 算法的十折交叉验证法对六个类的文档进行测试. 十折交叉验证法是将数据集分成十份,轮流将其中 9 份作为训练集,1 份作为测试集进行实验,每次实验都会得出相应的准确率,十次结果的准确率的平均值即为对算法精度的估计. 实验在分别运行 DPM 算法、文档频方法以及改进的 DPM 算法后比较对应性能的变化.

4.2 实验结果比较

用 NaïveBayes 作为分类器,TF-IDF 公式计算特征权重,在最好的情况下 DPM 算法、文档频方法以及改进后的 DPM 算法在每个类别中的分类准确率和召回率如表 2 所示.

表 2 最好情况下准确率和召回率的比较

类别	准确率/%			召回率/%		
	Improved-DPM	DPM	DF	Improved-DPM	DPM	DF
财经	91.2	90.7	89.4	88.3	89.4	86.9
教育	83.5	82.6	76.9	80.9	77.4	77.1
科技	95.9	95.3	95.2	93.4	92.6	90.3
人才	86.6	85.0	83.7	92.3	90.9	89.4
体育	98.0	98.2	97.1	98.0	96.0	94.3
艺术	90.2	87.1	85.4	92.3	92.3	88.6
平均	90.9	89.8	87.9	90.9	89.8	87.8

具体看来,不同的特征选择算法在不同的数据集上各有优势,但总体来说,改进后的 DPM 特征选择算法在分类性能上优于 DPM 算法和文档频方法。

比较在不同的特征数目下分类器的性能. 实验结果如表 3 所示.

表 3 特征维数变化下的 F_1 测试值比较

特征维数	DPM	DF	Improved-DPM
50	0.828	0.539	0.827
100	0.830	0.712	0.841
300	0.842	0.754	0.869
500	0.871	0.796	0.886
800	0.886	0.835	0.900
1 000	0.887	0.859	0.898
1 200	0.889	0.862	0.903
1 500	0.889	0.869	0.901
1 800	0.895	0.872	0.903
2 000	0.897	0.873	0.903
2 500	0.894	0.877	0.908
3 000	0.890	0.877	0.907
3 500	0.893	0.878	0.904
4 000	0.887	0.876	0.900
4 500	0.885	0.874	0.906
5 000	0.889	0.873	0.905

在特征维数变化的情况下,考察分类器的性能变化情况,可以很好地反映一个分类器对数据样本变化的敏感程度. 从表 3 可以看出,随着特征维数从小到大变化,分类器的 F_1 测试值也随着特征个数的增加而升高,并将到达一个相对稳定的水平. 同时改进后 DPM 的性能并没有因维数的变化而出现较大的波动,因此该算法具有比较稳定的性能. 从表 3 还可以看出,改进后 DPM 的 F_1 测试值在特征个数为 2 500 的时候达到最高,为 0.908,以后基本维持在 0.904 左右. DPM 和 DF 的 F_1 测试值分别在特征维数为 2 000 和 3 500 的时候达到最高,为 0.897 和 0.878,以后分别基本维持在 0.890 和 0.874 左右. 对比可知,改进后的 DPM 具有较好的特征选择效果,优于其他两种特征选择算法. 这是由于改进的 DPM 在选取特征时,由单词的文档频率和类别频次两个因素同时决定特征的类别区分能力,克服了只有一种因素决定的片面性,取得了较好的效果,因此对 DPM 算法的改进是有效的.

5 结束语

首先分析了 DPM 特征选择算法,针对其对低

文档频的特征不可靠这一不足,提出了一种改进的算法. 改进的 DPM 算法不但考虑了特征的文档频,而且还考虑了特征的类别频次. 经过文本分类系统实验验证,改进后的 DPM 特征选择算法性能要好于 DPM 算法.

参考文献:

[1] Chen Chih-Ming, Lee Hahn-Ming, Chang Yu-Jung. Two novel feature selection approaches for web page classification [J]. Expert Systems with Application, 2009(36): 260—272.

[2] Joachims T. Text categorization with support vector machines; learning with many relevant features [C]//proceedings of the 10th European Conference on Machine Learning. New York: Springer, 1998: 137—142.

[3] Lewis D D. Naive(Bayes) at forty; the independence assumption in information retrieval [C]//proceedings of the 10th European Conference On Machine Learning. New York: Springer, 1998: 4—15.

[4] Yang Y, Liu X. A re-examination of text categorization methods [C]//proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval. New York: ACM Press, 1999.

[5] 胡强. 基于两种特征影响度的特征选择[J]. 微电子学与计算机, 2010, 27(12): 65—69.

[6] Yiming Yang, Jan O. Pedersen. A comparative study on feature selection in text categorization [G]//ICML 1997: 412—420.

[7] 陈平, 刘晓霞, 李亚军. 文本分类中改进型互信息特征选择的研究 [J]. 微电子学与计算机, 2008, 25(6): 194—197.

[8] 谭松波, 王月粉. 中文文本分类语料—TanCorp V1. 0 [EB/OL]. [2011—03—29]. [http:// www. searchforum. org. cn/tansongbo/corpus. html](http://www.searchforum.org.cn/tansongbo/corpus.htm).

作者简介:

陈建华 女, (1988—), 硕士研究生. 研究方向为数据挖掘.

王治和 男, (1965—), 教授. 研究方向为数据挖掘.

蒋芸 女, (1970—), 副教授, 博士. 研究方向为数据挖掘、粗糙集.

许虎寅 男, (1986—), 硕士研究生. 研究方向为数据挖掘.

樊东辉 男, (1978—), 硕士研究生. 研究方向为数据挖掘.