基于类别相关的新文本特征提取方法*

林少波, 杨　丹, 徐　玲
(重庆大学 软件学院, 重庆 400030)

摘　要: 为了避免文本特征提取过程中负相关特征与弱相关特征产生的干扰,提出一个新的基于类别正相关和强相关(SP)的特征提取方法。通过结合正相关性因子和强相关因子,SP 方法能够有效地区别特征与类别正负相关性和强弱相关程度,通过优先选择正相关和强相关特征,避免了负相关和弱相关特征的干扰,从而有效地提取高质量的文本特征。实验结果表明,该方法具有强降维能力和良好的分类效果。

关键词: 正相关; 强相关; 文本分类; 特征降维; 特征提取
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)05-1680-04
doi:10.3969/j.issn.1001-3695.2012.05.021

New approach to feature selection for text categorization using class correlation

LIN Shao-bo, YANG Dan, XU Ling
(*School of Software Engineering*, *Chongqing University*, *Chongqing* 400030, *China*)

Abstract: This paper proposed a new approach of feature selection for text categorization, which was based on the strong class correlation and positive class correlation, named SP. SP could eliminate the effect of negative and poor correlation feature effectly. SP discriminated between the positive feature and the negative feature by positive correlation factor, and eliminated the effect of negative feature. SP discriminated between the strong class correlation of features and the poor class correlation of features by positive class correlation factor, and eliminated the effect of poor correlation feature. SP could select high quality features effectively by combining these two factors. The result of Experiment indicates that the proposed approach has a good performance at categorization and reducing high dimensional feature space.

Key words: positive correlation; strong correlation; text classification; feature reduction; feature selection

0　引言

文本自动分类是在给定分类体系的情况下,根据预定义类别中的文本数据信息总结归纳出分类的规律性,根据这个规律性将未标明类别的文本映射到预定义的类别[1]。

现有文本分类技术通常有两种文本表示模型,即布尔模型和向量空间模型[2]。向量模型中特征向量与文本集合中的文本一一对应,将文本集合中的词条作为向量中的特征项。布尔模型可以看做是向量模型的一种特例,根据特征项在文档中出现与否,特征项的权值只能取 1 或 0。布尔模型不能很好地体现文本特征的重要程度,通常情况下布尔模型的效果不如其他模型。而向量空间模型是一种不考虑词与词之间的上下文关系、出现的顺序和位置以及文本长度的词袋(bag of words)文本表示模型。故本文选择向量空间模型作为文本表示模型。

针对文本分类中文本特征向量空间的高维性问题,采用特征提取方法进行特征降维。该方法是基于过滤模型[3]的一种特征选择方法,过滤模型的基本思想是根据训练集数据的一般特性进行特征选择,通过构造某种特征评价函数,来统计文本特征空间中各个特征项的值,将各个特征项按其特征值排序,并根据设置的阈值选择出合适规模的对文本分类贡献较大的特征子集,以达到降低特征空间维度的目的。在特征选择过程中不包括任何学习算法。特征提取通过对特征向量空间的降维,不仅提高了分类速度,而且过滤了噪声数据,在提高精度的同时还有助于解决过拟合问题。

现有的特征提取方法包括文档频数(document frequency, DF)、信息增益(information gain, IG)、互信息(mutual information, MI)、χ^2 统计(Chi-square statistic, CHI)和 CC(correlation coefficient)等[4~8]。目前 CHI 和 IG 性能较好,DF 和 MI 性能较差[7]。CC 是基于相关系数的特征提取方法,是 CHI 公式的开平方式。两者的区别是,CHI 评价函数计算得到特征值具有非负性,等同了正相关和负相关特征的对文本分类的重要性,而 CC 通过对 CHI 公式的开平方,使得特征值具有正负性,区别了正相关和负相关特征对文本分类的重要性[8]。文献[9]使用选择带有较强类别信息的特征词,在线性分类器上有较好的准确率。文献[10]指出对于文本分类而言,特征的重要性主要由特征的正相关能力决定。过多地考虑特征项与类别的负相关度,不仅不能提高分类的精度,还可能对分类结果产生干扰。

本文提出了一种基于类别正相关和类别强相关的新特征提取方法 SP。该方法只选取与类别正相关和强相关的特征。

收稿日期: 2011-10-20; **修回日期**: 2011-11-30　　**基金项目**: 国家自然科学基金资助项目(60975015); 重庆市重点攻关资助项目(CSTC2009AB2230)

作者简介: 林少波(1986-),男,福建莆田人,硕士研究生,主要研究方向为文本分类、企业信息化及电子政务(linshaobo123@126.com);杨丹(1962-),男,重庆云阳人,教授,博导,博士,主要研究方向为图像处理、模式识别、机器学习;徐玲(1975-),女,安徽庐江人,讲师,博士,主要研究方向为数据挖掘、图像处理.

首先给出特征与类别相关性和相关度的概念,通过对 CHI 不能区别正相关特征与负相关特征重要程度的原因分析,将文本特征与类别正、负相关特征的重要性差异度作为正相关性因子,用来区别特征与类别正负相关性。通过对文本特征在文本集合中分布规律的分析,将文本特征类内和类间的相关度指标结合成为强相关度因子,用来区别特征与类别的强弱相关度。

1　特征与类别的相关性

特征与类别的相关性包括正相关性和负相关性。为了更好地描述特征与类别的正负相关性概念,这里将通过对文本集合的一般特性的分析,具体地描述特征与类别的正负相关性。文本集合这种特性可以由表 1 中四种基本元素 A、B、C、D 表示。为进一步解释,现作以下假设:在一个文本集合 H 中,假设存在一个类别 c_i,集合中除 c_i 以外的类别表示为$\overline{c_i}$,则按照文本所属类别划分,该集合 H 可以表示为$\{c_i,\overline{c_i}\}$。假设存在一个文本特征词 t_k,集合中除 t_k 外的特征词表示为$\overline{t_k}$,则按照是否包括特征词 t_k 划分,该集合 H 可以表示为$\{t_k,\overline{t_k}\}$。按照文本是否包括特征词 t_k,是否属于类别 c_i,可将文本集合 H 表示为$\{c_i,\overline{c_i}\}$和$\{t_k,\overline{t_k}\}$的笛卡尔乘积。该集合由表 1 所示的四个信息元素 A、B、C、D 构成。

表 1　信息基本元素表

	c_i	$\overline{c_i}$
t_k	A_i	B_i
$\overline{t_k}$	C_i	D_i

表 1 中,$A_i=(t_k,c_i)$表示不仅包含特征 t_k 而且属于类别 c_i 的文本数量;$B_i=(t_k,\overline{c_i})$表示包含特征 t_k 但是不属于类别 c_i 的文本数量;$C_i=(\overline{t_k},c_i)$表示不包含特征 t_k 但是属于类别 c_i 的文本数量;$D_i=(\overline{t_k},\overline{c_i})$表示不包含特征 t_k 并且不属于类别 c_i 的文本数量。文本总数量 $N=A+B+C+D$。

基于表 1 的描述,将特征选择方法根据特征与类别的相关性进行分类,包括正相关性和负相关性[11]。

定义 1　若特征项 t_k 当且仅当只出现在与类别 c_i 相关的文本中,则称特征项 t_k 与类别 c_i 正相关。

定义 2　若特征项 t_k 当且仅当只出现在与类别 c_i 不相关的文本中,则称特征项 t_k 与类别 c_i 负相关。

表 1 中 $A_i=(t_k,c_i)$和 $D_i=(\overline{t_k},\overline{c_i})$描述的是特征项 t_k 和类别 c_i 的正相关程度,$B_i=(t_k,\overline{c_i})$和 $C_i=(\overline{t_k},c_i)$描述的是特征项 t_k 和类别 c_i 的负相关程度。

2　特征与类别的相关度

特征与类别的相关度可分为强相关和弱相关。在此结合本文第 1 章的文本集合一般性描述来进一步阐述特征与类别的相关度概念。特征的相关度是衡量一个特征代表一个类别的程度。特征与类别的相关度评价可以基于以下两种指标:

a)若存在一个特征词,该特征词越集中出现在当前所在类别文档中,在其他类别文档中越少出现,那么这个特征词越能区别所在类别与文本集合中的其他类别。即该特征词越能代表当前类别,对分类贡献越大。这里将衡量类别间特征分散度的指标称为相关度的类间指标。

b)若存在一个特征词,在当前所在类别内,包含该特征词的文档在该类别中比例越高越能代表当前类别,对分类贡献越大。将衡量类内特征分散度的指标称为相关度的类内指标。

在此采用两种比值表示相关度的类间指标和类内指标:

a)$\frac{A_i}{B_i}$。如果特征词 t_k 只与类别 c_i 相关度高,$\frac{A_i}{B_i}$值越高,那么说明特征词 t_k 能很好地代表类别 c_i,那么$\frac{A_i}{B_i}$的值就高。

b)$\frac{A_i}{C_i}$。在类别 c_i 中有两个特征词 t_k 和 t_m,它们各自对应一个$\frac{A_i}{C_i}$值,那么$\frac{A_i}{C_i}$值越大的特征词越能代表类别 c_i。

鉴于以上概念的提出,结合类间指标和类内指标来刻画特征与类别的相关程度,区别强、弱相关特征。将特征与类别的强相关度因子 SCD(strong correlation degree)表示为

$$\mathrm{SCD}(t_k,c_i)=\frac{A_i}{B_i}\frac{A_i}{C_i} \tag{1}$$

若一个特征与一个类别相关度越强,则 SCD 值越大。SCD 优先选择与类别强相关的特征,从而能够有效地避免弱相关特征产生的噪声干扰。

3　SP 文本特征提取方法

基于以上文本特征与类别的相关性和相关度的概念的提出,采用强相关度因子 SCD 来区别特征与类别相关度的强弱,并对 CHI 无法体现正相关特征与负相关特征的重要性差异度的原因进行分析,寻求特征与类别间的正负相关性的具体表示方法。

CHI 统计方法是在数理统计中一种常用的检验两个变量独立性的方法。CHI 最基本的思想就是通过观察实际值与理论值的偏差来确定理论的正确与否。其运用在文本特征选择中,假设 t_k 与类别 c_i 之间是独立的[12]。这种独立关系类似于具有一维自由度的χ^2 分布,t_k 对于 c_i 的 CHI 统计量为

$$\chi^2(t_k,c_i)=\frac{N(A_iD_i-B_iC_i)^2}{(A_i+C_i)(B_i+D_i)(A_i+B_i)(C_i+D_i)} \tag{2}$$

CHI 统计方法用$\chi^2(t_k,c_i)$来度量特征项 t_k 和类别 c_i 之间的相关程度。$\chi^2(t_k,c_i)$值越大说明特征项 t_k 和类别 c_i 相关程度越高,此时特征项 t_k 所包含的与类别 c_i 相关的信息就越多。$\chi^2(t_k,c_i)=0$ 表示特征项 t_k 与类别 c_i 相互独立。特征项 t_k 与类别 c_i 的相关性越强,$\chi^2(t_k,c_i)$的值就越大。特征性选择的过程就是计算特征项 t_k 和类别 c_i 的$\chi^2(t_k,c_i)$值,然后按值大小排序,根据实际需要选取 CHI 值大的特征项。

由定义 1 和定义 2 可知,A_i 和 D_i 描述的是特征项 t_k 和类别 c_i 的正相关程度,B_i 和 C_i 描述的是特征项 t_k 和类别 c_i 的负相关程度。由概率论可知,假设 A_i 的值越大,D_i 的值越大,即特征 t_k 出现在 c_i 类的文本里面的概率越大,则 t_k 与 c_i 正相关度越大。假设 B_i 的值越大,C_i 的值越大,那么特征 t_k 出现在类别$\overline{c_i}$的概率越大。

$A_iD_i-B_iC_i$ 体现了特征 t_k 与类别 c_i 的正、负相关特征重要性差异度。若 $A_iD_i-B_iC_i>0$,则特征 t_k 与类别 c_i 体现出正相关性;若 $A_iD_i-B_iC_i<0$,则特征 t_k 与类别 c_i 体现出负相关性。由式(2)可知,分子中$(A_iD_i-B_iC_i)^2$ 使得 $A_iD_i-B_iC_i$ 呈现

非负性，若特征 t_k 与类别 c_i 体现出正相关度越大，那么 $A_iD_i - B_iC_i$ 必然为正值，且值将随正相关度的增加而增大，CHI 得出的特征值也将增大。若特征 t_k 与类别 c_i 体现出负相关性，那么 $A_iD_i - B_iC_i$ 必然为负值，而且值随着负相关度的增加而减小，但是 $(A_iD_i - B_iC_i)^2$ 的值将随之变大，CHI 得出的特征值也将变大。这样正、负相关特征的重要性差异度便无法体现。

通过以上分析，正、负相关特征重要性差异度能很好地刻画文本特征与类别的正负相关性。假定用正相关性因子 PC (positive correlation)来区别特征与类别间的正负相关性，则

$$\mathrm{PC}(t_k, c_i) = A_iD_i - B_iC_i \tag{3}$$

新文本特征提取方法的目的是选择与类别正相关并且强相关的特征词。综上所述，提出新的文本特征提取方法 SP (SCD&PNC)：

$$\mathrm{SP}(t_k, c_i) = \mathrm{SCD} \times \mathrm{PC} = \frac{A_i}{B_i}\frac{A_i}{C_i}(A_iD_i - B_iC_i) = (\frac{A_iD_i}{B_iC_i} - 1)A_i^2 \tag{4}$$

若一个特征词与类别正相关并且强相关，那么 SP 值一定越大，故 SP 优先选择与类别正相关并且强相关的特征词。

文本特征提取策略可以分为局部和全局特征提取。文献[12]指出在均衡数据中，全局特征提取优于局部特征提取，单独利用局部特征提取的缺点是它不能有效地选取到能代表所有类别的特征集合，忽视了全体训练样本和类别的整体性。文献[13]指出在偏斜数据集中使用加权局部特征选择优于全局特征，因为偏斜数据中各个类别间文档数量差距很大，基于全局提取的特征集合不能代表少数类。本文使用均衡数据集，采用全局特征提取策略。

基于全局特征提取策略的特征提取方法，需获取特征项 t_k 对整个文档集的全局特征值时，由于 SP 是基于类别正相关性与强相关性的特征提取方法，采用式(5)基于最大值的方式计算全局特征，度量各个特征对于分类的重要性，能够选择出对某一个类具有较好标志作用的特征项。

$$\mathrm{SP}_{\max}(t_k) = \max_{i=1}^{m}\{\mathrm{SP}(t_k, c_i)\} \tag{5}$$

SP 特征提取方法的优点在于：计算复杂度低，只选择与类别正相关并且强相关的特征，避免了弱相关特征和负相关特征的干扰。

4 实验结果及分析

本实验采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组的文本分类语料库[1]。该语料库共分为 20 个类别，测试集共 9 804 篇文章，训练集 9 804 篇文章。本文从中抽取 9 个类别作为训练集，包括艺术、历史、航空、计算机、环境、农业、经济、政治、体育。在原训练集中每个类别随机抽取 200 个训练文档，共计 1 800 个训练文档。在原测试集中每个类别随机抽取 100 个测试文档，共计 900 个测试文档。训练文档和测试文档的比例为 2∶1，通过预处理，测试文档包含 57 771个特征词。

实验采用 KNN 分类算法，实验效果的评价指标采用 Marco-Recall、Marco-Precision、和 Marco-F1。式(6) ~ (8)分别表示宏平均查全率、宏平均查准率、宏平均 F1 值。其中 recall_i 表示类 c_i 查全率，$\mathrm{precision}_i$ 表示类 c_i 查准率，$\mathrm{F1}_i$ 表示类 c_i 的 F1 值，$|C|$表示文本集合中类别总数。实验将 SP 方法与传统方法 CHI、CC、DF 进行比较。

$$\text{Marco-Recall} = \frac{\sum_{i=1}^{|C|}\mathrm{recall}_i}{|C|} \tag{6}$$

$$\text{Marco-Precision} = \frac{\sum_{i=1}^{|C|}\mathrm{precision}_i}{|C|} \tag{7}$$

$$\text{Marco-F1} = \frac{\sum_{i=1}^{|C|}\mathrm{F1}_i}{|C|} \tag{8}$$

如图 1 ~ 3 所示，CC 和 CHI 在提取的文本特征向量维数为 1 500 维时，分类效果最优；SP 在提取的文本特征向量维数为 1 500 维时，分类效果最优；DF 在提取的文本特征向量维数为 3 000 维时，分类效果最优。CC、CHI、SP 在分类效果达到最优时，Marco-Recall、Marco-Precision、Marco-F1 三项评价指标数值相当，DF 效果最差。

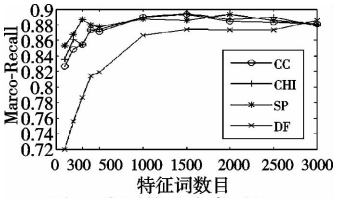

图1 宏平均查全率对比

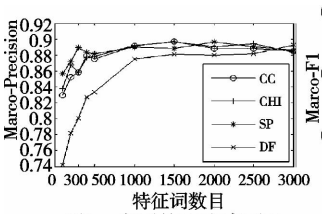

图2 宏平均查准率对比

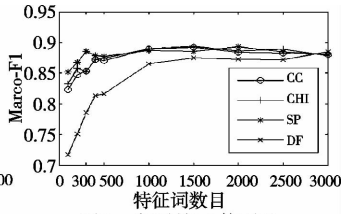

图3 宏平均F1值对比

当提取的文本特征向量维数为 300 维时，SP 的分类效果远好于 CC、CHI 和 DF，此时 Marco-Recall = 0.8867，Marco-Precision = 0.8894，Marco-F1 = 0.8858。通过观察可知，在一定范围内，随着文本特征向量维数的增加，分类效果会逐渐提高，但是当文本特征向量维数超过一定数值时将趋于平稳甚至下降，同时提高了计算代价。虽然 SP 在提取的文本特征向量维数为 2 000 时分类效果最好，此时 Marco-Recall = 0.8933，Marco-Precision = 0.8965，Marco-F1 = 0.8932，但是通过与 SP 提取的文本特征向量维数为 300 维的分类效果作比较，发现其提高的分类性能很有限。由以上分析可知，SP 提取的文本特征向量维数为 300 维，即特征降维达到 99.48% 时，已经使 KNN 具有一个良好的分类效果，可以有效对未知文本进行分类。

通过实验结果可知，SP 通过提取高质量的特征词，构造低维的特征向量，能够有效地降低特征空间维度，并且有效地表示各个类别的文本，反映了类别间的差异度。实验结果表明该方法应用在 KNN 分类算法上分类效果良好。

5 结束语

本文提出一个新的基于类别正相关和类别强相关的新特征提取方法 SP。与传统方法相比，该方法能够充分利用特征项在文本集合中的分布统计信息选择与文本类别正相关并且强相关的特征。该方法在降维能力上有突出表现，实验中在保证良好分类效果的情况下，可对经过预处理后的特征向量进行高达 99.49% 的降维，整体评价指标上表现优于 DF 方法，降维能力、分类效果比传统特征提取方法 CHI、CC 性能上有所提升。

参考文献:

[1] OGURA H,AMANO H,KONDO M. Feature selection with a measure of deviations from Poisson in text categorization[J]. Expert Systems with Applications, 2009,36(3):6826-6832.

[2] 李荣陆. 文本分类及其相关技术研究[D]. 上海:复旦大学, 2005.

[3] WANAS N M,SAID D A,HEGAZY N H,*et al*. A study of local and global thresholding techniques in text categorization[C]//Proc of the 5th Australasian Conference on Data Mining and Analystics. Darlinghurst,Australia:Australian Computer Society,2006:91-101.

[4] LI Shou-shan,XIA Rui,ZONG Cheng-qing,*et al*. A framework of feature selection methods for text categorization[C]//Proc of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg,PA:Association for Computational Linguistics,2009:692-700.

[5] LIU Hua-wen,SUN Ji-gui,LIU Lei,*et al*. Feature selection with dynamic mutual information[J]. Pattern Recognition,2009,42(7):1330-1339.

[6] 肖婷,唐雁. 改进的χ^2统计文本特征选择方法[J]. 计算机工程与应用,2009,45(14):136-137,140.

[7] NG H T,GOH W B,LOW K L. Feature selection, perceptron learning, and a usability case study for text categorization[C]//Proc of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1997:67-73.

[8] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning. San Francisco: Morgan Kanfmann, 1997:412-420.

[9] CUI Zi-feng, XU Bao-wen, ZHANG Wei-feng,*et al*. A new approach to feature selection for text categorization[J]. Wuhan University Journal of Natural Sciences,2006,1(5):1335-1339.

[10] GALAVOTTI L, SEBASTIANI F,SIMI M. Experiments on the use of feature selection and negative evidence in automated text categorization[C]//Proc of the 4th European Conference on Research and Advanced Technology for Digital Libraries. London: Springer-Verlag, 2000:59-68.

[11] ZHENG Zhao-hui, WU Xiao-yun, SRIHARI R. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2002,6(1):80-89.

[12] HOW B C,NARAYANAN K. An empirical study of feature selection for text categorization based on term weightage[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC:IEEE Computer Society, 2004: 599-602.

[13] SOUCY P,MIMEAU G W. Feature selection strategies for text categorization[C]//Proc of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artifical Intelligence. Berlin:Springer-Verlag,2003:505-509.