第 35 卷 第 2 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.2 Computer Engineering January 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)02—0194—03 文献标识码：A 中图分类号：TP391

文本分类中一种混合型特征降维方法

刘海峰[1,2]，王元元[1]，姚泽清[2]，张述祖[2]
(1. 解放军理工大学指挥自动化学院，南京 210007；2. 解放军理工大学理学院，南京 210007)

摘 要：提出一种基于特征选择和特征抽取的混合型文本特征降维方法，分析基于选择和抽取的特征降维方法各自的特点，借助特征项的类别分布差异信息对特征集进行初步选择。使用一种新的基于 PCA 的特征抽取方法对剩余特征集进行二次抽取，在最大限度减少信息损失的前提下实现了文本特征的有效降维。对文本的分类实验结果表明，该特征降维方法具有良好的分类效果。

关键词：文本分类；特征选择；特征抽取；主成分分析

Mixed Method of Reducing Feature in Text Classification

LIU Hai-feng[1,2], WANG Yuan-yuan[1], YAO Ze-qing[2], ZHANG Shu-zu[2]
(1. Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007;
2. Institute of Sciences, PLA University of Science and Technology, Nanjing 210007)

【Abstract】 A mixed method of reducing the text features based on feature selection and feature extraction is brought forward. The characteristics about feature selection and feature extraction are analyzed. Some features are chosen by using the sort distribution information. And a new way based on Principle Component Analysis(PCA) is used to extract the surplus features and realize the compression of features twice. In the precondition of the information loss least, the text feature decrease smart is completed. Test results show that this method has a better precision in the text categorization.

【Key words】 text classification; feature selection; feature extraction; Principle Component Analysis(PCA)

1 概述

作为机器学习和模式识别技术彼此渗透、相互交融的研究领域，文本自动分类技术具有迫切的现实需求和巨大的潜在应用前景。文本自动分类是指在事先确定的分类体系下将未知类属的文本划入相应类别的过程。目前随着中文信息处理技术特别是中文自动分词技术、中文文本分类语料库建设等基础研究设施的日臻成熟完善，国内中文文本分类研究取得很大进展，文本自动分类成为中文信息处理研究的热点之一。向量空间模型是目前文本表示的主流方式，在该模式下文本向量的高维性以及样本协方差阵所表现出的数据稀疏性是影响文本分类效率的主要瓶颈。寻找合理的特征降维方法成为进一步提高文本自动分类效率的关键。

2 文本特征降维的主要模式及其特点

总的说来，文本特征降维方法主要分为特征选择和特征抽取方法以及两者的合理结合。所谓特征选择一般是指依据某个准则从众多原始特征中选择部分最能反映模式类别统计特性的相关特征，也就是说要找到能表现文本内容的最优特征子集，本质上是对特征集合的约简。但是由于寻找最优解的代价太高，在实际问题中常常将寻找最优特征子集的目标降低为逐个寻找满足某个最优化准则的特征集合，然后将前 k 个最优特征构成所需的特征子集。一般说来，这种方法虽然不能得到最优特征子集，但由于其计算量小，模型构造相对简单而且选择出的特征子集基本能够满足分类需求，因此成为一种常用的方法。

特征选择主要是基于某个决策函数对特征项在文本之间的分布信息进行统计。常见的方法有文档频率(DF)、信息增益(IG)、互信息(MI)、χ^2 (Ch-S)统计、期望交叉熵(ECE)、文本证据权(WET)等模型。这些模型由于构造相对简单、易于理解而得到广泛应用。大体说来，这些方法的指导思想是通过对特征频数或特征与类别之间的相关性统计信息寻找分值较高的特征项，选择的标准是特征项对类别的表现能力。但是特征项的赋权大小与其对文本分类的作用未必有相应的同步关系。首先文本中使用的词语存在着潜在的语义关系，一词多义和多词同义现象非常普遍，而这一情况与基于选择的特征降维模型前提条件“特征项之间相互独立”相矛盾；其次即使假定各个原始特征之间相互独立，利用一定的最优准则得到的各个最优特征构成的特征子集未必就是最优特征子集[1]。因此，从方法上决定了特征选择模型在文本分类应用上难以达到十分理想的效果。

另一方面，特征抽取却是基于特征项之间的语义关系对文本特征集的一种压缩。在文本分类中这种模式着眼于处理特征选择方法的“特征项之间相互独立”在现实文本中不成立的问题。即着眼于处理特征项之间的一词多义、多词同义、近义等现象所引起的特征降维困难。特征抽取是通过映射 φ 把原始特征空间 R^n 的数据投影到低维空间 R^k ：

$$\varphi: x \to y, R^n \to R^k, \text{其中} k < n$$

进而以像集作为特征集进行文本分类的过程，其本质是完成测量空间到特征空间的变换。常用的特征抽取方法有主成分分析(PCA)、因子分析(FA)，潜在语义分析(LSA)等。特征抽

基金项目：国家自然科学基金资助项目(70571087)
作者简介：刘海峰(1962－)，男，副教授、博士研究生，主研方向：数据挖掘，文本分类；王元元，教授、博士生导师；姚泽清，教授；张述祖，副教授
收稿日期：2008-06-05 **E-mail：**liuhaifeng19620717@sina.com

取得到的特征含有了特征项之间的语义相关信息，是对特征集在语义层面下的维数压缩。

3 特征降维模型

特征选择和特征抽取分别从统计和变换的角度对特征维数进行压缩。前者注重于特征与文本之间的分布信息而很少利用特征之间的语义信息，后者侧重于特征之间、特征与文本之间的语义信息但存在高维矩阵的分解困难。2 种方法各有优点，也分别存在不足的方面，将两者有机结合构造多步骤或组合型的特征降维方法是一种很有前途的发展方向[2]。本文着眼于探讨这 2 种方法的结合：先以特征项在不同文本类里分布的差异为依据对特征集进行初步筛选；再借助信息论思想使用一种新的基于主成分分析的特征抽取方法对剩余的特征集进一步抽取；在最大限度减少信息损失的前提下前后 2 次对特征集进行有效压缩。随后的文本分类实验结果表明，这种特征降维方法在文本分类上具有较高的精度。

3.1 特征选择降维方法

文本的特征通常是指构成文本的词或词组。即使几百个文本的特征个数也常常有几万个，如此高维的文本向量使得文本的表示、文本之间相似度计算、文本类内散布矩阵、类间散布矩阵的处理等问题变得难以进行。但仔细分析可发现，从在文本集的分布上看，特征项大体分为 3 类[3]：第 1 类特征表现为大量地在某一类文本中出现而在其他类文本里却很少出现，这类特征对判定一个文本属于一个确定的类别具有重要作用；第 2 类特征常常在几个文本类别中出现而在其他类文本里很少出现，这类特征对判定一个文本属于几个确定的类别具有重要作用，而此时这类文本往往属于多个类别，比如一些交叉型学科的文本；第 3 类特征却在几乎所有文本类别中都出现，而这类特征对判定文本类别归属上所起的作用很小，但是这些特征在原始特征集 R^n 里却占有很大的比例。它们的存在使得特征空间维数增大，计算过程变得复杂，并且由于其参与文本相似度的计算而导致分类精度大大降低。因此对这部分特征项进行清除成为决定文本分类效率的关键。笔者从特征项在文本各个类别中的分布统计入手对特征集进行选择，具体方法如下：

设 C_i 表示训练文本集里的文本类别， n_i 为 C_i 内的文本数， t_i 表示 C_i 内含有特征项 T 的文本数， $i=1,2,\cdots,r$ ； r_t 表示含有特征项 T 的文本类别数， $1\le r_t\le r$ ，则 $f_i=\dfrac{t_i}{n_i}$ 表示 T 在 C_i 内相对于文本的分布频率。记

$$\lambda_i=\frac{t_i}{n_i}+0.1 \tag{1}$$

$$\theta_T=\mathrm{lb}[\frac{\max\{\lambda_1,\lambda_2,\cdots,\lambda_r\}}{\min\{\lambda_1,\lambda_2,\cdots,\lambda_r\}}\times\frac{r}{r_t}] \tag{2}$$

则由于 $\dfrac{\max\{\lambda_1,\lambda_2,\cdots,\lambda_r\}}{\min\{\lambda_1,\lambda_2,\cdots,\lambda_r\}}\times\dfrac{r}{r_t}\ge 1$，易知 $\theta_T\in[0,\mathrm{lb}11+\mathrm{lb}r]$。这里 θ_T 体现了特征项 T 在不同文本类别之间分布状况的差异：θ_T 越大，说明特征项 T 的出现越集中在少数几个文本类内，对文本的类别判定作用越大。特别的，当 T 仅在一个文本类别中出现且该类别的所有训练文本均含有 T 时，有

$$\max\{\lambda_1,\lambda_2,\cdots,\lambda_r\}=1.1,\ \min\{\lambda_1,\lambda_2,\cdots,\lambda_r\}=0.1,\ r_t=1$$

此时 $\theta_T=\mathrm{lb}11+\mathrm{lb}r$ 。

反之 θ_T 越小，说明特征项 T 在不同文本类别上的分布越广泛，对文本的类别判定作用越小。而在极端情况下：当特征项在所有文本类别里均出现时，假如此时不同训练文本类所含文本数相同： $n_i=n_j, i,j=1,2,\cdots,r$ ，则

$$\max\{\lambda_1,\lambda_2,\cdots,\lambda_r\}=\min\{\lambda_1,\lambda_2,\cdots,\lambda_r\},\ \ r_t=r$$

从而 $\theta_T=0$。这时 T 对于文本类别判定没有作用。而在式(1)中加入修正常数 0.1 是为了避免 $\min\{\lambda_1,\lambda_2,\cdots,\lambda_r\}=0$ 时式(2)没有意义情况出现。

对于文本特征项集合 $S=\{T_1,T_2,\cdots,T_n\}$ ，使用式(2)计算各特征项相应的统计值 θ_T ，并按照从大到小顺序排列：

$$\theta_{Ti_1}\ge\theta_{Ti_2}\ge\cdots\ge\theta_{Ti_n} \tag{3}$$

其中， $i_1,i_2,\cdots,i_n$ 为 $1,2,\cdots,n$ 的一个排列，取式(3)中前 K 个值相应的 K 个特征项($k<<n$) $T_{i_1},T_{i_2},\cdots T_{i_k}$ 组成原始特征集合的一个有效子集对文本进行标注，完成对原始特征集基于类别分布差异的特征选择，将原始特征集 R^n 维数从 n 维降到 k 维。

3.2 二次特征降维

特征选择主要是基于特征项的频数以及特征项在文本之间分布的统计信息进行特征筛选，没有或者很少利用特征之间的语义信息。其实自然语言中语义的准确表达不仅取决于词汇本身的恰当使用，也取决于上下文对词义的界定，如果忽视上下文语境的限制，仅以孤立的关键字来表示文本的内容，势必影响文本分类的准确率。特征抽取是基于特征项之间的语义相关度对文本特征集的一种压缩模式。其中主成分分析法在模式识别、文本分类方面取得了令人满意的效果，本文在上节特征选择的基础上使用一种新的基于信息论的 PCA 改进方法对选择出的特征集进行进一步的压缩。

3.2.1 PCA 的基本原理与方法

主成分分析(PCA)是一种基于多维统计分析的线性鉴别方法，是寻求有效线性变换的经典方法之一。在文本分类中，PCA 的目的是在最小均方差意义下求解代表原始文本特征集的最佳投影方向，本质上是一种基于目标统计特性的最佳正交变换，使得变换后产生的新特征的各个分量正交或不相关。准则函数定义为

$$\varphi(\xi)=\xi^{\mathrm{T}}\boldsymbol{S}\,\xi \tag{4}$$

其中，

$$\boldsymbol{S}=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})(x_{ij}-\overline{x})^{\mathrm{T}} \tag{5}$$

为样本的总体散布矩阵，这里 $\overline{x}_i=\dfrac{1}{n_i}\sum\limits_{j=1}^{n_i}x_{ij}$ 为第 i 类样本均值，$\overline{x}=\dfrac{1}{N}\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{n_i}x_{ij}$ 为总体样本均值，易知 $\boldsymbol{S}$ 为非负定矩阵，特征值为非负实数。

借助Fisher鉴别法的思想[4]，寻找一组正交单位向量 $\xi_1,\xi_2,\cdots,\xi_s$ ；使得式(4)在这组正交向量下达到最大值，本质上是使得变换后所得特征的类间散布量与类内散布量之和达到最大。可以证明[5]$\boldsymbol{S}$的$u(u\le s)$个最大正特征值 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_u>0$ 对应的标准特征向量 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_u$ 即为最佳投影向量，取 $\boldsymbol{P}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_u)$ ，则投影变换为 $\boldsymbol{Y}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}$ ，样本在 $\boldsymbol{\alpha}_1$ 方向上投影后方差最大，在 $\boldsymbol{\alpha}_2$ 方向上投影后方差次之，依次递减；也就是说在 $\boldsymbol{\alpha}_1$ 方向上投影后特征项集所损失的信息量最少，即 λ_1 所承载的信息量最大。

3.2.2 一种新的基于 PCA 思想的特征抽取方法

根据上面分析， $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_r$ 所承载的信息量依次递减，对于式(5)，取样本 x_{ij} 为第i类簇(相应第i类文本)的 第j个特征，对$\boldsymbol{S}$所有正特征值 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_u>0$ 作如下变 换

[6]：$\mu_i=\frac{\lambda_i}{\sum_{i=1}^{u}\lambda_i}$, $i=1,2,\cdots,u$，则 $0<\mu_i<1$；定义信息函数

$$I(\lambda_i)=\mathrm{lb}(1+\mu_i),\ i=1,2,\cdots,u$$

该函数为增函数。$I(\lambda_i)$ 取值越大，λ_i 所含有的信息量越大，此时若 $I(\lambda_1)\geq I(\lambda_2)\geq\cdots\geq I(\lambda_t)$，且使得

$$\sum_{i=1}^{t}I(\lambda_i)/\sum_{i=1}^{u}I(\lambda_i)>1-\beta\% \tag{6}$$

其中，β 为参数，常取 $\beta\in(0,50)$，则取 $\boldsymbol{P}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_t)$ 为投影矩阵，作投影变换：

$$\boldsymbol{Y}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{X} \tag{7}$$

后，特征项所含的信息损失量 $I_{\mathrm{loss}}<\beta\%$，实验中取 $\beta=15$。

3.3 模型概括

把这种特征选择和特征抽取相结合的特征降维模型概括如下：

首先使用式(3)对原始特征集 R^n 进行选择，得到 k 维特征子集 R^k，然后使用式(6)确定投影矩阵 $\boldsymbol{P}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_t)$ 中所含投影向量个数 t，并对 R^k 中相应的文本向量按照式(7)进行投影压缩至 t 维，从而通过这种特征选择和特征抽取相结合的方式，在最大限度减少信息损失的前提下实现了特征空间的有效降维。

4 实验

4.1 分类器选取

KNN 方法是一种简单有效的非参数方法，在准确率和召回率方面表现出众成为文本分类中常用的分类器。该算法的具体步骤为：

(1)将待分类文本 $\boldsymbol{d}$ 表示成和训练文本构成的维数一致的特征向量。

(2)根据距离函数计算待分类文本向量 $\boldsymbol{d}$ 和训练文本向量的相似度，可以使用两向量之间欧氏距离 $|\boldsymbol{d}-\boldsymbol{d}_i|$ 计算；选择与 $\boldsymbol{d}$ 相似度最大(距离最小)的 k 个文本作为 $\boldsymbol{d}$ 的 k 个最近邻。

(3)根据 $\boldsymbol{d}$ 的 k 个最近邻依次计算文本类别 $c_1,c_2,\cdots,c_r$ 相应的权重，计算公式为

$$p(x,c_j)=\sum_{i=1}^{k}sim(\boldsymbol{d}_i,\boldsymbol{d})\delta(\boldsymbol{d}_i,\boldsymbol{d})$$

其中，$sim(\boldsymbol{d}_i,\boldsymbol{d})$ 表示文本 $\boldsymbol{d}_i$ 与文本 $\boldsymbol{d}$ 之间相似度，使用常用的向量夹角余弦计算；

$$\delta(\boldsymbol{d}_i,\boldsymbol{d})=\begin{cases}1 & \boldsymbol{d}_i\text{是属于}c_j\text{的文本}\\ 0 & \boldsymbol{d}_i\text{不属于}c_j\end{cases}$$

为示性函数。

(4)将待分类文本 $\boldsymbol{d}$ 归入权重最大的类别。

4.2 实验结果及其分析

本文对上述方法的分类效果进行了实验。实验数据为新浪网上下载的 3 580 篇文本，其中分为房地产(580 篇)、金融(560 篇)、体育(630 篇)、计算机(602 篇)、音乐(505 篇)以及旅游(703 篇)。实验时采用 4 分交叉实验法，将 3 580 篇文本平均分为 4 份，3 份为训练集，1 份为测试集；每份轮流作为测试集循环测试 4 次，取平均值为测试结果。经过剔除特高频词、低频词及停用词等文本预处理后的原始特征集特征维数为 2 301，使用式(3)进行特征选择后剩余特征数为 450 个，使用式(6)、式(7)进行投影变换，将特征维数压缩到 120 个，效果评估函数使用常用的查准率、查全率和 F_1 测试值：

查准率=分类的正确文本数/实际分类文本数；

查全率=分类的正确文本数/应有文本数；

$$F_1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}}$$

分类结果统计如表 1 所示。

表 1 实验结果统计 (%)

类 别	查准率	查全率	F_1测试值
房地产	78.4	83.3	80.8
金融	83.2	81.3	82.2
体育	90.3	87.2	88.7
计算机	89.6	92.1	90.8
音乐	85.3	88.7	87.0
旅游	89.2	88.5	88.8

由表 1 可知，平均查准率为 86.0%，查全率为 86.9%，F_1 测试值为 87.7%。其中，房地产与金融两类的 F_1 值略低一点，可能是由于这 2 类文本里部分文本分界不是那么明显所至，效果还是令人满意的。

5 结束语

特征降维问题是文本处理所必须面对的主要问题之一，是制约文本分类效率的瓶颈。本文着眼于研究特征选择和特征抽取方法的结合途径，先用一种基于特征项的类别分布差异信息对特征集进行初步选择，再使用基于主成分分析思想的特征抽取方法对选择出的特征集进一步抽取，在最大限度减少信息损失的前提下先后 2 次对特征集进行压缩。对中文文本分类实验结果表明，本文提出的特征降维方法在文本分类的准确性方面效果较好。如何将 2 种模型进行更合理的结合，构造组合型的特征降维模型是今后要继续深入研究的工作之一。

参考文献

[1] Cover T M. The Best Two Independent Measurements Are Not the Two Best[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1974, 4(1): 116-117.

[2] Makrehchi M, Kamel M S. Text Classification Using Small Number of Features[C]//Proc. of the 4th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition. [S. l.]: IEEE Press, 2005: 580-589.

[3] 陈治平, 林亚平, 彭 雅, 等. 基于最小类别差异的无关信息预处理算法[J]. 电子学报, 2003, 31(11): 1750-1753.

[4] 宋枫溪, 刘树海, 杨静宇, 等. 最大散度差分类器及其在文本分类中的应用[J]. 计算机工程, 2005, 31(5): 8-10.

[5] Jin Zhong, Yang Jingyu, Hu Zhongshan, et al. Face Recognition Based on Uncorrelated Discriminant Transformation[J]. Pattern Recognition, 2001, 34(7): 1405-1416.

[6] 于世飞, 靳奉祥, 王 健, 等. 一种新的基于信息论的 PCA 特征压缩算法[J]. 小型微型计算机系统, 2004, 25(4): 694-697.