文本分类中一种基于正交变换的特征降维方法*)

刘海峰[1,2] 王元元[2] 张学仁[1] 刘守生[1]
(解放军理工大学理学院[1] 解放军理工大学指挥自动化学院[2] 南京 210007)

摘 要 本文讨论了一种基于正交变换的文本特征降维方法。分析了基于特征选择和特征抽取的特征降维方法各自特点,借助矩阵的分解论证了基于Fisher准则函数的特征降维模式的原理与理论基础,讨论了PCA与SVD两种模式的相互关系。实验结果表明这种特征降维模式在文本分类的准确性方面效果较好。

关键词 文本分类,特征抽取,特征降维,正交变换,奇异值分解

A Method of Reducing Features Based on Orthogonal Transforming in Text Classification

LIU Hai-feng[1,2] WANG Yuan-yuan[2] ZHANG Xue-ren[1] LIU Shou-sheng[1]
(Institute of Sciences[1], Institute of Command Automation[2], PLAUST, Nanjing 210007, China)

Abstract Discuss a method of reducing the text features based on orthogonal transforming. Then, we analyze the characteristic about feature selection and feature extraction. In virtue of decomposing the matrix, we demonstrate the principle and the theory foundation about the reduce of dimensions based on Fisher. We discuss the relation of PCA and SVD. The test shows that this method has a better precision in the text categorization.

Keywords Text classification, Feature extraction, Feature selection, Orthogonal transforming, SVD

1 引言

文本自动分类是指在给定的分类体系下,对未知类别的文本根据其特征自动判断其类别归属的过程。文本自动分类是文本挖掘领域的重要组成部分,在自然语言处理及模式识别等领域的应用日益广泛,在提高信息利用的有效性和准确性上具有重要的现实意义。国外的文本分类研究起步较早,以卡内基集团为路透社开发的Construe系统以及麻省理工学院为白宫开发的邮件分类系统等为代表,文本分类研究已经进入实用化阶段。而国内中文文本分类研究随着中文信息处理技术特别是自动分词技术、中文文本语料库建设等日渐成熟完善也取得很大进展[1],基于机器学习的中文文本自动分类研究成为中文信息处理研究的热点之一。

2 文本分类中特征降维的主要方法及其存在问题

文本分类过程主要分为三个步骤:文本的向量表示、文本特征降维以及分类器构造。训练集样本量大、类别分布倾斜以及文本特征的维数很高是文本分类的显著特点。寻找有效的特征降维方法,抑制噪音对分类的影响成为能否快速高效进行文本分类的关键。总的说来,特征降维方法可分为特征选择和特征抽取两种类型。

2.1 特征选择方法及其相关问题

特征选择一般是利用某个评价函数独立对原始特征项逐一评分,然后从中选出分值最高的前k个特征项,其结果是得到原始特征集的一个子集,本质上是对特征集合的约简。特征选择方法因为评分模型构造相对简单、多样(如基于词频、信息增益、互信息、期望交叉熵、文本证据权等方法)易于理解而得到广泛应用。但是该方法存在两个主要问题:一是该模型没有或是很少考虑特征项之间的相关性信息;二是一些特征项获得的权值大小与其对文本分类作用未必有相应的同步关系:也就是说特征选择方法确定的对分类"价值较高"的特征在不同类别文本的类属判定上未必体现出相应的重要性,对于数据偏斜严重的样本集上这个问题尤为突出[2]。Cover指出,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[3]。尽管如此,自动文本分类的大量研究表明,基于单一最优准则的特征选择方法仍然具有很高的效率[4]。

2.2 特征抽取的主要方法及其特点

特征抽取一般是通过映射把测量空间的数据投影到特征空间:

$$\varphi: x \to y, R^n \to R^d$$

其中,$d<n$;x为原始向量,y为x在φ下的像。

获得原始特征集相应的像集,进而对像集进行评估,其本质是完成测量空间到特征空间的变换[5]。常用的特征抽取方法有主成分分析(Principal Component Analysis, PCA)[6-9],又称(Karhunen-Loeve, K-L)变换,因子分析(Factor Analysis, FA),潜在语义分析(Latent Semantic Analysis, LSA)等[10]。特征抽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。

2.3 基于Fisher准则的文本特征降维原理与方法

主成分分析(Principal Component Analysis, PCA)是一种线性鉴别分析方法。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轴,本质上是一种基于目标统计特性的最佳正交变换[9],使得变换后产生的新特征的各个分量正交或不相关。

设$C_i(i=1,2,\cdots,r)$表示训练文本集里的文本类别,C_i

*)国家自然科学基金资助项目(编号:70571087)。**刘海峰** 副教授,博士生,研究方向为人工智能、文本分类;**王元元** 教授,博导,研究方向为人工智能、计算机软件与理论;**张学仁** 副教授,研究方向为人工智能、信息检索;**刘守生** 副教授,博士,研究方向为人工智能、遗传算法。

内的文本数为 n_i，$i=1,2,\cdots,r$；$N=\sum_{k=1}^{r}n_i$ 为训练集里文本数，$x_{ij}=(w_{ij,1},w_{ij,2},\cdots,w_{ij,n})$ 表示类 C_i 中第 j 个文本，$i=1,2,\cdots,r$；$j=1,2,\cdots,n_j$；$w_{ij,p}$ 表示特征项 T_p 在文本 x_{ij} 中的权重，使用常用的 tf-idf 公式[11]进行赋值。

训练文本类间散度反映了类别样本之间差异的平均，类间散度大说明不同类别样本之间平均间距大；而类内散度反映了类内样本之间密集程度，类内散度小意味着同一类别的样本之间密集程度高。类间散度和类内散度从两个侧面反映了样本的可分性。对训练集进行分类应遵循的原则是应使得类内样本的分散程度最小而类间样本的分散程度最大，为此，定义训练集的类内散布矩阵 S_w、类间散布矩阵 S_b 以及总体散布矩阵 S 分别为相应样本的协方差矩阵[6]：

$$S_w=\sum_{i=1}^{r}p(C_i)\frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T \tag{1}$$

$$S_b=\sum_{i=1}^{r}p(C_i)(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T \tag{2}$$

$$S=\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})^T \tag{3}$$

其中 $\bar{x}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$ 为第 i 类样本均值，$\bar{x}=\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}$ 为总体样本均值，$p(C_i)$ 为第 i 类样本出现的概率，这里以该类文本的频率计算：$p(C_i)=\frac{n_i}{N}$，

因此(1)、(2)式改写为：

$$S_w=\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T \tag{4}$$

$$S_b=\frac{1}{N}\sum_{i=1}^{r}n_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T \tag{5}$$

从定义易知上述矩阵均为非负定阵[12]。

Fisher 准则函数定义为

$$\varphi(y)=\frac{y^TS_by}{y^TS_wy} \tag{6}$$

作为样本的投影方向。也就是说选择使得 $\varphi(y)$ 达到极大值的向量 y 作为样本 x_{ij} 的最佳投影方向，使得 x_{ij} 在 y 方向上投影后的类间离散度最大而类内离散度最小。

也可以使用(3)式来构造确定投影方向的准则函数：

$$\varphi(\xi)=\xi^TS\xi \tag{7}$$

使得 $\varphi(\xi)$ 达到极大值，这里取 ξ 为最大化(7)式的单位向量。

通常情况下，由于训练集的文本类别数较大，样本的分布也不均匀，单一的最优投影方向很难满足分类精度要求，借助 Fisher 鉴别法的思想[12,13]，可以寻找一组正交的单位投影向量 $\xi_1,\xi_2,\cdots,\xi_s$；使得(7)式在这组正交向量下达到最大值。

由于 S 为实对称矩阵，故 S 亦为 Hermite 阵，故有[14]：

性质 1 S 的特征值为实数。

性质 2 S 的相异特征值对应的特征向量必定正交。

定义 1 满足内积 $(\sigma(\alpha),\beta)=(\alpha,\sigma(\beta))$ 的欧氏空间的线性变换 σ 称为对称变换。

引理 1 设 σ 为对称变换，V 是 σ 的不变子空间，则 V 的正交补 V^* 也是 σ 的不变子空间。

证明：设 $\alpha\in V^*$，则 $\forall\beta\in V$，有 $\sigma(\beta)\in V$，故 $(\alpha,\sigma(\beta))=0$。由于 σ 为对称变换，故 $(\sigma(\alpha),\beta)=0$，即有 $\sigma(\alpha)\perp V$，因而 $\sigma(\alpha)\in V^*$，故 V^* 也是 σ 的不变子空间，证毕。

性质 3 对于(7)式中 S，存在 n 阶正交矩阵 T，使得 $T^{-1}ST=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$

证明：根据实对称矩阵与对称变换的关系[15]，只须证明 σ 有 n 个特征向量构成 R^n 的一组标准正交基。现对维数 n 作归纳法。

显然 $n=1$ 时结论成立；

假设 $n-1$ 时结论成立；

对于 R^n，σ 必有一单位特征向量 α_1，使得 $\sigma(\alpha_1)=\lambda_1\alpha_1$。以 $L(\alpha_1)$ 表示 α_1 生成的子空间，$L^*(\alpha_1)$ 表示 $L(\alpha_1)$ 的正交补，由引理 1，$L^*(\alpha_1)$ 必为 σ 的 $n-1$ 维不变子空间，且在 $L^*(\alpha_1)$ 上 σ 仍为对称变换，故由归纳假设，在 $L^*(\alpha_1)$ 上 σ 有 $n-1$ 个特征向量 $\alpha_2,\alpha_3,\cdots,\alpha_n$ 构成 $L^*(\alpha_1)$ 的标准正交基，从而 σ 的 n 个单位特征向量 $\alpha_1,\alpha_2,\cdots,\alpha_n$ 构成了 R^n 的一组标准正交基。证毕。

此时，$\alpha_1,\alpha_2,\cdots,\alpha_n$ 为 S 的 n 个单位特征向量。

定义 2 设 A 为 n 阶 Hermite 阵，B 为 n 阶正定阵，若有向量 $X\neq0$ 及 $\lambda\in R$，使得 $AX=\lambda BX$，称 λ 为 A(关于 B)的广义特征值。

考虑到(7)式实质上是单位特征向量下的 Rayleigh 熵，由 Rayleigh 熵的性质知[14]：

性质 4 设 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 为实对称矩阵 A 关于 B 的广义特征值，且 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$，则

$$\lambda_1=\max_{0\neq x\in R^n}\frac{X^TAX}{X^TBX},\ \lambda_n=\min_{0\neq x\in R^n}\frac{X^TAX}{X^TBX} \tag{8}$$

此时，取 B 为单位阵，X 为单位向量，$A=S$，则得

$$\lambda_1=\max_{x\in R^n}X^TSX,\lambda_n=\min_{x\in R^n}X^TSX,|X|=1 \tag{9}$$

矩阵 S 的前 n 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$ 相应的单位特征向量依次为投影变换的最佳投影方向，因此 S 的 n 个单位特征向量对准则函数的“贡献程度”单调下降，从而可以取其前 s 个最大特征值 $\lambda_1,\lambda_2,\cdots,\lambda_s$ 所对应的 s 个标准正交特征向量 $\alpha_1,\alpha_2,\cdots,\alpha_s$ 组成(7)式的最优投影矩阵：

$$Q=(\alpha_1,\alpha_2,\cdots,\alpha_s)\in R^{n\times s} \tag{10}$$

此时，对于训练文本 $x_{ij}=(w_{ij,1},w_{ij,2},\cdots,w_{ij,n})$，以(10)式作为投影变换，得 x_{ij} 的投影向量：

$$y_{ij}=Q^Tx_{ij} \tag{11}$$

其中 y_{ij} 为 s 维向量。由于 $s<n$，因而达到文本特征降维目的。投影特征 $\alpha_1,\alpha_2,\cdots,\alpha_s$ 构成 x_{ij} 的主成分。

2.4 PCA 与 SVD 的区别与联系

在文本分类中，奇异值分解[16](SVD)也是一种基于特征相关性的特征抽取方法。通过对文本-特征项矩阵 A_{mn} 进行投影分析，得到文本向量在投影基上的投影坐标。与 PCA 相比，前者是对任意的矩阵 A_{mn} 构成的对称阵 A^TA 以及 AA^T 求解特征值与特征向量，从而得到文本向量在特征向量上的投影；后者是直接对实对称协方差阵 S 进行相似变换得到文本向量在 S 的特征向量上的投影。这两种方法本质上讲都是通过对矩阵的正交分解，借助变换矩阵进行原始空间到低维空间的线性变换。当对正定矩阵进行奇异值分解时，已经证明 SVD 与 PCA 之间具有等价性[7,9]。

3 试验结果及其分析

3.1 分类器选取：一种改进的 KNN 方法

KNN 方法是一种简单有效的非参数方法，在查准率和查全率方面表现出众成为文本分类中常用的分类器[17]。

该算法的具体步骤为：

(1)将待分类文本 d 表示成和训练文本构成的维数一致

(下转第 162 页)

ary algorithms on binary constraint satisfaction problems. IEEE Transactions on Evolutionary Computation, 2003, 7(5): 424-444
[2] Manuel B, Daniel K. Locally consistent constraint satisfaction problems with binary constraints. Lecture Notes in Computer Science, 2005, 3787: 295-306
[3] 杨轻云,孙吉贵,张居阳. 最大度二元约束满足问题粒子群算法. 计算机研究与发展,2006,43(3): 436-441
[4] Michel A, Khaled H-H, Guillaume M, et al. Mass customiza t-ion and configuration: Requirement analysis and constraint based modeling propositions. Integrated Computer-Aided Engineering, 2003, 10 (2): 177-189
[5] Fahiem B, Chen Xingquang, Peter V B, et al. W Binary vs. n-on-binary constraints. Artificial Intelligence, 2002, 140(1/2): 1-37
[6] Nikolaos S, Kostas S. Binary encodings of non-binary constraint satisfaction problems: Algorithms and experimental results. Journal of Artificial Intelligence Research, 2005, 24(7): 641-684
[7] Bessiere C, Reqin J C. MAC and combined heuristics: Two reasons to forsake FC (and CBJ?) on hard problems//Proceedings of Principles and Practice of Constraint Programming - CP96. 1996: 61-75
[8] Bessiere C, Meseguer P, Freuder E C, et al. On forward checking for non-binary constraint satisfaction. Aritficial Intelligence, 2002, 141(1/2): 205～224
[9] Lal A. Neighborhood interchangeability for non-binary CSPs & application to databases. Ph. D. Dissertation. Lincoln: University of Nebraska, 2005
[10] Liu Zhe. Algorithms for Constraint Satisfaction Problems(CSPs). Master's Thesis. Canada: University of Waterloo, 1998
[11] 施伯乐,丁宝康,汪卫. 数据库系统教程(第2版). 北京: 高等教育出版社, 2005
[12] Lal A, Choueiry B Y, Freuder E C. Neighborhood interchangeability and dynamic bundling for non-binary finite CSPs//Proceedings of the 20th National Conference on Artificial Intelligence and the 17th Innovative Applications of Artificial Intelligence Conference, 2005: 397-404

(上接第 126 页)

的特征向量;

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

(3)根据 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:

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

其中,$\mathrm{sim}(d_i,d)$ 表示文本 d_i 与文本 d 之间相似度,

$$\delta(d_i,d)=\begin{cases}1, d_i \text{ 是属于 } c_j \text{ 的文本}\\ 0, d_i \text{ 不属于 } c_j\end{cases}$$ 为示性函数。

(4)将待分类文本 d 归入权重最大的类别。

考虑到 KNN 分类器对训练集数据倾斜问题的敏感性,我们使用如下距离公式[18]对 KNN 算法进行改进,即以

$$d(d_i,d)=\frac{|d_i-d|}{\sqrt{M(d_i)}\sqrt{M(d)}} \qquad (12)$$

替换算法中步骤(2)的欧式距离。这里 $M(d_i)$、$M(d)$ 表示 d_i、d 与其它文本之间欧式距离的平均值。

(12)式使得样本分布密集区域的样本之间距离被放大,而稀疏区域的样本之间距离被缩小,从而有效降低了由于样本点分布不均而对分类器的影响。

3.2 试验结果及其分析

本文对上述方法的分类效率进行了试验。试验数据为新浪网上下载的 3460 篇文本,其中分为文学(560 篇)、金融(579 篇)、体育(651 篇)、计算机(572 篇)、新闻(495 篇)以及军事(603 篇)。试验时采用 3 分交叉试验法,将 3460 篇文本平均分为 3 份,2 份为训练集,1 份为测试集;每份轮流作为测试集循环测试 3 次,取平均值为测试结果。经过文本预处理后的原始特征集特征维数为 1321,使用(7)式进行特征抽取,使用(11)式进行投影变换,取投影空间维数 $s=85$,效果评估函数使用常用的查准率、查全率和 F_1 测试值:

表 1 试验结果统计

类 别	文学	金融	体育	计算机	新闻	军事
查准率	86.1%	89.2%	91.3%	84.1%	85.7%	89.2%
查全率	87.3%	85.4%	82.5%	88.6%	90.2%	84.7%
F_1 测试值	86.7%	87.3%	86.7%	85.8%	87.9%	86.9%

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 1 表所示。

平均查准率为 87.6%,查全率为 86.5%;F_1 测试值为 87.1%,效果还是令人满意的。

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。本文讨论了 PCA 在文本分类中的应用。论证了投影特征的选取方法及其相关的理论依据。讨论了 PCA 与 SVD 这两种基于特征抽取的降维方法异同点以及在分解矩阵为正定阵的条件下两者之间的等价性。特征选择和特征抽取从不同的立足点出发对文本特征进行压缩,两种模式各有其优点和不足。如何将两种模型进行合理的结合,构造组合型的特征降维模型应该是提高特征降维效率的思路之一,这也是我们今后要进行的研究工作。

参 考 文 献

[1] 廖莎莎,等. 中文文本分类中基于概念屏蔽层的特征提取方法. 中文信息学报[J], 20(3): 22-28
[2] 苏金树,等. 基于机器学习的文本分类技术研究进展. 软件学报[J],2006(9): 1848-1859
[3] Cover T M. The best two independent measurements are not the two best. IEEE Transactions on Systems, Man, and Cybernetics, 1974, 4: 116-117
[4] Dumais S, Platt J, Heckerman D, et al. Inductive learning algorithms and representations for text categorization//Proceedings of the CIKM-98, 7th ACM International Conference on Information and Knowledge Management. Bethesda, MD, 1998: 148-155
[5] 宋枫溪,等. 统计模式识别中的维数削减与低损降维. 计算机学报[J],2005(11): 1915-1922
[6] 陈伏兵,等. 小样本情况下 Fisher 线性鉴别分析的理论及其验证. 中国图象图形学报[J], 10(8): 984-991
[7] 吴春国,等. 关于 SVD 与 PCA 等价性研究. 计算机学报[J], 2004(2): 286-288
[8] 张生亮,等. 基于类间散布矩阵的二维主分量分析. 计算机工程[J],2006(6): 44-46
[9] 张志佳,等. 基于线性投影的代数空间降维分析. 计算机工程[J],2005(11): 25-27
[10] 刘海峰,等. 基于潜在语义空间的文本检索问题研究. 情报科学[J], 2007(5): 748-753
[11] 刘海峰,等. 基于 Web 的文本检索位置加权模型研究. 情报科学[J],2007(3): 451-455
[12] 王文胜,等. 一种基于奇异值分解的特征抽取方法. 电子与信息学报[J],2005(2): 294-297
[13] Duda R O, Hart P E, Store D G. Pattern Classification, 2[nd] Edition. New York: John Wiley & Sons, 2001: 117-123
[14] 张明淳. 工程矩阵理论[M]. 南京: 东南大学出版社, 1995: 116-133
[15] 王萼芳,等. 高等代数[M]. 北京: 高等教育出版社, 2003: 377-381
[16] 周水庚,等. 隐含语义索引及其在中文文本信息处理中的应用[J]. 小型微型计算机系统, 2001(22): 240-243
[17] Yang Yiming, Liu Xin. A re-examination of text categorization methods//Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 42-49
[18] 王和勇,等. 基于聚类和改进距离的方法在数据降维中的应用. 计算机研究与发展[J], 2006, 43(8): 1485-1489