文章编号: 1001- 2486(2008) 05- 0135- 04

基于 Fisher 线性判别模型的文本特征选择算法*

刘 健, 钱 猛, 张维明
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 在采用向量空间模型表示方法的文本分类系统中, 维数约简是必要的步骤, 特征选择方法由于计算复杂度较低而被广泛采用。本文基于 Fisher 线性判别模型提出了一种新的文本特征选择算法, 将其求解过程转换为一个特征项优化组合的问题, 避免了复杂的矩阵变换运算。实验表明, 该方法与信息增益、卡方统计方法比较, 具有较明显的优势。

关键词: Fisher 线性判别模型; 文本分类; 特征选择
中图分类号: TP391 文献标识码: B

A Fisher Linear Discriminant Model-Based Text Feature Selection Algorithm

LIU Jian, QIAN Meng, ZHANG Wei-ming
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Dimension reducing is very important in VSM based text classification system. Feature selection is more suitable for text data because of its efficiency. A new feature selection algorithm is proposed in this paper on the basis of Fisher linear discriminant model, which converts the solution process to feature optimization problem and avoids the complex matrix operations. The experiment shows that the new algorithm has good performance and is better than IG and CHI method.

Key words: fisher linear discriminant model; text classification; feature selection

文本分类是基于内容的信息检索和数据挖掘的重要基础, 在自然语言处理与理解、信息组织与管理、内容过滤等领域都有着广泛的应用。向量空间模型(Vector Space Model) 是目前文本表示的主要方法[1], 这种方法的特点是文本向量的维数很高, 通常一个文本向量可以达到数万维的量级。很多文献[1-3] 指出降维不仅能大量降低处理开销, 而且可以提高分类的效果, 因此对高维的文本向量进行降维是必要的工作, 特征降维的方法主要有特征选择和特征抽取两种。两者的性能没有显著的差异, 特征选择方法由于计算复杂度较低而被广泛采用。

基于 Fisher 线性判别模型的特征抽取方法在许多应用中取得了很好的证明, 但是该方法的主要不足在于其计算规模受到限制, 复杂度与初始特征集合的数目相关, 因此当特征数目很多时, 需要的计算量与存储规模都急剧提升[4]。本文利用 Fisher 线性判别模型, 从特征选择的角度出发, 通过推导, 提出了一种新的特征选择算法, 并与目前常用的特征选择算法进行了实验比较, 实验结果表明, 该方法具有较明显的优势。

1 Fisher 线性判别模型

Fisher 线性判别模型的基本思想是通过对样本的变换, 将样本投影到一条直线上, 使样本的投影能分得最好, 也就是说变换后的样本类别间离散度达到最高, 类内的样本离散度达到最低, 从而提高各个类别之间的区分能力。

设训练集中类别分别为 $c_1, c_2, \cdots, c_L$, 类别总数为 L。类内散布矩阵 S_a 与类间散布矩阵 S_b 可以分

* 收稿日期: 2008- 01- 12
基金项目: 国家自然科学基金资助项目(70371008)
作者简介: 刘健(1975—), 男, 工程师, 博士生。

别定义为

$$S_a = \sum_{j=1}^{L} E[(X - m_j)(X - m_j)^{\mathrm{T}} \mid c_j] \tag{1}$$

$$S_b = \sum_{j=1}^{L} P(c_j)(m_j - m_0)(m_j - m_0)^{\mathrm{T}} \tag{2}$$

其中 m_j 为类别 c_j 的均值, $P(c_j)$ 为其先验概率(通常取值为 $P(c_j) = n_j/N$, n_j 为训练集中某一类别样本数, N 为训练集中全部样本数), m_0 为训练集总体均值。

定义目标函数

$$J(w) = \frac{w^{\mathrm{T}} S_b w}{w^{\mathrm{T}} S_a w} \tag{3}$$

寻找变换矩阵 $w = (w_1, w_2, \cdots, w_m)^{\mathrm{T}}$, 使得函数 $J(w)$ 取得极大值。

2 FS特征选择算法

求解变换矩阵 w 的方法一般是利用特征矩阵进行分解变换, 但文本向量维数动辄达到数千、上万甚至十万的量级, 对特征矩阵进行变换, 无疑计算开销非常大。从特征选择的角度考虑, 维数约减的目的是找到一个特征子集 T^k 使得 $J(w)$ 取得极大值, 因此完全可以将求解过程转换为一个特征项优化组合的问题, 取得其中一个近似的最优解, 从而避免矩阵变换带来的计算量。本文的思路是对变换矩阵 w 增加限定条件, 即假设 w 为元素均为1或0的二元矩阵, 由此寻找一组特征使得目标函数取得极大值。

首先引入惩罚函数

$$\theta(w) = \begin{cases} 1, & w_{ij}^2 > 0 \\ 0, & w^{ij} = 0 \end{cases} \tag{4}$$

利用 $\theta(w)$ 对 w 进行变换, 得到 $w = \theta(w)$, 由于 w 中元素取值都为1或者0, 代入目标函数, 得到

$$J(w) = \frac{w^{\mathrm{T}} S_b w}{w^{\mathrm{T}} S_a w} = \frac{\sum_{i=1}^{k} S_b(t_i)}{\sum_{i=1}^{k} S_a(t_i)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{c} \frac{n_j}{n}(m_j^i - m^i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{c}\sum_{l=1}^{n_j}(x_i^l - m_j^i)^2} \tag{5}$$

至此, 问题已经转换为在特征空间 T^n 中搜索出一个子空间 T^k ($T^k \subseteq T^n$), 使得函数(5)在 T^k 上的取值大于在其他子空间 $T^{k'}$ 上的取值。这等价于将每个特征项 t_i 按照函数 $f(t_i) = \frac{S_b(t_i)}{S_a(t_i)}$ 值由大到小进行排序, 由此序列构成的新的有序特征空间是使得目标函数递增速度最快(或者递减速度最慢)的序列, 也就是说按照此序列, 从特征空间中选择前 k 个特征项, 必然可以使目标函数的取值不小于在其他 k 个特征项上的取值。

根据以上论述, 可以确定特征评估函数为

$$f(t_i) = \frac{S_b(t_i)}{S_a(t_i)} = \frac{\sum_{j=1}^{c} \frac{n_j}{n}(m_j^i - m_0^i)^2}{\sum_{j=1}^{c}\sum_{l=1}^{n_j}(x_i^l - m_j^i)^2} \tag{6}$$

按照每个特征的 $f(t_i)$ 值由大到小排列, 前 k 个特征项就是所需要的特征。FS特征选择算法可以归纳为:

(1) 计算训练样本集总体均值: $m_j = \frac{1}{n_j}\sum_{x_i \in c_j} X_i$

(2) 计算训练样本集总体均值: $m_0 = \frac{1}{N}\sum_{i=1}^{n} X_i = \frac{1}{N}\sum_{j=1}^{c} n_j M_j$

(3) 计算每个特征项对应的评估值: $f(t_i) = \frac{S_b(t_i)}{S_a(t_i)}$

(4)选取 k 个具有最大$f(t_i)$评估值的特征项作为最终特征集合。

3　实验与讨论

实验选用的语料是中文文本分类语料库—TanCorpV1.0[5],该语料库分为两个层次,共包含文本14150篇,第一层为12个类别,第二层为60个类别。依照第一层类别标准,平均抽取10个类别共5000篇文档,其中3000篇作为训练集,其他文档平均分为两个测试集,测试集1包含500篇文档,测试集2包含1500篇文档。具体情况如表1所示。

表1　试验数据基本情况

Tab.1　Experiment Data

编号	类型	训练集		测试集1		测试集2	
		样本数	总词数	样本数	总词数	样本数	总词数
1.	人才	300	12764	50	4352	150	8389
2.	体育	300	10727	50	1878	150	5158
3.	卫生	300	10588	50	2513	150	6463
4.	娱乐	300	15175	50	2808	150	8926
5.	房产	300	10883	50	1720	150	4293
6.	教育	300	11828	50	2325	150	6619
7.	汽车	300	10503	50	2019	150	4482
8.	电脑	300	10375	50	2394	150	5263
9.	艺术	300	19523	50	4172	150	11159
10.	财经	300	13942	50	1995	150	6491
合计		3000	43551	500	12241	1500	26971

文本分类中通常用准确率(precision)和召回率(recall)这两个评价指标进行评估,本文采用综合分类率 F_1 进行评价,该指标综合考虑准确率与召回率,具体计算公式如下:

$$F_1=\frac{2\times precision\times recall}{precision+recall}$$

本文使用 *tf-idf* 作为权重表示方法,为了使实验结果有可比性,将卡方统计,信息增益和本文提出的FS算法在matlab平台上编程实现,采用LIBSVM[6]作为分类器对这几种方法选择出的特征进行了分类实验。同时,为了检测这些方法在特征维数变换情况下的性能,选取了不同特征维数,将综合分类率 F_1 的变化情况进行了比较,具体实验结果如图1所示。

从实验结果可以看出,本文提出的FS算法与另两种算法比较具有明显的优势,除了在测试集1上特征维数小于300时 F_1 值略低于卡方统计外,其他均高于另外两种算法。在降维性能方面,FS算法在维数较低的时候已经取得了较好的分类能力,尤其是在特征维数为1500时达到了最优性能,在测试集1上达到89%,在测试集2上达到83.5%。从实验曲线的变化趋势看,FS算法也趋于稳定。由此说明,FS特征选择算法性能稳定、鉴别能力较强。

4　结束语

本文基于Fisher线性判别模型,针对该模型计算复杂的问题,从特征选择的角度对其进行了优化,提出一种新的特征选择算法用于文本分类,该算法在维数约简与分类性能上均取得了满意的效果,但是在特征维数较少的情况下,不如卡方统计算法,因此,在未来的工作中,将进一步研究卡方统计算法的原理,优化FS算法。

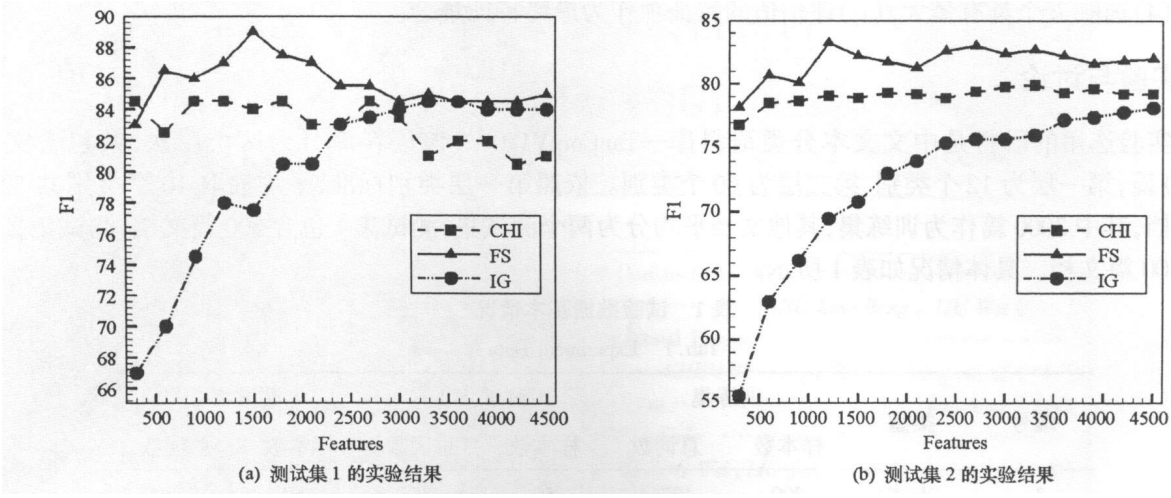

图 1 实验结果
Fig. 1 Experiment result

参 考 文 献:

[1] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报 2006, 17(9): 1848- 1859.
[2] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1- 47.
[3] Chen W, Chang X, Wang H, et al. Automatic Word Clustering for Text Categorization Using Global Information[C]// Proc. of the Information Retrieval Technology, Asia Information Retrieval Symp. (AIRS 2004). Beijing: Springer-verlag, 2004. 1- 11.
[4] 封举富, 时建新. 基因选择的快速 Fisher 优化模型[J]. 北京大学学报(自然科学版), 2005, 141(1): 122- 128.
[5] 谭松波, 王月粉. 中文文本分类语料库- TanCorpV1.0[DB].
[6] LIBSVM[CP]. http://www.csie.ntu.edu.tw/~ cjlin/libsvm/

(上接第 134 页)

参 考 文 献:

[1] 王子才, 等. 仿真系统的校核、验证与验收(VV&A): 现状与未来[J]. 系统仿真学报, 1999, 11(5): 321- 325.
[2] Base Object Model (BOM) Template Specification[Z]. SISO- STD- 003.1- DRAFT- V0.11, 2005.
[3] SISO- STD- 003.1. Guide for Base Object Model (BOM) Use and Implementation [S]. SISO, 2006.
[4] 龚建兴, 等. 构建可扩展的 HLA 联邦成员架构[J]. 系统仿真学报, 2006, 18(11): 3126- 3130.
[5] Defense Modeling and Simulation Office(DMSO). High Level Architecture Federation Development and Execution Process (FEDEP) Model Version 1.4[S]. DMSO, 1999.
[6] IEEE. IEEE Draft Recommended Practice for High Level Architecture (HLA)- Federation Development and Execution Process (FEDEP). IEEE P1516.3™[S]. IEEE, 2003.
[7] 曹星平. HLA 仿真系统的校核验证与确认研究[D]. 长沙: 国防科技大学, 2004.
[8] Weisel E W, Petty M D, Mielke R R. Validity of Models and Classes of Models in Semantic Composability[C]// Proceedings of the Fall 2003 Simulation Interoperability Workshop, 03F- SIW- 073.
[9] 孙世霞. 复杂大系统建模与仿真的可信性评估研究[D]. 长沙: 国防科技大学, 2005.