统计模式识别中的维数削减与低损降维

宋枫溪[1),2)] 高秀梅[3)] 刘树海[2)] 杨静宇[4)]

[1)](哈尔滨工业大学深圳研究生院 深圳 518000)

[2)](炮兵学院二系 合肥 230031)

[3)](淮阴师范学院计算机系 淮阴 223001)

[4)](南京理工大学计算机系 南京 210094)

摘 要 较为全面地回顾了统计模式识别中常用的一些特征选择、特征提取等主流特征降维方法,介绍了它们各自的特点及其适用范围,在此基础上,提出了一种新的基于最优分类器——贝叶斯分类器的可用于自动文本分类及其它大样本模式分类的特征选择方法——低损降维.在标准数据集 Reuters-21578 上进行的仿真实验结果表明,与互信息、χ^2 统计量以及文档频率这三种主流文本特征选择方法相比,低损降维的降维效果与互信息、χ^2 统计量相当,而优于文档频率.

关键词 维数削减;特征选择;特征抽取;低损降维;文本分类

中图法分类号 TP18

Dimensionality Reduction in Statistical Pattern Recognition and Low Loss Dimensionality Reduction

SONG Feng-Xi[1),2)] GAO Xiu-Mei[3)] LIU Shu-Hai[2)] YANG Jing-Yu[4)]

[1)](*Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen* 518000)

[2)](*New Star Research Institute of Applied Technology in Hefei City, Hefei* 230031)

[3)](*Department of Computer, Huaiyin Teachers College, Huaiyin* 223001)

[4)](*Department of Computer, Nanjing University of Science and Technology, Nanjing* 210094)

Abstract First, authors review the prevailing feature selection methods such as Exhaustive Search, Genetic Algorithm, Sequential Forward Floating Selection, and Best Individual Features, and feature extraction approaches such as Principal Component Analysis, Fisher Discriminant Analysis, and Projection Pursuit for feature space dimensionality reduction in statistical pattern recognition. Second, authors discuss the characteristics and the applicable domains of all these techniques. Third, authors propose a novel feature selection method based on so-called optimal classifier, Bayesian classifier. The new feature selection method, i.e. the low loss dimensionality reduction (LLDR), is applied in automatic text categorization and compared with the prevailing feature selection methods such as Mutual Information (MI), Chi-square Statistic (CHI), and Document Frequency (DF) in automatic text categorization. Experimental results performed on the well known dataset Reuters-21578 show that the ability for dimensionality reduction of LLDR compared with those of MI and CHI, and higher than that of DF. Considering that LLDR is more computational efficient than MI and CHI, LLDR is a promising feature selection method for automatic text categorization.

Keywords dimensionality reduction; feature selection; feature extraction; low loss dimensionality reduction; text categorization

收稿日期:2002-10-22;**修改稿收到日期**:2005-09-25.**宋枫溪**,男,1964 年生,博士,教授,主要研究领域为模式识别理论与应用. E-mail: fx_song@sina.com. **高秀梅**,女,1968 年生,博士,副教授,主要研究方向为模式识别与机器学习. **刘树海**,男,1942 年生,教授,博士生导师,主要研究领域为战场数据融合系统. **杨静宇**,男,1941 年生,教授,博士生导师,主要研究领域为模式识别与智能系统.

1 引 言

统计模式识别方法是模式识别理论中的主流方法,统计模式识别技术已成功应用于指纹识别、印刷体字符识别、语音识别、车牌识别等领域.统计模式识别在人脸识别、手写体字符识别、自动文本分类、多媒体数据挖掘等领域的应用研究也取得了长足进展[1].

统计模式识别的基本思想是首先将模式样本表示成线性空间中的向量,即特征向量.然后用训练样本对事先选定的分类算法或学习算法进行训练,直接或间接地提取出蕴涵在训练样本中有关各个模式类的统计特性,并根据这些特性确定出分类准则.最后依据这些准则对未知模式样本进行分类决策.显然,模式表示的准确如否,将严重影响模式识别效果.与句法(或结构)模式识别只需要少量的关键特征不同,统计模式识别则依赖于大量的非关键特征,即统计特征.

为了提高统计模式识别的正确识别率,人们通常需要采集数量巨大的原始特征,使得原始特征空间或输入空间的维数可能高达几千维或几万维.如果直接在输入空间上进行分类器训练,就可能带来两个棘手的问题:(1)很多在低维空间具有良好性能的分类算法在计算上变得不可行;(2)在训练样本容量一定的前提下,特征维数的增加将使得样本统计特性的估计变得更加困难,从而降低分类器的推广能力或泛化能力,呈现所谓的"过学习"或"过训练"的现象.

要避免出现"过学习"的情况,用于统计分类器训练的训练样本个数必须随着特征维数的增长而呈指数增长,从而造成人们所说的"维数灾难"(curse of dimensionality)[2].事实上,就两类分类问题,Hughes 给出了不同条件下贝叶斯分类器的期望识别率与特征度量复杂度以及训练样本容量之间的定量关系[3].

考虑以下这样一个特殊的两类分类问题.类先验概率均为 1/2,用于模式分类的特征共有 d 个,每个特征均为二元特征,即每个特征仅取两种不同的特征值.训练样本容量为 tn,训练样本中属于第一类和第二类的样本个数分别为 N_1,N_2($N_1\approx N_2\approx tn/2$).$P(d,tn)$表示贝叶斯分类器在训练样本容量为 tn,特征空间维数为 d 时的平均识别率,由文献[3]不难导出以下关系:

$$P(d,tn)=\sum_{r=0}^{tn/2}\sum_{s=0}^{tn/2}\left(\prod_{i=1}^{2^d-2}\frac{tn/2-r+i}{tn/2+i}\right)\left(\prod_{j=1}^{2^d-2}\frac{tn/2-s+j}{tn/2+j}\right)\cdot\frac{g(r,s)}{(tn/2+2^d-1)^2} \tag{1}$$

其中

$$g(r,s)=2^d(2^d-1)^2\frac{\max(s,r)}{tn+2(2^d-1)} \tag{2}$$

由式(1)和(2)我们可以绘出,不同训练样本容量下,贝叶斯分类器的平均识别率与特征维数之间的函数关系图(见图 1).从图 1 中可以看出,当特征个数增加到一定程度之后,继续增加特征维数将导致贝叶斯分类器平均识别率的持续下降.图 1 是在类先验概率均为 1/2 的条件下得到的,在一般情况下我们仍有类似的结论.

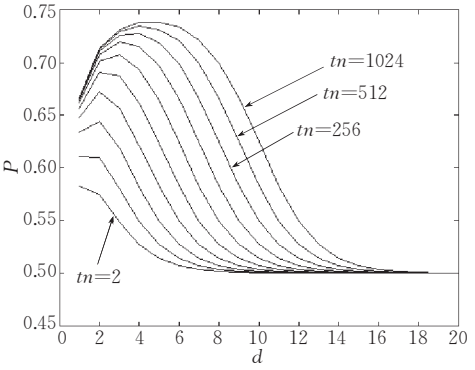

图 1 平均识别率随特征维数的变化规律

需要特别指出的是,图 1 所示的规律不仅适用于各类贝叶斯分类器,如朴素贝叶斯(Naive Bayes)、高斯(Gauss)分类器等,同样也适用于 knn(k 近邻)、SVM(支持向量机)等间接利用训练样本的统计特性进行分类的所谓几何分类器(geometric classifiers).虽然有文献报道 SVM 的推广能力几乎不受原始特征空间维数大小的影响[4],我们在相同数据集上的实验结果表明,当训练样本中正例个数较少时,适当降低特征空间维数有利于提高 SVM 分类器的推广能力.如对标准数据集 Retuers-21578① 中正例个数较少的后 80 类文本进行模式分类,当特征维数从 200 维降到 100 维或 50 维时,这 80 个类别的平均 BEP 值(break-even-point)则从 62.1%分别上升到 63.4%和 65.0%.

因此,如何降低原始特征维数,即如何进行维数削减(dimensionality reduction),一直成为统计模式识别的一个重要研究课题,得到了众多研究者的密切关注.

① Lewis D.. Reuters Collection. http://www.research.att.com/~lewis/reuters21578.html

2　特征选择与特征抽取

维数削减的根本任务就是将分散在各个原始特征中的有关模式类别的统计信息有效地集中起来，以达到提高正确识别率和降低计算工作量的目的. 维数削减有两种基本途径：特征选择(feature selection)与特征抽取(feature extraction). 特征选择——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的统计特性的相关特征. 特征抽取——依据某一原则构造输入空间到新的特征空间的一个变换，从而将分散在众多原始特征中的分类信息或鉴别信息集中到少量的新的特征上来.

2.1　特征选择

设 Y 是由所有 d 个原始特征构成的集合，r 是需要选出的特征个数，$J:2^Y\to R$ 为 Y 的幂集到实数集合的一个映射. 则特征选择问题可以表示为以下组合优化问题：

$$\max_{X\subset Y,\#(X)=r} J(X),$$

其中 $\#(X)$ 表示集合 X 的势，而判别函数 $J(X)$ 则反映了特征集合 X 的鉴别能力，它可以是反映不同类别数据之间可分性的巴氏(Bhattacharyya)距离[5]或 Chernoff 上界[6]，也可以是某一特定学习算法在这一特征子集上的正确识别率.

特征选择方法可分为筛选(filter)和复选(wrapper)两大类[7]. 特征筛选根据判别函数所得到的最优特征子集仅依赖于训练样本的统计特性，而与分类器所采用的学习算法无关. 相反，特征复选则依据分类器的学习算法在不同特征子集上的实际分类效果(正确识别率)，来评判各个特征子集的优劣. 显然，特征复选的结果不仅与训练样本的统计特性有关，而且与测试样本的统计特性和学习算法密切相关，因而比特征筛选复杂得多. 相对而言，特征筛选用得更多一些.

无论是特征筛选还是特征复选，从 d 个原始特征中选出势为 r 的最优特征子集的最简单也是最可靠的方法是逐一评估所有 C_d^r 个不同的特征子集，从中挑选出使判别函数 $J(X)$ 达到最大的那一个特征子集. 不容置疑，穷举法(Exhaustive Search，ES)是一种能确保找到最优特征子集的特征选择算法. 但令人遗憾的是，穷举法仅适用于非常有限的场合，当 d 较大时，该算法在计算上根本不可行(比如，取 $d=24$，$r=12$，则 $C_d^r\approx 2.7\times 10^6$). 另外一种可以保证找到最优特征子集的特征选择算法为分枝定界法(Branch-and-Bound Search，BB)[8]，其前提是判别函数为特征个数的单调增函数. 在现实世界中，这一前提往往难以得到满足. 另外，分枝定界法的算法复杂度与 d 仍是指数关系，当 d 较大时该算法仍不可行.

由于寻找最优解的代价太高，人们转向研究能得到较好次优解的特征选择算法. 基于各种启发式规则的次优搜索算法被相继提了出来. 比较著名的有基于贪婪算法的循序前向选择(Sequential Forward Selection，SFS)、循序后向选择(Sequential Backward Selection，SBS)和增减选择("Plus l-take away r" Selection，PTA)以及在这些算法基础上改进得到的循序前向浮动搜索(Sequential Forward Floating Selection，SFFS)、循序后向浮动搜索(Sequential Backward Floating Selection，SBFS)[9]. 此外有基于遗传算法的特征选择(Genetic Algorithm，GA)[10,11]、基于 Tabu 搜索的特征选择[12~14]以及基于数学规划的特征选择[15,16]等等. 不同的比较实验表明[17]①，当 $d<50$ 时，SFSS 和 SBFS 要优于 SFS、SBS、PTA 以及 GA 等算法. 而当 $d\geq 50$ 时，则基于遗传算法的特征选择方法体现出较大的优越性.

当原始特征维数不是很大时(如 $d<200$)，这些搜索算法在计算上是基本可行的. 但是，当特征维数高达几千维，甚至几万维时，它们就根本行不通了. 在自动文本分类、多媒体数据挖掘以及人脸识别等诸多应用领域中，原始特征的维数超过一万维是非常普遍的现象. 为了有效解决高维输入空间的降维问题，人们往往需要将寻找最优特征子集的目标降低为逐个寻找满足一定最优化准则的单个特征，然后将前 r 个最优特征拿来构成所需的特征子集. 这种方法显然不能保证找到最优特征子集，甚至也不能保证找到一个较好的次优特征子集. 但由于其计算量非常小，从而使得高维特征空间通过特征选择进行降维成为可能. Cover 曾经指出，即使在各个原始特征相互独立的假设条件下，各个最优特征构成的子集未必是最优特征子集[18]，但是自动文本分类的大量研究实践表明，基于单个特征优劣的特征选择方法仍是非常有效的[19]. 在众多单个特征的优选算法中，以互信息②(Mutual Information，MI)[20]、χ^2 统计(Chi-square Statistic，CHI)[21]以及文档频率

① Jain A.，Zongker D.. Feature Selection：Evaluation，Application，and Small Sample Performance. http://citeseer.nj.nec.com/context/293242/0

② 有的文献称信息增量.

(Document Frequency,DF)特征选择算法最为有效[22].

用互信息从 d 个原始特征中选取 r 个特征的过程如下:

首先,对于原始特征集中的每一个特征 f,用式(3)计算其与类别标号的互信息

$$MI(f,\omega)=\sum_{x}\sum_{\omega}p(x,\omega)\log\frac{p(x,\omega)}{p(x)p(\omega)} \quad (3)$$

其中,x 为 f 的观测值,ω 为 x 的类别标号.

然后,将互信息最大的 r 个特征挑选出来,构成所需的特征子集.

2.2 特征抽取

从 d 个原始特征中抽取 r 个新的特征,就是构造 $\varphi:R^d\to R^r$ 这样一个映射,使得 $J(\varphi(X))$ 达到最大. φ 可以是线性的也可以是非线性的. 多元统计理论中的很多方法被用来进行线性特征抽取,如主成分分析(Principal Component Analysis,PCA)、因子分析(Factor Analysis,FA)、多维标度(Multi-dimensional Scaling,MS)、投影追踪(Projection Pursuit,PP)[23]和 Fisher 鉴别分析(Fisher Discriminant Analysis,FDA)等[24]. 其中,以主成分分析和 Fisher 鉴别分析最为著名. 前者充分保留了原始数据中的二阶矩信息,是原始数据的一种最佳简约表示. 但是,其未能利用原始数据中的类别信息,使得降维后的数据有时反而不利于模式分类. 鉴别分析则克服了 PCA 的缺点,通过使 Rayleigh 商达到最大,从而同时实现类间散度最大和类内散度最小[25]. 近几年人们又提出了统计不相关的鉴别分析方法[26,27],以克服经典方法中的不足. 除了上述通用线性特征抽取方法外,还有一些用于特定应用领域的特征抽取方法,如自动文本分类中常用的基于矩阵奇异值分解的特征抽取方法[28].

在非线性特征抽取方法中,常用的有基于自组织映射(Self Organizing Maps,SOM)的特征抽取方法[29],基于核的主成分分析和基于核的鉴别分析(Kernel PCA 和 Kernel FDA,分别简记为 KPCA 和 KFDA)[30,31],前者通过自组织神经网络,后两者通过核函数(kernel function)实现输入空间到特征空间的非线性映射. 基于核的特征抽取及分类方法在国际上得到了广泛研究,取得了丰硕成果,近年来也引起了国内部分学者的关注.

在人脸识别、手写体字符识别、自动文本分类、多媒体数据挖掘等热点研究领域,不同特征抽取和选择算法得到了研究和应用. 相对来说,人脸识别与手写体字符识别使用特征抽取的机会更多一些,而自动文本分类和多媒体数据挖掘使用特征选择的机会更多一些.

图 2 概括了用于特征降维的一些主流方法. 图中 LLDR 代表本文提出的低损降维算法,具体内容见下一节.

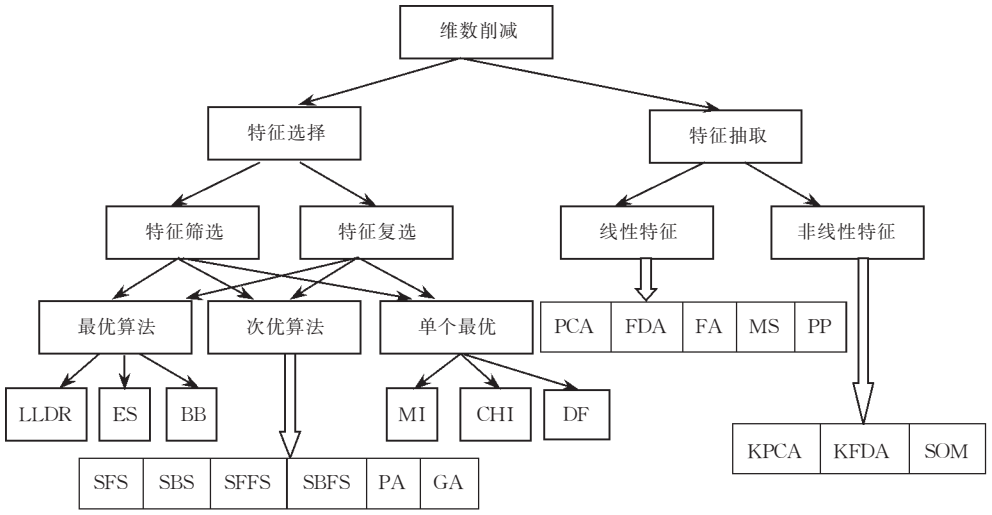

图 2 用于模式识别维数削减的一些常用算法

3 低损降维

低损降维(Low Loss Dimensionality Reduction,LLDR)是本文作者基于最优分类器,即贝叶斯分类器提出的一种新的特征选择方法.

我们考虑以下这么一个多类分类问题. 有 c 个模式类 $\omega_1,\omega_2,\cdots,\omega_c$,各个类的先验概率分别为 $P_1,P_2,\cdots,P_c$. 任务是将任意一个未知模式判别为这 c 类中的某一类. 假设 $f_1,f_2,\cdots,f_d$ 是 d 个模式特征,

对于任意给定的模式 s，其在特征集 $F=\{f_1,f_2,\cdots,f_d\}$ 下的观测值（即特征向量）$\boldsymbol{x}=(x_1,x_2,\cdots,x_d)$ 是一个 d 维随机变量. $\boldsymbol{x}$ 的所有可能取值即构成了所谓的特征空间，记做 $\chi=\chi_1\times\chi_2\times\cdots\times\chi_d$. 这里 χ_i 为变量 x_i 的取值范围，$i=1,2,\cdots,c$. 设 $p(\boldsymbol{x})$ 和 $p(\boldsymbol{x}|\omega_j)(j=1,2,\cdots,c)$ 为 $\boldsymbol{x}$ 的概率密度函数和类条件概率密度函数，贝叶斯分类器将模式 s 判别为模式类 i，当且仅当 s 的特征向量 $\boldsymbol{x}$ 满足条件 $p(\omega_i|\boldsymbol{x})=\max\limits_{1\le j\le c}p(\omega_j|\boldsymbol{x})$.

记 R_j 为 $\{\boldsymbol{x}\in\chi\mid p(\omega_j|\boldsymbol{x})=\max\limits_{1\le i\le c}p(\omega_i|\boldsymbol{x})\}$，则贝叶斯分类器的正确分类概率为

$$P_F=\sum_{j=1}^{c}P_jp(R_j|\omega_j).$$

定义 1.　特征 f 为冗余特征当且仅当

$$P_{F-\{f\}}=P_F.$$

定义 2.　特征 f 与特征集 $F-\{f\}$ 及类别 $\Omega\omega$ 相互独立，当且仅当对于任意给定的模式 s，其在 f 下的特征值 x 与其在 $F-\{f\}$ 下的特征向量 $\boldsymbol{x}^*$ 及其类别标号 ω 相互独立，即

$$p(x,\boldsymbol{x}^*,\omega)=p(x)p(\boldsymbol{x}^*,\omega).$$

定理 1.　如果特征 f 与 $F-\{f\}$ 及 ω 相互独立，则 f 为冗余特征.

证明.　令 $R_j^*=\{\boldsymbol{x}^*\in\chi^*\mid p(\omega_j|\boldsymbol{x}^*)=\max\limits_{1\le i\le c}p(\omega_i|\boldsymbol{x}^*)\}$，且

$$R_j=\{\boldsymbol{x}\in\chi\mid p(\omega_j|\boldsymbol{x})=\max_{1\le i\le c}p(\omega_i|\boldsymbol{x})\}.$$

由于

$$\begin{aligned}\boldsymbol{x}\in R_j&\Leftrightarrow\forall i,1\le i\le c,p(\boldsymbol{x}|\omega_j)\ge p(\boldsymbol{x}|\omega_i)\\&\Leftrightarrow\forall i,1\le i\le c,p(x,\boldsymbol{x}^*|\omega_j)\ge p(x,\boldsymbol{x}^*|\omega_i)\\&\Leftrightarrow\forall i,1\le i\le c,p(\boldsymbol{x}^*|\omega_j)\ge p(\boldsymbol{x}^*|\omega_i)\\&\Leftrightarrow\boldsymbol{x}^*\in R_j^*.\end{aligned}$$

于是我们有

$$P_F=\sum_{j=1}^{c}P_jp(R_j|\omega_j)=\sum_{j=1}^{c}P_jp(R_j^*|\omega_j)=P_{F-\{f\}},$$

即特征 f 为冗余特征.

推论 1.　如果特征 f 具有 0-1 分布，则 f 为冗余特征.

证明.　由于 0-1 分布的随机变量与任意随机变量独立，因而与 $F-\{f\}$ 和 ω 独立.

对于任意的二元特征 f，其分布率均可以写成以下形式：

f	x_1	x_2
P	p	$1-p$

当 p 接近 0 或 1 时，该概率分布可近似看成 0-1 分布. 于是相应的特征可以从特征集中剔除掉. 通常当 $p<0.01$ 或 $p>0.99$ 时，相应的特征即可剔除.

基于低损降维的特征选择算法.

设 Y 是原始特征集合，tn 为训练样本个数，r 是需要选择的特征个数，基于低损降维的特征选择过程如下：

1. 对于每一模式类 ω，统计出训练集中该模式类的正例个数 N_ω；

2. 对于特征集合 Y 中的每一个特征 f，统计出其在正例中出现的频数 $ptf(f)$；

3. 计算

$$lldr(f,\omega)=\max\{ptf(f)/N_\omega,(tn-ptf(f))/(tn-N_\omega)\}.$$

4. 取使得 $lldr(f,\omega)$ 达到最大的前 r 个特征.

4　文本分类实验

4.1　数据集及文本表示

我们采用 Lewis 编辑的路透社财经新闻语料库 Reuters-21578，遵循“ModApte”分割方式，将 Reuters-21578 分成训练集和测试集两部分，分别包含 9603 个训练样本和 3299 个测试样本. 略去没有类别标号或无正文的那些样本，实际用于训练和测试的样本个数分别为 7063 和 2742. 这里我们仅考虑前 10 类的分类问题，具体情况见表 1.

表 1　前 10 类训练和测试样本正例个数

类别	训练正例数	测试正例数
Earn	2709	1044
Acq	1488	643
Money-fx	460	141
Grain	394	134
Crude	349	161
Trade	337	112
Interest	289	100
Ship	191	85
Wheat	198	66
Corn	159	48

首先将训练集中所有不同的单词挑选出来，从中去除功能词（stop word），然后用 Porter 取词根算法取出这些单词的词根[32]，这样得到 15569 个不同的词根. 每篇文档表示为一个 15569 维向量，即特征向量. 其中每一个分量仅取 0，1 两种值. 0 表示对应的特征（即词根）没有在该文档中出现，1 表示对应的特征（即词根）在该文档中至少出现 1 次.

4.2　分类器及性能评估

研究表明，就文本分类而言，SVM 与 k 近邻相当或略优于它，明显优于其它分类器，如朴素贝叶

斯、决策树、最小距离等[19,33]. 这里,我们采用 SVM 作为评估各种特征选择算法优劣的标准分类器,Matlab 程序由 Ma 等提供①.

同其它模式分类不同,文本分类中的各个样本可能同属多个不同的类别. 常见的做法是将多类文本分类问题转化为多个不同的两类分类问题. 每次分类仅判断给定的某个样本究竟属不属于某个特定的类别. 评估分类效果分两步进行,第一步用 BEP 值或 $F1$ 值评估各个分类器的分类效果,然后用微平均(micro average)或宏平均(macro average)综合各个分类器的分类效果[34]. 这里,我们采用 BEP 值和微平均. 记

a——判断为正例的正例个数,

b——判断为正例的反例个数,

c——判断为反例的正例个数,

则查全率为

$$recall=\frac{a}{a+c} \tag{4}$$

查准率为

$$precision=\frac{a}{a+b} \tag{5}$$

查全率与查准率是互为制约的关系,一般来说提高查全率必然会降低查准率,反之亦然. 因此,要客观评估文本分类器的分类效果必须同时考察这两个指标. 所谓 BEP 值,即查全率与查准率相等时的值. 而微平均的计算则是将(4),(5)两式中的 a,b,c 分别改为所有判断为正例的正例个数和,所有判断为正例的反例个数和,所有判断为反例的正例个数和,然后取其 BEP 值.

4.3 实验结果

从表 2、表 3 和图 3 不难看出,LLDR 的识别效果与 MI 和 CHI 相当,但是优于 DF. 另外,由算法本身易知 LLDR 的计算复杂度要低于 MI 和 CHI.

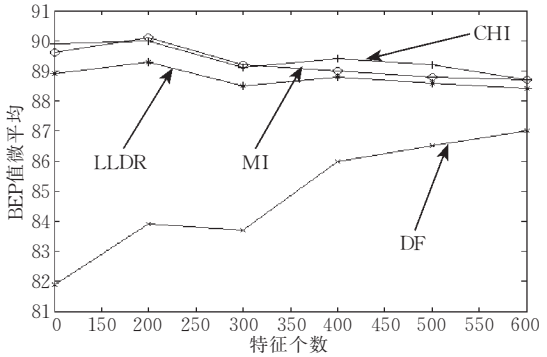

图 3 不同特征选择算法的前 10 类 BEP 值微平均随特征个数变化曲线

表 2 不同特征选择算法及不同数量特征下"Acq"类的 BEP 值的比较

r	BEP 值(%)			
	MI	CHI	LLDR	DF
100	93.0	92.5	91.1	89.9
200	93.6	93.3	94.4	92.4
300	93.5	93.5	93.8	93.5
400	93.5	93.5	93.0	92.7
500	93.3	93.5	93.3	92.9
600	94.4	93.0	92.5	92.4

表 3 不同特征选择算法及不同数量特征下前 10 类 BEP 值微平均的比较

r	BEP 值微平均(%)			
	MI	CHI	LLDR	DF
100	89.6	89.9	88.9	81.9
200	90.1	90.0	89.3	83.9
300	89.2	89.1	88.5	83.7
400	89.0	89.4	88.8	86.0
500	88.8	89.2	88.6	86.5
600	88.7	88.7	88.4	87.0

参 考 文 献

1 Jain A. K., Duin R. P. W., Mao Jian-Chang. Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis Machine Intelligence, 2000, 22(1): 4～37

2 Bishop C. M.. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995

3 Hughes G. F.. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory, 1968, 14(1): 55～63

4 Joachims T.. Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning (ECML), Chemnitz, DE, 1998, 137～142

5 Xuan Guo-Rong, Chai Pei-Qi. Feature selection based on Bhattacharyya distance. Pattern Recognition & Artificial Intelligence, 1996, 9(4): 324～329(in Chinese)
(宣国荣,柴佩琪. 基于巴氏距离的特征选择. 模式识别与人工智能, 1996, 9(4): 324～329)

6 Xuan Guo-Rong, Chai Pei-Qi. Feature selection based on Chernoff upper bound. Pattern Recognition & Artificial Intelligence, 1996, 9(1): 26～30(in Chinese)
(宣国荣,柴佩琪. 基于 Chernoff 上界的特征选择. 模式识别与人工智能, 1996, 9(1): 26～30)

7 Wang Hui, Bell D., Murtagh F.. Axiomatic approach to fea-

① Ma Jun-Shui, Zhao Yi, Ahalt Stanley. OSU SVM Classifier Matlab Toolbox (ver 3.00). http://eewww.eng.ohio-state.edu/～maj/osu_svm/

ture subset selection based on relevance. IEEE Transactions on Pattern Analysis Machine Intelligence, 1999, 21(3): 271～277

8 Narendra P. M., Fukunaga K.. A branch and bound algorithm for feature subset selection. IEEE Transactions on Computers, 1977, 26(9): 917～922

9 Pudil P., Novovicova J., Kittler J.. Floating search methods in feature selection. Pattern Recognition Letters, 1994, 15(11): 1119～1125

10 Siedlecki W., Sklansky J.. On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence, 1988, 2(2): 197～220

11 Liu Wei-Quan, Wang Ming-Hui, Zhong Yi-Xin. Feature dimensionality compression in hand written digit recognition by means of genetic algorithm. Pattern Recognition & Artificial Intelligence, 1996, 9(1): 45～51(in Chinese)
(刘伟权,王明会,钟义信. 利用遗传算法实现手写体数字识别中的特征维数的压缩. 模式识别与人工智能, 1996, 9(1): 45～51)

12 Glover F.. Tabu search I. ORSA Journal on Computing, 1989, 1: 190～206

13 Glover F.. Tabu search II. ORSA Journal on Computing, 1989, 2: 4～32

14 Zhang Hong-Bin, Sun Guang-Yu. Applications of Tabu search in feature selection. Acta Automatica Sinica, 1999, 25(4): 457～466(in Chinese)
(张鸿宾,孙广煜. Tabu 搜索在特征选择中的应用. 自动化学报, 1999, 25(4): 457～466)

15 Foroutan I., Sklansky J.. Feature selection for automatic classification of non Gaussian data. IEEE Transactions on System, Man and Cybernetics, 1987, 17(2): 187～198

16 Zhang Xin-Hua. A dynamic programming method for feature selection. Acta Automatica Sinica, 1998, 24(5): 675～680(in Chinese)
(章新华. 一种特征选择的动态规划方法. 自动化学报, 1998, 24(5): 675～680)

17 Kudo M., Sklansky J.. Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, 2000, 33(1): 25～41

18 Cover T. M. The best two independent measurements are not the two best. IEEE Transactions on Systems, Man, and Cybernetics, 1974, 4: 116～117

19 Dumais S., Platt J., Heckerman D., Sahami M.. Inductive learning algorithms and representations for text categorization. In: Proceedings of the CIKM-98, 7th ACM International Conference on Information and Knowledge Management, Bethesda, MD, 1998, 148～155

20 Lewis D. D.. An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval, Copenhagen, 1992, 246～254

21 Schutze H., Hull D. A., Pedersen J. O.. A comparison of classifiers and document representations for the routing problem. In: Proceedings of the SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, Seattle, WA, 1995, 229～237

22 Yang Y., Pedersen J. O.. A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference(ICML'97), Nashville, TN, 1997, 412～420

23 Intrator N.. Localized exploratory projection pursuit. In: Proceedings of the 23rd Symposium on the Interface, Seattle, WA, 1991, 237～240

24 Fang Kai-Tai. Applied multivariate statistical analysis. Shanghai: East China Normal University Press, 1989(in Chinese)
(方开泰. 实用多元统计分析. 上海:华东师范大学出版社, 1989)

25 Foley D. H., Sammon J. W.. An optimal set of discriminant vectors. IEEE Transactions on Computing, 1975, 24(3): 281～289

26 Jin Zhong, Yang Jing-Yu, Hu Zhong-Shan, Lou Zhen. Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition, 2001, 34(7): 1405～1416

27 Jin Zhong, Yang Jing-Yu, Lu Jian-Feng. An optimal set of uncorrelated discriminant features. Chinese Journal of Computers, 1999, 22(10): 1105～1108(in Chinese)
(金　忠,杨静宇,陆建峰. 一种具有统计不相关的最佳鉴别矢量集. 计算机学报, 1999, 22(10): 1105～1108)

28 Wiener E., Pedersen J. O., Weigend A. S.. A neural network approach to topic spotting. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, 1995, 317～332

29 Haykin Smon. Neural Networks: A Comprehensive Foundation. Second Edition. Beijing: Tsinghua University Press, 2001

30 Scholkopf B., Smola A., Mulle K. R.. Nonlinear component analysis as a kernel eigenvalue problem. Max-Planck-Institute, Germany: Technical Report No. 44, 1996

31 Yang Jian, Frangi Alejandro F., Yang Jing-Yu, Zhang David, Jin Zhong. KPCA plus lda: A complete kernel fisher discriminant framework for feature extraction and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(2): 230～244

32 Porter M. F.. An algorithm for suffix striping. Program, 1980, 14(3): 130～137

33 Yang Yi-Ming. An evaluation of statistical approaches to text categorization. Information Retrieval, 1999, 1(1～2): 69～90

34 Sebastiani F.. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1～47

SONG Feng-Xi, born in 1964, Ph.D., professor. His current research interests include pattern recognition theory and its applications.

GAO Xiu-Mei, born in 1968, Ph.D., associate professor. Her current research interests include pattern recognition and machine learning.

LIU Shu-Hai, born in 1942, professor, Ph.D. supervisor. His current research interests include fusion system of battlefield data.

YANG Jing-Yu, born in 1941, professor, Ph.D. supervisor. His current research interests include pattern recognition and intelligent systems.

Background

One of the major research interests of the research group is "Studies on Some Essential Problems in Automatic Text Categorization". It includes text representation, text classifier design, evaluation of text categorization system performance, and feature selection. We have published a series of papers in this field.

This paper focuses on the problem of curse of dimensionality which text classifiers usually have to confront with.