# 特征向量的归一化比较性研究

肖汉光[1],蔡从中[2]
XIAO Han-guang[1],CAI Cong-zhong[2]

1.重庆工学院 数理学院,重庆 400054
2.重庆大学 数理学院,重庆 400044
1.School of Mathematics and Physics,Chongqing Institute of Technology,Chongqing 400054,China
2.School of Mathematics and Physics,Chongqing University,Chongqing 400044,China
E-mail:simenxiao1211@163.com

**XIAO Han-guang,CAI Cong-zhong.Comparison study of normalization of feature vector.Computer Engineering and Applications,2009,45(22):117-119.**

**Abstract**: Feature extraction and the parameter optimization of classifiers are two key methods for the improvement of the classification accuracy.The paper uses normalization method for the feature transformation based on the public database UCI. KNN,PNN and SVM are employed for classification.The effects of normalization on the accuracy of classification and parameter optimization are discussed.The results of experiment show normalization improved effectively the accuracies of classifiers,especially for SVM,reduce the searching range of the parameters of classifiers and the training periods.
**Key words**: normalization;feature vector;parameter optimization;Support Vector Machine(SVM)

**摘 要**: 特征提取和分类器的参数优化是提高分类准确率的主要途径, 对公用数据库 UCI 的相关数据进行特征向量的归一化处理,采用 KNN、PNN 和 SVM 进行分类。讨论了特征归一化对分类准确率和分类器参数的影响。实验结果表明:归一化能有效提高分类器的分类准确率,SVM 尤为明显,且参数的寻优范围缩小,缩短训练周期。
**关键词**:归一化;特征向量;参数优化;支持向量机
**DOI**:10.3778/j.issn.1002-8331.2009.22.038 **文章编号**:1002-8331(2009)22-0117-03 **文献标识码**:A **中图分类号**:TP18

## 1 前言

归一化处理在模式识别中应用十分广泛,其用途主要分为两类:一类是归一化作为特征提取前的预处理技术;另一类是归一化对特征提取后的特征向量进行特征变换。归一化作为数据预处理技术常用于特征的产生和提取,如人脸识别、虹膜识别、车牌识别和手写字体识别等,其主要作用为统一识别对象的大小和尺寸[1-2]。由于归一化预处理的好坏直接影响特征生成和提取的效果,所以归一化预处理技术始终是研究者讨论的热点。从广义上讲,特征向量的特征归一化是一种特征变换。由于识别对象的不同,其特征向量的特征分量在数量级上有较大的差别。在代价函数中,大值特征分量比小值特征分量的影响更大,但并不能反映大值特征分量更重要,所以需要对特征进行数量级统一,即特征归一化。由于未采用特征归一化的特征向量能得到较为满意的结果, 所以特征归一化往往容易被忽视,讨论特征归一化对识别率影响的研究相对较少。但从提高识别率的角度看,特征归一化是值得讨论的。归一化后的特征向量在特征空间中的分布将发生相应的改变,一方面,不同的分类器对该变化都有不同的响应,即分类器的识别准确率发生不同程度的变化,另一方面,相关分类器的参数优化范围也发生改变,这也将影响寻优时间。

对 14 个分类问题进行特征归一化, 讨论其对三种常用分类器的分类准确率的影响。对于多类分类问题采用一对多的分类策略。为减小数据分组对分类准确率的影响,采用了 5 次交叉验证的测试方法。

## 2 归一化方法

设训练样本集为$\{\boldsymbol{x}_i\}$,测试样本集为$\{\boldsymbol{x}_i'\}$,训练样本集所有样本各分量的最大值、最小值和平均值分别构成向量$\boldsymbol{x}_{\max}$、$\boldsymbol{x}_{\min}$和$\bar{\boldsymbol{x}}$。对训练样本集中的任一样本$\boldsymbol{x}_i$进行如式(1)或式(1')归一化,测试集样本$\boldsymbol{x}_i'$的归一化和$\boldsymbol{x}_i$相同。

$$\boldsymbol{x}_i=\frac{2\boldsymbol{x}_i-\boldsymbol{x}_{\max}-\boldsymbol{x}_{\min}}{\boldsymbol{x}_{\max}-\boldsymbol{x}_{\min}} \tag{1}$$

**基金项目**:国家教育部新世纪人才支持计划 (the New Century Excellent Talent Foundation from MOE of China under Grant No.NCET-07-0903);重庆市自然科学基金(the Natural Science Foundation of Chongqing City of China under Grant No.2006BB5240);重庆工学院青年教师科研基金(the Young Teacher Scientific Research Foundation of Chongqing Institute of Technology under Grant No.20062D39)。
**作者简介**:肖汉光(1980-),男,硕士,新加坡国立大学访问学者,主要研究方向:机器学习、模式识别等;蔡从中(1966-),男,博士,研究员,博士生导师,主要研究方向:人工智能和机器学习、计算物理学、计算生物信息学等。
**收稿日期**:2008-04-28 **修回日期**:2008-09-16

或 $$x_i=\frac{N(x_i-\bar{x})}{\sum_{i}^{N}(x_i-\bar{x})} \quad (1')$$

其中,$N$ 为训练(或测试)样本个数归一化后,训练和测试样本的分量值介于-1 和+1 之间。或采用式(2)将数据归一化到[0,1]。

$$x_i=\frac{|x_i-x_{\min}|}{|x_{\max}-x_{\min}|} \quad (2)$$

## 3 分类原理

### 3.1 KNN 的分类原理

KNN 和其他分类方法相比是最简单但准确率较高的分类器。该方法遵从的假设为:同类样本在特征空间中距离相近,而异类的样本距离较远。若给定一待分类的 $L$ 维样本 $x'=(x_1',x_2',\cdots,x_L')$,计算其与训练样本$\{x_i\}$(即已知类别的样本)的相似度或距离,如式(1)为待测样本与测试集中第 $i$ 个样本欧氏距离。

$$S_i=\|x'-x_i\| \quad (3)$$

由 $K$ 个最相似或接近的样本根据自身类别进行少数服从多数的投票决定待识别样本的类别。一般 $K$ 取 1 到 $N$($N$ 为训练样本的样本数)。

### 3.2 PNN 的分类原理

PNN 是根据贝叶斯最优决策规则而设计的分类方法,由输入层、径向基层、比较层和输出层组成[3]。当待测样本输入到输入层,和径向基层的所有神经元进行运算,计算其与神经元的距离,神经元一般设定为训练集中的各样本。在比较层中进行距离比较,计算待测样本与所有正和负样本神经元的平均距离,若与正样本神经元的平均距离小于负样本神经元的平均距离,则输出为正类别,反之为负类别。实际 PNN 相当 $K$ 为 $N$($N$ 为训练集的样本数)时的 KNN,但计算距离的表达式略有不同。式(2)为径向基层中计算待测样本与神经元的距离公式。

$$S_i=\exp(-\frac{\|x'-x_i\|}{2\sigma^2}) \quad (4)$$

其中 $g=-1/2\sigma^2$ 为伽玛参数,在训练 PNN 时,需进行该参数优化,一般采用网格搜索法。

### 3.3 SVM 的分类原理

支持向量机(Support Vector Machine,SVM)建立在统计学习理论的 VC 维(Vapnik Chervonenks Dimension)理论和结构风险最小原理(Structural Risk Minimization)基础上,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力[4-5]。

以两类(正样本和负样本)分类问题为例,在线性可分的情况下,SVM 构建一个超平面 $H$:

$$w\cdot x+b=0 \quad (5)$$

式中,$w$ 为权重向量,$x$ 为特征向量,$b$ 为一参数。该超平面以最大边界的形式将正负样本区分开。该超平面的构建是通过寻找向量 $w$ 和参数 $b$,使其在满足条件

$$w\cdot x_i+b\geqslant 0,(\text{对正样本},y=+1) \quad (6)$$

$$w\cdot x_i+b<0,(\text{对负样本},y=-1) \quad (7)$$

时,$\|w\|^2$ 达到最小。式中 $x_i$ 代表第 $i$ 个训练样本的特征向量,$\|w\|^2$ 代表权重向量 $w$ 的欧几里德范数,$y$ 为样本类别标记。在求出 $w$ 和 $b$ 后,通过决策函数

$$y_i=sign[w\cdot x_i+b] \quad (8)$$

判断向量 $x_i$ 所对应测试样本的类别。若决策函数值为+1,该样本属于正样本;否则,属于负样本。

在线性不可分的情况下,SVM 利用核函数 $K(x_i,x_j)$将特征向量映射到一个高维空间。在此高维空间中,线性不可分问题被转化为线性可分问题,其决策函数为:

$$y_j=sign[\sum_{i=1}^{l}\alpha_i y_i K(x_i,x_j)+b] \quad (9)$$

上式中,$l$ 为训练样本数,系数 $\alpha_i$ 和 $b$ 应使拉格朗日表达式:

$$\sum_{i}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \quad (10)$$

达到最大值,且应满足:

$$C>\alpha_i\geqslant 0 \text{ 和 } \sum_{j=1}^{l}\alpha_i y_i=0 \quad (11)$$

其中,$C$ 为错误惩罚参数,它控制对错误分类样本的惩罚程度,$C$ 越大支持向量的个数越多,最优超平面越复杂。

核函数 $K(x_i,x_j)$一般取径向基函数:

$$K(x_i,x_j)=\exp(\frac{-\|x_i,x_j\|^2}{2\sigma^2}) \quad (12)$$

一般训练过程中需要对径向基函数中的参数 $g=-1/2\sigma^2$ 进行优化,大多采用的方法为网格搜索法。

## 4 实验及分析

本实验数据来自公用数据库 UCI(http://archive.ics.uci.edu/ml),选择了具有代表性的 14 个分类问题,如表 1 所示,其中二类和多类分类问题分别为 5 个和 9 个。本实验归一化采用式(1)。对于多类分类问题采用一对多的方法,即轮流选择其中一类样本为正样本,其他类别作为负样本。每次训练和测试采用 5 次交叉验证,即将正负样本分为 5 等份,轮流选择其中一份作为测试集,其他 4 份作为训练集,每等份中均有适量的正负样本。

在 KNN 的测试中,选择 $K$ 为 3,距离公式采用欧氏距离。在 PNN 和 SVM 的训练中,$g$ 搜索值为:[0.000 1:0.000 1:0.001,0.002:0.001:0.01,0.02:0.01:0.1,0.2:0.1:1],即不同数量级上等公差搜索,其中 SVM 训练中 $C$ 取 10 000。

在每次交叉验证中,设 $TP$(True Positive)代表在测试集中被判断正确的正样本个数;$FN$(False Negative)代表在测试集中被错判为正样本的个数;$TN$(True Negative)代表在测试集中被判断正确的负样本个数;$FP$(False Positive)代表在测试集中被错判为负样本的个数。

第 $j$ 个分类问题的 5 次交叉验证中第 $i$ 次交叉验证的测试准确率公式为:

$$Q_j^i=\frac{TP+TN}{TP+FN+TN+FP} \quad (13)$$

分类器对第 $j$ 个分类问题的测试准确率为:

$$Q=\frac{1}{C\times 5}\sum_{j=1}^{C}\sum_{i=1}^{5}Q_j^i \quad (14)$$

其中,$C$ 为第 $j$ 个分类问题的类别数。

归一化前后,KNN、PNN 和 SVM 对不同分类问题的 5 次交叉验证的测试准确率如表 1 所示。从表 1 可以看出特征归一化

**表 1 归一化前后 KNN、PNN 和 SVM 对 5 次交叉验证的平均准确率**

| Database | Num | Dim | Class | KNN | | PNN | | SVM | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $Q_1$/(%) | $Q_2$/(%) | $Q_1$/(%) | $Q_2$/(%) | $Q_1$/(%) | $Q_2$/(%) |
| auto-mpg | 398 | 8 | 3 | 81.6 | 83.2 | 82.7 | 85.6 | 84.2 | 90.4 |
| breast-w | 699 | 10 | 2 | 96.6 | 96.6 | 97.4 | 97.4 | 97.1 | 97.3 |
| clear1 | 476 | 167 | 2 | 85.1 | 83.0 | 87.0 | 88.0 | 85.1 | 96.2 |
| diabetes | 768 | 9 | 2 | 70.4 | 72.9 | 73.6 | 77.2 | 74.8 | 78.5 |
| flag | 194 | 28 | 6 | 81.1 | 82.8 | 78.4 | 82.9 | 86.7 | 88.4 |
| glass | 214 | 10 | 6 | 89.1 | 89.5 | 91.9 | 90.4 | 92.1 | 92.3 |
| hayes-roth | 132 | 5 | 3 | 74.2 | 71.4 | 84.8 | 84.3 | 90.4 | 90.6 |
| heart-cleveland | 303 | 14 | 5 | 81.3 | 83.6 | 78.5 | 82.5 | 85.3 | 87.9 |
| heart-statlog | 270 | 14 | 2 | 66.7 | 78.9 | 68.9 | 83.7 | 73.7 | 84.8 |
| iris | 150 | 5 | 3 | 97.3 | 96.9 | 98.2 | 98.0 | 98.0 | 98.4 |
| segment | 2 310 | 20 | 7 | 98.7 | 98.8 | 99.0 | 99.2 | 99.0 | 99.4 |
| sonar | 208 | 61 | 2 | 82.7 | 83.1 | 88.9 | 90.9 | 92.3 | 92.8 |
| vehicle | 846 | 19 | 4 | 83.1 | 85.4 | 82.8 | 85.3 | 89.5 | 92.9 |
| wine | 178 | 14 | 3 | 80.1 | 97.4 | 85.0 | 98.5 | 91.9 | 99.2 |

注:Num、Dim 和 Class 分别代表样本总数、向量维数和类别数,$Q_1$ 和 $Q_2$ 分别代表分类器归一化前和归一化后的分类准确率。

**表 2 不同 $g$ 参数搜索范围下,归一化前后 PNN 和 SVM 的分类平均准确率和归一化对平均准确率**

| $g$ | Normalization | PNN | | SVM | |
|---|---|---|---|---|---|
| | | $\overline{Q}$/(%) | $\Delta\overline{Q}$/(%) | $\overline{Q}$/(%) | $\Delta\overline{Q}$/(%) |
| [0.000 1,1] | NO | 85.5 | 3.4 | 88.3 | 3.8 |
| | YES | 88.9 | | 92.1 | |
| [0.000 01,10] | NO | 85.8 | 3.3 | 90.1 | 1.7 |
| | YES | 89.1 | | 92.2 | |

后 KNN 和 PNN 对除少数分类问题的测试准确率略有下降外,其他分类问题的测试准确率都有一定的提高。相比之下,特征归一化后 SVM 对所有分类问题的测试准确率均高于特征归一化前,平均提高 3.8%。另外,对同一分类问题,SVM 得到了绝大多数的最高准确率,除 breast-w 外。表 2 为扩大 $g$ 参数的搜索范围前后,归一化前后 PNN 和 SVM 对各分类问题的平均测试准确率。从表 2 的 $\Delta\overline{Q}$ 可以看出 $g$ 参数的搜索范围改变对特征归一化前后的 PNN 的分类准确率影响较 SVM 小。另外从表 2 中 SVM 的 $\overline{Q}$ 可以看出 $g$ 参数搜索范围的改变对特征归一化前的分类准确率影响加大,而对归一化后的分类准确率影响较小。

归一化前后,PNN 和 SVM 的 $g$ 参数在各次交叉验证的训练中最优值概率分布如图 1 所示。从图中可以看出,经过归一化后,SVM 和 PNN 的 $g$ 参数分布较归一化前 $g$ 参数的搜索范围得到了减小。对 SVM 而言,在归一化前 $g$ 参数的最优值搜索范围很大,约 43%的最优值处在搜索边界上,归一化后约 32%的最优值处在搜索边界上。但是,在特征归一化后,扩大 $g$ 参数的搜索范围对 SVM 的分类准确率提高并不太大,从表 2 中可以得出该结果。

## 5 结论

通过对 14 个分类问题的研究表明:在特征归一化后,相对

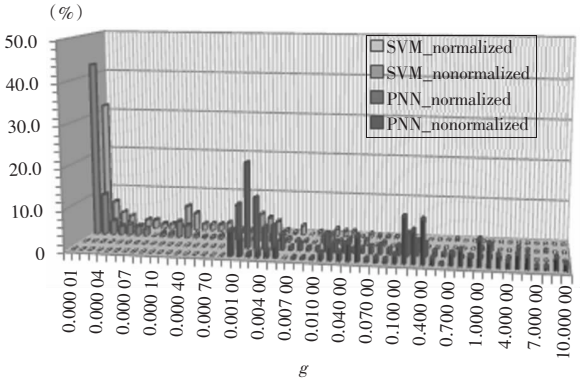


**图 1 归一化前后,PNN 和 SVM 的 $g$ 参数在各次交叉验证的训练中最优值的概率分布**

KNN 和 PNN 能提高大部分分类问题的识别率而言,SVM 识别率提高更为普遍和明显,并且 SVM 获得了绝大多数分类问题的最高准确率。在分类器的识别率改善的同时,特征归一化使得 SVM 和 PNN 的最优参数搜索范围变小,大大缩减了分类器的训练时间。

## 参考文献:

[1] 王先梅,王宏,王粉花.基于归一化背景方向特征的脱机手写汉字识别[J].计算机工程与应用,2007,43(30):190-192.

[2] 刘小平,赖剑煌,张智斌.基于小波子带图像的人脸光照归一化方法[J].中山大学学报:自然科学版,2007,46(5):25-28.

[3] Specht D F.Probabilistic neural networks[J].Neural Networks,1990,3:109-118.

[4] Vapnik V.The nature of statistical learning theory[M].New York:Springer,1995.

[5] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.