第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)23—0172—03 文献标识码：A 中图分类号：TP18

基于Box-Cox变换的分类器性能改进

李建刚，吴小俊
(江南大学信息工程学院，无锡 214122)

摘 要：贝叶斯分类器、最小距离分类器、近邻分类器和BP网络等是比较常用的分类器，为提高这些分类器的性能，引入了Box-Cox变换的思想。将Box-Cox变换用于数据正态化处理技术，并对常用分类器的性能进行改进。实验结果显示，通过引入Box-Cox变换，分类器的分类正确率有较大的提高。

关键词：Box-Cox变换；贝叶斯分类器；近邻分类器；最小距离分类器；BP神经网络

Improvement of Classifier Performance Based on Box-Cox Transformation

LI Jian-gang, WU Xiao-jun
(School of Information Engineering, Jiangnan University, Wuxi 214122)

【Abstract】Bayes classifier, minimum distance classifier, nearest neighbour classifier and back propagation neural networks are widely used classifiers. In order to improve their performance, this paper introduces the idea of Box-Cox transformation. Box-Cox transformation, which can transform the data and make the distribution nearer normal distribution, is a simple but quite effective data processing technology. Experiment results show that, as the result of the introducation of Box-Cox transformation, the accurate rate of classifiers is improved remarkably.

【Key words】Box-Cox transformation; Bayes classifier; nearest neighbour classifier; minimum distance classifier; BP neural network

1 概述

在模式识别中，分类决策就是在特征空间中用统计方法把被识别对象归为某一类。通常，人们在样本训练集的基础上确定某个判别规则，使按这种判别规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小[1]。通过训练样本确定的这个判别规则，通常称为分类器。迄今为止，已经发展了很多的分类器，例如决策树[1]、朴素贝叶斯、神经网络、支持向量机等。这些分类器分别采用不同的学习算法建立模型，各个模型应该尽量拟合输入数据集的属性集与类别之间的关系。解决分类问题的过程一般要分为2个阶段：(1)采用学习算法，通过对训练集进行归纳学习得到分类模型；(2)将已经学习得到的分类模型用于测试集，对测试集中未知类别的实例进行分类。很显然，通过训练集得到的分类模型未必是最佳的，这就会导致对测试集的分类可能会产生错误。而人们希望得到性能更好的分类器，为了达到这个目标，通常有2种做法：发明新的分类器和对现有的分类器的进行改进。一直以来，人们研究的重点主要集中在前者，并提出了一些新算法思想，但对常用的一些分类器的性能改进就被人们忽略了。

在众多的分类器中，常用的有最小距离分类器、近邻分类器[2]、贝叶斯分类器等，另外一些简单的神经网络(如BP网络)在类型判别上的应用也是十分普遍的。在统计分析中，很多重要的结论都是在样本总体具有相同方差的正态分布这个假设情况下得到的，但是在通常情况下这种假设未必成立的。一种可行的方法是变换样本总体，使其更加接近假设的条件[3]。1964年，Box和Cox为了降低样本的偏离正态性，发明了一种基于参数的数学变换公式，这就是通常所说的Box-Cox变换，现在这种技术已经被广泛的研究[3]。本文引入了Box-Cox变换，通过Box-Cox变换在同等条件下对一些常用的分类器的性能进行改进。

2 基于Box-Cox变换的分类器研究

2.1 Box-Cox变换

假设样本 x 是一维的，x^{λ} 是经过Box-Cox变换以后的样本(λ 为变换参数)，则它们之间的关系为

$$x^{\lambda}=\begin{cases}\dfrac{\exp(\lambda x)-1}{\lambda} & \lambda\neq 0\\ x & \lambda=0\end{cases} \tag{1}$$

假设有 N 个样本观测值 $x_1, x_2, \cdots, x_k, \cdots, x_N$，在该样本集是正态分布的前提下，可以用极大似然估计法估计出式(1)中的变换参数 λ；但是实际操作中不实用，一般情况下采用如下 $\bar{\lambda}$ 表示 λ：

$$\bar{\lambda}=\frac{6m_3}{3(m_2)^2-7m_4} \tag{2}$$

其中，$\bar{x}=\frac{1}{N}\sum_{k=1}^{N}x_k$；$m_r=\frac{1}{N}\sum_{k=1}^{N}(x_k-\bar{x})^r$ (r=2,3,4)。在本文实验中，$\bar{\lambda}$ 是每一类的每一维计算出来的。

基金项目：国家自然科学基金资助项目(60472060, 60572034)；2006年教育部新世纪优秀人才计划基金资助项目(NCEG-06-0487)；江苏省自然科学基金资助项目(BK2006081)；江南大学创新团队研究计划基金资助项目(JNIRT0702)

作者简介：李建刚(1984－)，男，硕士研究生，主研方向：人脸识别；吴小俊，教授、博士、博士生导师

收稿日期：2009-07-09 **E-mail：**lijianganghenan@163.com.cn

2.2 基于Box-Cox变换的正态贝叶斯分类器

假设 M 类总体样本 ω_1，$\omega_2,\cdots,\omega_M$，每一类的均值为 μ_1，$\mu_2,\cdots,\mu_M$；则对于测试样本 x：

$$g_i(x)=\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\sum_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+\frac{1}{2}\ln|\sum_i|-\ln P(\omega_i) \tag{3}$$

其中，$i=1,2,\cdots,M$；$\sum_i$ 是协方差矩阵。式(3)是基于正态分布的贝叶斯分类器[4]。考虑到 Box-Cox 变换对样本集的影响，把式(3)表示成 Box-Cox 变换以后的形式：

$$g_i'(x)=\frac{1}{2}(\boldsymbol{x}^{\overline{\lambda}_i}-\boldsymbol{\mu}_i^{\overline{\lambda}_i})^{\mathrm{T}}\sum_i^{\overline{\lambda}_i-1}(\boldsymbol{x}^{\overline{\lambda}_i}-\boldsymbol{\mu}_i^{\overline{\lambda}_i})+\frac{1}{2}\ln\left|\sum_i^{\overline{\lambda}_i}\right|-\sum_{j=1}^{l}\overline{\lambda}_{ij}x_j \tag{4}$$

其中，$\boldsymbol{x}^{\overline{\lambda}_i}$，$\boldsymbol{\mu}_i^{\overline{\lambda}_i}$，$\sum_i^{\overline{\lambda}_i}$ 分别是与 x，μ_i，$\sum_i$ 相对应的经过 Box-Cox 变换后的向量和矩阵，$\overline{\lambda}_{ij}$ 是第 i 类第 j 维的 Box-Cox 变换参数。

2.3 基于Box-Cox变换的最小距离分类器

在各种线性分类器中，最小距离分类器是最被普遍使用的分类器之一。如果有 M 类总体样本 ω_1，$\omega_2,\cdots,\omega_M$，每一类的均值为 μ_1，$\mu_2,\cdots,\mu_M$，对于其中任意的样本 x，相对于第 i 类的距离

$$d_i=|\boldsymbol{x}-\boldsymbol{\mu}_i| \tag{5}$$

对所有的 d_i 取最小的，就判别 x 属于第 i 类。

本文引入了 Box-Cox 变换，距离的定义就变成：

$$d_i'=\left|\boldsymbol{x}^{\overline{\lambda}_i}-\boldsymbol{\mu}_i^{\overline{\lambda}_i}\right| \tag{6}$$

进而通过式(6)进行分类识别。

2.4 基于Box-Cox变换的近邻分类器

假定有 M 类总体样本 ω_1，$\omega_2,\cdots,\omega_i,\cdots,\omega_M$，$\omega_i$ 类有标明类别的样本 N_i 个，可以规定 ω_i 类的判别函数为

$$g_i(x)=\min_k\left|x-x_i^k\right|,k=1,2,\cdots,N_i \tag{7}$$

其中，x_i^k 的角标 i 表示 ω_i 类；k 表示 ω_i 类 N_i 个样本中的第 k 个。依据式(7)，如果有

$$g_j(x)=\min_i\left(g_i(x)\right),i=1,2,\cdots,M \tag{8}$$

则判定 x 属于 ω_i 类。

通过 Box-Cox 变换，可以规定 ω_i 类的判别函数为

$$g_i'(x)=\min_k\left|\boldsymbol{x}^{\overline{\lambda}_i}-x_i^{k\,\overline{\lambda}_i}\right|,k=1,2,\cdots,N_i \tag{9}$$

其中，$\boldsymbol{x}^{\overline{\lambda}_i}$ 是 x 经过 Box-Cox 变换得到的；$\overline{\lambda}_i$ 是对应的变换参数；$x_i^{k\,\overline{\lambda}_i}$ 是 x_i^k 经过 Box-Cox 变换得到的。因此，如果有

$$g_j'(x)=\min_i\left(g_i'(x)\right),i=1,2,\cdots,M$$

(1 0)

则判定 x 属于 ω_i 类。

2.5 Box-Cox变换在BP网络中的应用

BP 神经网络在结构上类似于多层感知器，是一种多层前馈神经网络。它的名字来源于在神经网络训练中，调整网络权值的训练算法是误差反向传播学系算法，即 BP 算法。BP 学习算法是 Rumelhart 等人于 1986 年提出的。自此以后，由于结构简单、可调参数多、训练算法多、可操控性好，BP 神经网络获得了广泛的实际应用。

BP 网络的性能不仅取决各个神经元连接权值初始值和训练函数的选取，而且训练样本对整个网络连接权值的调整以及相对于测试样本的输出都有一定的影响。Box-Cox 变换正是对同一类样本进行正态化处理，从而使分散的数据集中在一个较小的区域，进而使 BP 网络的收敛速度加快，分类准确性大幅提高。

3 实验结果与分析

3.1 实验

本实验选用的实验数据是 IRIS 数据集。IRIS 是一个 3 类 4 维数据集，共 150 个数据，它是一个典型的数据分类实验样本。

实验步骤(贝叶斯分类器、最小距离分类器、近邻分类器)：

Box-Cox 变换前：首先对训练集用 LDA[5]求出鉴别子空间，其次让测试集在该鉴别子空间上进行投影，最后分别用贝叶斯分类器、最小距离分类器、近邻分类器进行分类。

Box-Cox 变换后：首先对从训练样本通过式(2)计算出 Box-Cox 变换系数，然后对总共的 150 个数据进行 Box-Cox 变换，其次按照如上相同的步骤进行实验，只是最后的分类器使用 Box-Cox 变换以后的形式。

实验步骤(BP 网络)：

Box-Cox 变换前：用 IRIS 数据集每一类的前 25 个组成训练集，其余的组成测试集。首先建立一个三层的 BP 神经网络，中间层和输出层的神经元个数分别为 5 和 1，传递函数的可以任意选择，这个实验中使用输入层到中间层的传递函数为正切函数，中间层到输出层的传递函数为线性函数。然后用训练集和目标输出(每一类的类别 1, 2, 3)训练网络，最后用测试集作为输入进行测试。

Box-Cox 变换后：先用训练集通过式(2)计算出变换参数，然后对训练集和测试集都进行变换，其余步骤与 Box-Cox 变换前相同。

3.2 结果与分析

贝叶斯分类器、最小距离分类器、近邻分类器的实验效果如图 1~图 3 所示。

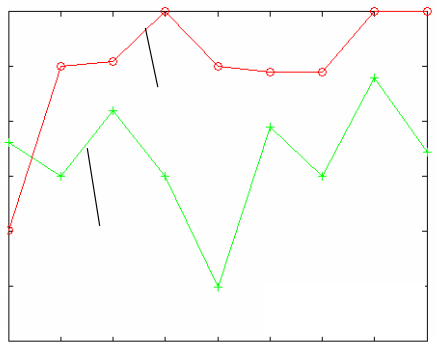

图 1 贝叶斯分类器分类效果对比

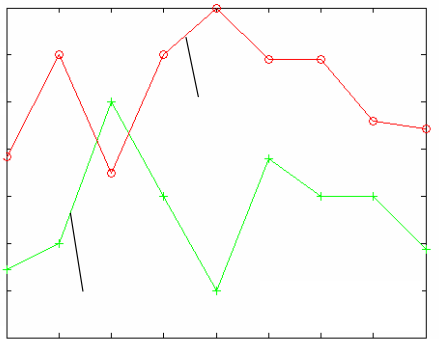

图 2　最小距离分类器分类效果对比

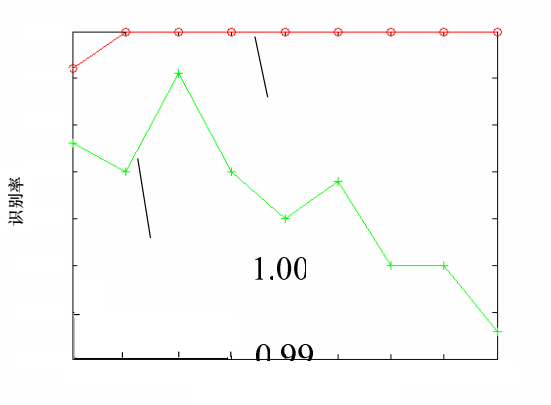

图 3　近邻分类器分类效果对比

从图 1~图 3 可以看出，由于 Box-Cox 变换，IRIS 数据每一类的分布都更加接近于正态化，每一类都更加集中在自己的一个较小的区域内，交叉区域相对减少，并最终导致数据间的类别更加清晰。反映在分类结果上就是分类正确率有了明显的提高，分类错误率将随之降低。另外，通过本文的实验过程可以看出，实验是在先验概率已知的情况下，选择性抽样进行训练测试。所以，得到的正确识别率估计是正确识别率的无偏估计，是在最似然估计意义下的最好估计，进而证明了本文实验结果的可靠性和正确性。

关于 Box-Cox 变换应用于 BP 网络的实验，给出其中一次的分类效果对比图如图 4、图 5 所示。

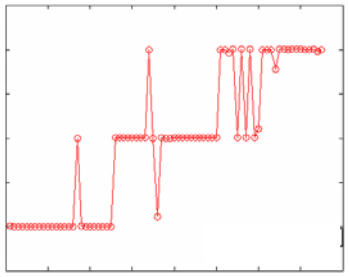

图 4　Box-Cox 变换前分类效果

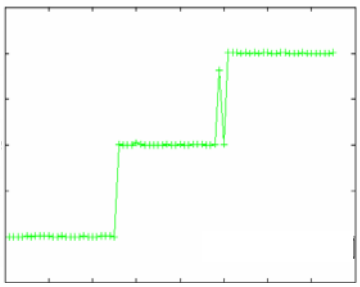

图 5　Box-Cox 变换后分类效果

为了使实验的结果更具有说服力，经过反复多次的实验。每一轮连续做 50 次，连续做 5 轮；实验中不仅统计出输出值与目标输出的误差，还统计出这 5 轮每轮 50 次实验的耗时，如表 1 所示(用 BC 表示 Box-Cox 变换)。

表 1　Box-Cox 变换前后误差、耗时对比　　s

轮次	误差		耗时	
	BC 前	BC 后	BC 前	BC 后
1	0.152 20	0.006 70	58.235 00	40.625 00
2	0.096 70	0.007 90	54.422 00	46.859 00
3	0.094 00	0.005 20	57.547 00	49.125 00
4	0.127 70	0.009 90	57.563 00	49.156 00
5	0.083 30	0.006 40	60.156 00	45.172 00
均值	0.110 78	0.007 22	57.584 60	46.187 20

通过表 1 可以看出，Box-Cox 变换以后，输出值与目标输出更加接近，分类的准确性有很大的提高。

另外，不难看出在如上类型判别准确率提高、误差降低的同时，由于训练时连接权值调整的加快和整个神经网络收敛速度的加快，使得每 50 次实验的总耗时减少，从而提高了神经网络的速度。

4　结束语

本文把 Box-Cox 变换用于各种分类器的性能改进，推导出了基于 Box-Cox 变换的各种分类器表达形式，并且在 IRIS 数据集上的对于各个分类器性能的实验。通过大量实验可以得到如下结论：各种分类器的识别率普遍得到提高，类型判别普遍稳定在相当高的正确率；特别使得神经网络分类器输出值更加准确化，输出值与目标输出的误差降低到一个很低的水平。通过 Box-Cox 变换使得 BP 神经网络的训练速度有了很大的提高，从而在提高识别率的同时，加快了 BP 网络类型判别的速度。

参考文献

[1] 边肇祺, 张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社, 2000.

[2] Myles J P, Hand D J. The Muti-class Metric Problem in Nearest Neighbour Discrimination Rules[J]. Pattern Recognition, 1990, 23(11): 1291-1297.

[3] Heiden R V D, Groen F C A. Box-Cox Metric for Nearest Neighbour Classification Improvement[J]. Pattern Recognition, 1997, 30(2): 273-279.

[4] Ujiie H, Omachi S, Aso H. A Discriminant Function Considering Normality Improvement of the Distribution[C]//Proceedings of the 16th International Conference on Pattern Recognition. Quebec, Canada: [s. n.], 2002: 224-227.

[5] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces Fisherfaces: Recognition Using Class Specific Linear Projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.

编辑　索书志

（上接第 165 页）

[7] NFC Forum. NFC Data Exchange Format(NDEF)[J/OL]. (2006-07-24). http://www.nfc-forum.org/specs/.

[8] NFC Forum. NFC Record Type Definition(RTD)[J/OL]. (2006-07-24). http://www.nfc-forum.org/specs/.

[9] 许海翔, 伏京生. 近场通信技术促进智能卡应用的前景展望[J]. 金卡工程, 2008, 12(2): 23-25.

编辑　张　帆