第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)05—0188—03 文献标识码：A 中图分类号：TP181

基于核的最小距离分类法的参数选择方法

邱潇钰，张化祥
（山东师范大学信息科学与工程学院，济南 250014）

摘 要：在基于核函数的最小距离分类方法对数据集进行分类过程中，目标函数的核函数参数选择直接影响分类器的分类成功率。该文提出一种选择应用目标函数来选择适当参数的方法。实验结果表明，与单纯的基于核的最小距离分类法相比，选择最优核函数的参数可以提高分类器的成功率。

关键词：最小距离分类法；数据集；核函数

Parameter Selection Method Based on Kernel Nearest Neighbor Classification

QIU Xiao-yu, ZHANG Hua-xiang
（School of Information Science and Engineering, Shandong Normal University, Jinan 250014）

【Abstract】This paper proposes kernel nearest neighbor to classify the data sets. The parameter of the kernel function is the most difficulty question of this research topic, so this paper proposes a hybrid approach to use the target function to choose an adaptive parameter. Through testing the approach on a typical classification data sets, and the preliminary results demonstrate that, the target function can provide an adaptive parameter to optimize the kernel function for classification in various domains, especially compared with other kernel-based nearest neighbor classification methods.

【Key words】 nearest neighbor classification; data sets; kernel function

1 概述

在模式识别中，最小距离分类器(NN)是非常简单并且很常用的技术。目前这种方法演变为基于核函数的最小距离分类器(KNN)以应对复杂的非线性的方式分类。像其他基于核的方法一样，KNN也是受核函数的参数的影响。本文研究使用目标函数选择最优的核函数参数来提高的KNN的分类效率。在众多核函数参数的选择方法中，最经典的参数的选择方法有 2 种[1-3]：(1)选择一系列的核参数，并且反复地在算法中运行。从而选择一个执行效果比较好的参数来作为选择的参数。因为参数范围的原因，所以这种方法找不到最优点。(2)比较著名的“cross-validation”，即把函数参数的范围放到更大的序列中去，但是这种方法的时间消耗太大，不适用于实际情况。文献[4]提出的设计新的目标函数来对KNN的参数选择，取得了很好的效果，但是计算的复杂度太大，不适用于大量的数据。本文研究使用目标函数求取核函数参数值的问题，计算代价比张道强提出的目标函数要小，计算效果与它持平。通过使用适当目标函数进行一种预测，使得非线性的数据集通过一个非线性函数 $\phi(x)$，映射到一个更高位的特征空间，然后使用最小距离分类法对新的特征空间中的点进行分类。该方法的特点是使用目标函数选择合适的参数，达到最优化分类的目的。

2 最小核距离的分类方法

模式识别中，经常被给予 l 个训练数据，$\{(x_i,y_i)\}_{i=1}^{l}$，其中，$x_i \in R^q$，并且 $y_i \in (0,1)$。为了对新的数据项 x_i 进行分类，将采用最近距离分类法，即利用训练数据得到每种类别的特征参数 Φ_i。以此为类别标准，分类时计算出每个像元的特征参数 Φ_i 与类别标准两者在特征空间里的距离，比较它到各种类别特征参数 Φ_i 的距离，将该像元划并到距离最小的类别。有多种方法计算空间的距离，常用的有绝对距离 $D^2(y,\Phi_i)=\|y-\Phi_i\|^2$ 和欧氏距离 $D_\varepsilon(y,\Phi_i)=\|y-\Phi_i\|$。

在适当的特征空间中，核函数用人们熟知的线性分类算法对非线性关系进行分类。尽管通过非线性函数将样本数据映射到具有高维甚至为无穷位的特征空间，并在其中构造最优分类超平面，但在具体求解时并不需要先行计算该线性函数 $\phi(x)$，而只需计算核函数，从而避免特征空间维数的灾难问题[5]。核函数的选择必须满足Merce条件[6]。或者可以认为：如果 $\phi(x),\phi(x')$是输入空间的一个点 x,x'的映射，那么核函数通过计算 $\phi(x)\times\phi(x')$，从而避免了计算新特征空间中的点 $\phi(x),\phi(x')$。即

$$k(x,x')=\phi(x)\times\phi(x') \tag{1}$$

常用的核函数有：

(1)多项式分布

$$k(x,x')=(x^{\mathrm{T}}x'+1)^d$$

(2)高斯分布

$$k(x,x')=\exp\left[-\frac{\|x-x'\|^2}{2\sigma^2}\right]$$

基金项目：山东省科技攻关计划基金资助项目(2005GG4210002)；山东省青年科学家科研奖励基金资助项目(2006BS01020)

作者简介：邱潇钰(1982－)，男，硕士研究生，主研方向：数据挖掘，机器学习；张化祥，教授、博士

收稿日期：2007-03-30 **E-mail：**xiaoyu402@163.com

(3)双曲 tan*h* 函数：

$$k(x,x')=\tanh\left(\lambda\left\|x-x'\right\|\right)$$

在新的特征空间中，最小分类法的数学表达方式：

$$h(x)=\arg\min_{1\le j\le C} D^2\left(\phi(x),\phi(\Phi_i)\right) \tag{2}$$

其中，$h(x)$是在新的特征空间中的分类的标准，即基于核函数的最小距离分类法。

对于最小距离分类问题，在输入特征空间中，每个类的中心为 $\Phi_i(i=1,2,\cdots,C)$，则在新的特征空间中的点 $\phi(x)$，到 i 类中心 $\phi(\Phi_i)$ 的距离定义为

$$D^2\left(\phi(x),\phi(\Phi_i)\right)=\left\|\phi(x)-\phi(\Phi_i)\right\|^2=\phi(x)\phi^{\mathrm{T}}(x)-2\phi(x)\phi(\Phi_i)+\phi(\Phi_i)\phi^{\mathrm{T}}(\Phi_i)=2-2\exp\left[-\frac{\left\|x-\Phi_i\right\|^2}{2\sigma^2}\right] \tag{3}$$

其中，$\Phi_i=\frac{1}{n_i}\sum_{i=1}^{n_i}\Phi(x_i)$，$n_i$ 为 i 类中的训练数据的样本数。

对式(3)分析中，可以发现在核函数中，参数 σ 的选择直接影响到分类器的性能。由此可知目标函数的选择是一个关键的设计问题

3 目标函数

目标函数的作用是为 KNN 选择最优的核参数，从而使得 KNN 的分类成功率达到最优。通常目标函数的使用是一个权衡的过程：一方面希望在更高维的空间中，各个类之间的距离达到最大；另一方面又希望带权重的各数据到各自中心的距离达到最小。

$$f(\sigma)=\sum_{i=1}^{C}\left[\frac{n_i}{n}\sum_{j=1}^{n_i}D^2\left(\phi(x_j),\phi(\Phi_i)\right)\right]-\frac{1}{2}\left[\sum_{i=1}^{C}\sum_{j=1}^{C}D^2\left(\phi(\Phi_i),\phi(\Phi_j)\right)\right]\quad(c\neq 2) \tag{4}$$

其中，c是数据集中类的个数；n_i 为 i 类中的训练数据的样本数；$\phi(x_j)$ 为 x_j 在新特征空间中的映射；$\phi(\Phi_i)$ $\phi(\Phi_j)$ 为 Φ_i 在特征空间的映射。其中式(4)右边第 1 项表示特征空间中，带权重的各数据到各自中心的距离之和。为了方便比较，计算各个点到各个类中心的平均距离的加和。用n_i/n来控制第 1 项中的加和各数。第 2 项表示特征空间中类间的距离之和。控制第 2 项加和的个数与第 1 项加和的个数相同，这样可以使目标函数的作用更加明显。那么存在一个特例：当 $c=2$ 时，第 1 项是两个项的加和，而第 2 项是一个类的加和。

将式(2)、式(3)带入式(4)中得到

$$f(\sigma)=\sum_{i=1}^{C}\left[\frac{n_i}{n}\sum_{j=1}^{n_i}\left(2-2\exp\left[-\frac{\left\|x_j-\Phi_i\right\|^2}{2\sigma^2}\right]\right)\right]-\frac{1}{2}\left[\sum_{i=1}^{C}\sum_{j=1}^{C}\left(2-2\exp\left[-\frac{\left\|\Phi_i-\Phi_j\right\|^2}{2\sigma^2}\right]\right)\right] \tag{5}$$

目标函数 $f(\sigma)$的最小化，即使得第 1 项类内距离最小化，第 2 项类间距离最大化，以 $f(\sigma)$作为目标函数学习参数值。将选择训练数据集对 $f(\sigma)$进行训练，在 σ 等于适当的值的时候，$f(\sigma)$最小化。

4 实验

整个实验过程将分 2 步来完成：(1)使用训练集对目标函数式(5)进行训练，从而得出适当的参数 σ 值。(2)使用最小距离分类器结合参数 σ 对矩阵中的点进行分类。

4.1 数据集

选用的数据集来自 the UCI Machine Learning Database Repository，分别是：

(1)Iris data：数据个数 L=150，属性个数 Q=4，类的个数 J=3。

(2)Wine data：数据个数 L=137，属性个数 Q=13，类的个数 J=3。

(3)Bupa data：数据个数 L=292，属性个数 Q=6，类的个数 J=2。

(4)Cmc data：数据个数 L=120，属性个数 Q=9，类的个数 J=3。

4.2 训练集和测试集

训练集和测试集如表 1 所示。

表 1 训练集和测试集

	训练集	测试集
Iris data	105	45
Bupa data	152	140
Wine data	90	45
Cmc data	60	60

在具体的求解过程中，构建一个 Gram 矩阵，它包含的是所有配对数据点上对核函数的求值结果。这个矩阵具有信息瓶颈的作用，因为它能提供给核算法所有的信息。

4.3 目标函数对核参数的选择

目标函数对各个训练数据集的参数预测图像，如图 1~图 4 所示。图中的横坐标代表 $\sigma(1:0.1)$，纵坐标代表目标函数 $f(\sigma)$的值。

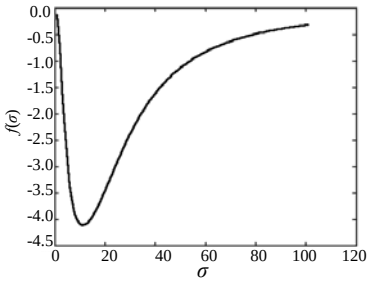

图 1 目标函数对 Iris data 的参数预测图像

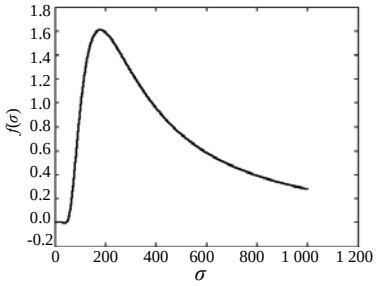

图 2 目标函数对 Bupa data 的参数预测图像

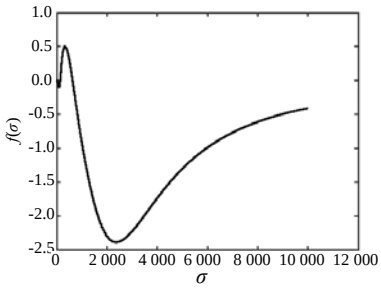

图 3 目标函数对 Wine data 的参数预测图像

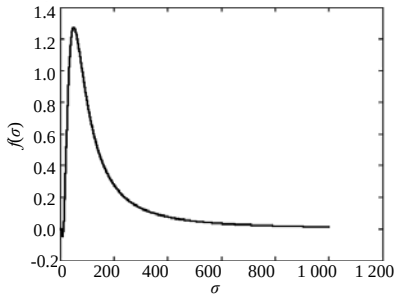

图 4　目标函数对 Cmc data 的参数预测图像

把图像中的 x 轴的单位 1 定义为实际数值 0.1。在图像中明显地看出在 x 值相对较小的时候存在一个最小值，然后当 $x(0\to+\infty)$ 时，$f(\sigma)$无限趋近于最小值。所以得到了一个这样的假设：当 x 值足够大时，它与最优化的点的分类效果是一样的。为了验证在图像中得出的假设，将在试验中加以证实。

因为各个数据集的自身特点，所以选择的参数有很大不同。3 个类数据集的最低点比较靠后，2 个类数据集的最低点比较靠前。这是因为目标函数的第 1 项实际上是各个点到各自中心的平均距离，3 个类就是 3 个平均距离的加和，2 个类是 2 个平均距离的加和，所以数值相对较小的比较大的更依赖参数值的变化。

由此可见，对于 Iris data，选择 $\sigma=1.1$ ；对于 Bupa data，选择 $\sigma=4.1$ ；对于 Wine data，选择 $\sigma=237.8$ ；对于 Cmc data，选择 $\sigma=0.9$。

4.4　实验结果及分析

为了验证使用目标函数后，基于核函数的最小分类法的效果。将不同 σ 值的核函数的最小距离分类法，最小距离分类法进行对比，如表 2 所示。

表 2　实验结果

	KNN/(%)	σ	最小距离分类法/(%)
Iris data	95.56	1.1	95.56
	84.44	0.1	
	95.56	500	
	95.56	800	
	95.56	1 100	
Bupa data	56.43	4.1	55.71
	49.30	0.5	
	57.14	600	
	57.14	900	
	57.14	1 200	
Wine data	64.44	237.8	59.44
	33.33	0.7	
	64.44	800	
	64.44	1 100	
	64.44	1 400	
Cmc data	33.33	0.9	21.71
	33.33	0.01	
	28.33	500	
	28.33	800	
	28.33	1 200	

从表 2 中容易看出：

(1)在 σ 取不同值比较中，很容易看到当 $\sigma(0\to+\infty)$ 时，基于核函数的最小分类法的分配正确率趋近于最优点。并且到当值比较大时，它的分类效果与在最优点是相近的。这证明了我们在目标函数图像中的假设，也证明目标函数的准确性。

(2)对比最小距离分类法，基于核函数的最小分类法对于某些数据集时分类效果要好一些。分析其中的原因：目标函数为 KNN 选择了最优的参数，所以，使得分类数据要比最小距离分类法效果要好。

(3)在新的特征空间中，对于训练集中含有较多实例的问题类型，由于问题类型的特征信息比较集中，各个类之间的绝对距离相对较远，因此属于这些类型的问题分类相对比较准确，比如对 Iris data 的问题分类准确率比较高，达到了 95.56%。而对于训练集中含有相对分散的问题类型，目标函数得不到充分的问题类型的特征信息，所以可能会造成这一类的问题分类准确率过于低下，比如属于 Bupa data 和 Cmc data 的问题分类准确率较低，只有 56.43%和 33.33%。

5　结束语

本文采用了使用基于核函数的最小距离分类器来进行分类，并目标函数来预测核函数的参数，取得了良好的效果。不过此问题还有很多需要改进的地方，如果能进一步地选择合适的目标函数，并结合其他优秀的分类方法，还可以进一步地优化分类效果使分类器呈现更好的性能。

参考文献

[1] Peng Jing, Heisterkamp D R. Adaptive Quasiconformal Kernel Nearest Neighbor Classification[J]. IEEE Trans. on Pattern Anal. Mach. Intell., 2004, 26(5): 656-661.

[2] Zhang Daoqing, Chen Songcan. Clustering Incomplete Data Using Kernelbased Fuzzy C-means Algorithm[J]. Neural Process. Lett., 2003, 18(3): 155-162.

[3] Wang Lei, Chan Kap Luk. Learning Kernel Parameters by Using Classseparability Measure[C]//Proc. of NIPS'02 Workshop on Kernel Machines. Whistier, Canada: [s. n.], 2002.

[4] Zhang Daoqiang, Chen Songcan, Zhou Zhihua. Learning the Kernel Parameters in Kernel Minimum Distance Classifier[J]. Pattern Recognition, 2006, 39(1): 133-135.

[5] Boser B, Guyon I, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[C]//Proc. of the 5th Annual ACM Workshop on Computational Learning Theory. New York: ACM Press, 1992: 144-152.

[6] Vapnik V N. The Nature of Statistical Learning[M]. Berlin: Springer, 1995.

（上接第 171 页）

参考文献

[1] Tumer D, Entwisle S. Symantec Internet Security Thread Report Trends[R]. Symantec Inc., 2006.

[2] Szor P, Ferrie P. Hunting for Metamorphic[C]//Proc. of the 11th International Virus Bulletin Conference. Prague, Czech Republic: [s. n.], 2001.

[3] Walenstein A, Mathur R. Normalizing Metamorphic Malware Using Term Rewriting[C]//Proc. of the 6th IEEE Workshop on Source Code Analysis and Manipulation. Philadelphia, PA, USA: [s. n.], 2006.

[4] Christodorescu M, Jha S. Semantics-aware Malware Detection[C]// Proc. of IEEE Symposium on Security and Privacy. Oakland, California, USA: [s. n.], 2005.

[5] Finjan Inc. Behavior-based Security[Z]. 2006.