基于 SRM 自组织多区域覆盖的可拒绝近邻分类算法研究

胡正平　贾千文
(燕山大学信息科学与工程学院　秦皇岛　066004)

摘　要：该文依据区分与划分相结合的可拒绝模式识别思路，提出了高维空间海量训练样本情况下的基于结构风险最小化决策的自组织多区域多球覆盖可拒绝近邻分类算法。该方法利用同类样本之间相互接近的特性，通过结构风险最小化紧几何覆盖策略，选择训练样本，通过自组织多区域多球覆盖模型构成同类样本的划分性描述，达到拒绝识别非训练类样本的目的，最后通过k近邻相互区分性比较确定真实类别。仿真实验结果表明该文的思路是合理可行的，在实际应用领域具有一定价值。

关键词：可拒绝模式识别；结构风险最小化原理；近邻分类
中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1009-5896(2009)02-0293-04

Rejecting Nearest Neighbor Classifier Based on Structural Risk Minimization Principle Self-organization Multiple Region Covering Model

Hu Zheng-ping　　Jia Qian-wen
(*School of Information Science and Engineering*, *Yanshan University*, *Qinhuangdao* 066004, *China*)

Abstract: According to the mode of rejecting pattern recognition principles, which is based on "matter description" and "matter separation" in uniform statistical pattern recognition, a rejecting nearest neighbor classifier based on structural risk minimization self-organization multiple region covering model is presented in this paper. This new model is better closer to the actual instance, rather than traditional statistical pattern recognition only using "optimal separation"as its main principle. Firstly, the optimal samples are selected from the training samples based on structural risk minimization, which is used for same class pattern matter description. Then the kNN distinguish is as a following step to identified the exact class. The simulation experimental results show that this method is valid and efficient.

Key words: Rejecting pattern recognition; Structural risk minimization; Nearest neighbor classifier

1　引言

数十年来人们在研究模式识别问题中，考虑的出发点都是在若干类别的最佳分类划分上，其根本原因或许在于用这样的数学描述与处理方法最具有一般性、通用性。但众所周知，即使基于目前最先进的模式识别理论基础上的识别机，其实际效果却仍然远不能令人满意。为解决此问题，中国学者王守觉教授提出了仿生神经网络，其基本出发点就在于把模式识别问题看成模式的“认识”，而不是分类划分，不是模式分类；是一类一类样本的“认识”，而不是多类样本的划分。他们建立与传统模式识别的“最优分类”界面的概念有所区别的基本数学模型，并构造了仿生神经网络[1]，核覆盖网络[2]等模型，并在实际应用中取得较好的效果。目前广受关注的支持向量机分类器因为其优越的性能成为近年研究的热点，SVM算法最为关键环节是基于结构风险的最大间隔划分，但是经典的SVM分类器总是假定测试样本属于两类训练样本之一，然而在很多分类识别问题中，往往存在许多非训练类目标样本，这时经典SVM分类器往往会给出错误的判决。为了解决可拒绝模式识别问题，需要将目前这两个最新的模式识别思想结合起来：认识事物有利于拒绝非训练类样本，最大间隔划分有利于区分真实类别。

最近邻分类是基于样本间距离的一种分类方法，最初的近邻法是由Cover和Hart于1968年提出的，由于该方法在理论上进行了深入的分析，方法简单有效且易于实现，直至现在仍是模式识别中重要的非参数方法之一[3]。近邻分类可以从高维空间几何学的角度，将判别过程理解为：以样本空间中每一类别的所有样本点为中心，以阈值为半径做超球体。判断待测试样本落入哪些超球体内，再计算待测试样本与这几个超球体球心的距离。最终将测试样本判别为与它最近的超球体球心同类。正是基于近邻分类优点，本文提出了基于结构风险最小化的自组织多区域多球覆盖可拒绝近邻分类算法，一方面利用同类样本之间相互接近的特性，通过结构风险最小化策略，选择训练样本，通过自组织多区域多球覆

2007-08-27 收到，2008-05-11 改回
河北省自然科学基金(F2008000891)，燕山大学博士基金项目(B287)和北京大学视觉与听觉信息处理国家重点实验室开放基金（0507）资助课题

盖构成同类样本的紧致性覆盖模型，进而达到拒绝识别非训练类样本的目的。最后再通过k近邻相互区分性比较达到识别真实类别的目的。

针对如何构造拒识-判决分类器问题，许多学者提出了一些有效的解决问题的思路。文献[4−6]提出了支持向量域描述(Support Vector Domain Description, SVDD)主要用来进行数据描述和剔除奇异点，SVDD的基本思想就是计算包含一组数据的最小超球体边界来对数据的分布范围进行描述。SVDD的不足之处在于对于非规则形状分布的覆盖模型不够紧密，存在不少冗余区域。文献[7]提出基于支持向量表示-鉴别机(Support Vector Representation and Discrimination Machine SVRDM)的具有拒识能力的分类方法，该方法的优点是一步实现可拒识分类器，缺点是控制拒绝能力的门限参数选择比较复杂，不具有自适应能力。文献[8]提出另外一种类似于本文的两步实现策略，首先将两类归于一大类，第1步决定是否拒绝，第2步进行分类。 该方法缺点是第一步需要知道全部非目标样本并进行训练，一方面计算量巨大，另一方面在实际应用中往往不容易收集到全部的非目标样本。而本文提出的方法不需要知道全部非目标样本，甚至只需要知道部分两类目标样本本身。文献[9]提出了基于模板等预处理结合SVM的分类方法，预处理的过程就是数据粗选择过程，该方法对于可模板化描述的问题是合理有效的，缺陷是对于不能模板化的高维数据无能为力，同时自适应门限选择困难。

2 可拒绝近邻分类系统组成

本文构造的可拒绝近邻分类模型如图 1 所示，该系统主要由两个模型组成：基于结构风险最小的近邻覆盖模型(认识层)和基于间隔划分的近邻区分模型，各自模型的参数通过样本训练而得到。

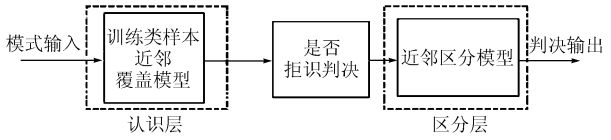

图 1 可拒绝近邻分类模型系统组成原理框图

3 可拒绝近邻分类模型

从上面给出的可拒绝近邻分类模型可以看出：基于结构风险最小的近邻覆盖模型(认识层)是以样本在特征空间的分布的最佳覆盖作为目标；而区分层则把不同类样本在特征空间中的最佳划分作为目标，下面利用近邻策略分别构建认识层与区分层的数学模型。

3.1 认识层模型

传统最近邻认识层规则规则如下：假定有 c 个类别：$\omega_1,\omega_2,\cdots,\omega_c$ 的模式识别问题，每类由标明类别的样本 $N_i(i=1,2,\cdots,c)$ 个，规定 ω_i 类的判别函数为

$$g_i(\boldsymbol{x})=\mathop{\mathrm{Min}}_{k}\left\|\boldsymbol{x}-\boldsymbol{x}_i^k\right\|,\quad k=1,2,\cdots,N_i \tag{1}$$

其中 $\boldsymbol{x}_i^k$ 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。

近邻不设门限时，决策规则可以写为，若

$$g_j(\boldsymbol{x})=\min_i g_i(\boldsymbol{x}),\quad i=1,2,\cdots,c \tag{2}$$

则 $\boldsymbol{x}\in\omega_j$。即对未知样本 $\boldsymbol{x}$，只要比较 $\boldsymbol{x}$ 与所有已知类别之间的欧式距离，并判决 $\boldsymbol{x}$ 与离它最近的样本同类，该思路对于非训练类样本缺乏拒绝判决的能力。

近邻设门限 th 时，决策规则可以写为，若同时满足下面式(3)，式(4)：

$$g_j(\boldsymbol{x})=\min_i g_i(\boldsymbol{x}),\quad i=1,2,\cdots,c \tag{3}$$

$$g_j(\boldsymbol{x})\le \mathrm{th} \tag{4}$$

则认为 $\boldsymbol{x}\in\omega_j$，否则测试样本属于非训练类样本。

上面的传统近邻认识层模型，主要存在两方面的缺陷：一是门限设置缺乏自适应性，二是样本数量大，导致计算效率比较低。鉴于此，提出自组织多区域多球思路，可以构造更好的认识层覆盖模型。示意图如图 2 所示，覆盖(b)属于更加合理的覆盖模型。

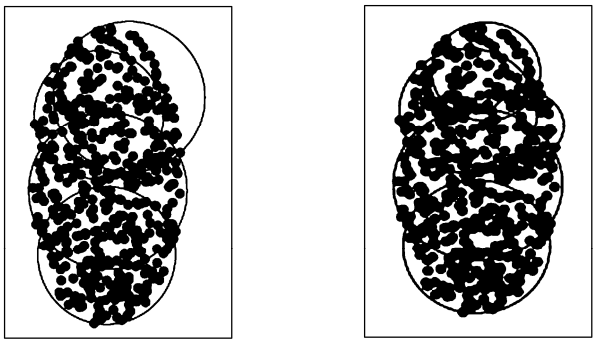

图 2 覆盖模型示意图

同类样本之间存在具有相互接近的性质，先假设同类样本的分布可以包括多个聚集区域(仿生模式识别认为同类样本只属于一个区域)，进而可以假设同一聚集区域的两个同类样本点之间的无数条曲线中，至少存在一条这样的曲线，满足曲线上每一点都与这两个样本点同类，公式表示：

$$L^k=\{c_i\in A\mid c_i\in L^{(k)}\},\quad i=1,2,\cdots,\infty \tag{5}$$

这里 c_i 为曲线 $L^{(k)}$ 上某一点，A 为 A 类模式，$L^{(k)}$ 为同类样本的一个区域。当两个样本点在样本空间位置接近时，可以用两个点的连线来逼近满足条件的曲线，根据连续性原理可知，曲线离散后的每一个点仍然数据这类模式，最后以离散的点为球心，以参数 r 为半径的作超球体，这样若干个离散的超球体就形成了对于区域 $L^{(k)}$ 的一个覆盖。

超球体描述为

$$p_i^{(k)} = \{(\boldsymbol{x} - \boldsymbol{a}_i^{(k)})^2 \le r_i^{(k)} \mid \boldsymbol{x} \in R^n, \quad i = 1,2,\cdots,N \tag{6}$$

将所有区域的超球体的并用于覆盖 A 类模式的分布区域，则 A 类模式的覆盖 $P_{(A)}$ 为

$$P_{(A)} = \bigcup_{k=1}^{M} \bigcup_{i=1}^{N} p_i^{(k)} \tag{7}$$

这里借鉴了仿生模式识别的原理，但是与仿生模式识别的超球覆盖模型相比，本文的模型主要有两方面的区别：

(1)超球体的半径是变化的，与区域分布形状具有自适应性。

(2)认为同类事物可以有多个区域覆盖模型，而不是一个区域分布，这样的模型具有更加广泛的适用性。

为保证对于非训练类样本的准确拒绝，需要构造训练类样本的结构风险最小的紧致性覆盖模型：既要保证无冗余的覆盖，又要保证模型的复杂度尽可能的低(置信度高)。为此，本文构造的算法步骤如下：

(1)采用模式聚类方法将训练类样本无监督聚类，聚类准则为

$$E = \sum_{n=1}^{N} \sum_{i=1}^{K_n} (\boldsymbol{x}_i^n - \boldsymbol{m}^n)^{\mathrm{T}} (\boldsymbol{x}_i^n - \boldsymbol{m}^n) \tag{8}$$

这里 N 为聚类组数，K_n 为类别为 n 的样本数目，$\boldsymbol{m}^n$ 为类别 n 的中心，$\boldsymbol{x}_i^n$ 为类别为 n 的训练样本。

(2)对于上面聚类得到每一组样本，构造最小半径的超球进行覆盖。

(3)对于上面得到的覆盖模型进行冗余度检验，冗余度定义为

$$R = r^M / K_n \tag{9}$$

M 为特征空间维数，K_n 为超球包含的样本数目。当 R 很大时，说明覆盖冗余度大，需要增加聚类类别数。重复步骤(1)，步骤(2)，步骤(3)直到达到规定的要求。

3.2 区分层模型

目前研究表明：近邻距离分类器的性能主要集中在最近邻类与次近邻类上[10,11]，因此本文利用近邻综合决策构造近邻区分模型，基本思路如下：

首先计算输入模式 $\boldsymbol{x}$ 的最近邻类别 w_i 以及最近邻距离 d_i $(i = 1,2,\cdots,c)$，然后引入广义置信度模型：

$$y(w_i / \boldsymbol{x}) = 1 - \frac{d_i}{\min\limits_{j \ne i} d_j}, \quad i = 1,2,\cdots,c \tag{10}$$

然后建立广义置信度与对应判决的自适应变换函数，自适应变换函数根据训练样本构造的神经网络模型训练而得到。

进而后验概率可以通过下式估计而得到：

$$\hat{p}(w_i / \boldsymbol{x}) = f(y(w_i / \boldsymbol{x})), \quad i = 1,2,\cdots,c \tag{11}$$

这里函数 $f(\cdot)$ 由神经网络训练得到，为自适应变换函数。

一般地，当广义置信度 $y(w_i / \boldsymbol{x})$ 取值较大时，可以认为样本属于 w_i 的后验概率也比较大，反之，后验概率也比较小。

4　仿真实验

为了验证本文提出算法的有效性和合理性，该文利用 MATLAB7.0 进行了实验仿真。为了验证本文提出的算法的性能，本文进行了两组实验，一组对随机产生的满足高斯分布的三类数据(一类作为非目标样本)进行实验，第二组对手写数字(一组代表非目标样本)进行分类实验。

(1)高斯分布样本点分类实验　随机产生了 3 类 3 维样本的数据点。每类样本数目 800 点，满足多元正态分布，样本点分布参数如表 1 所示。

表 1　合成样本点参数分布

类别	均值 1	均值 2	均值 3	标准偏差
1	0.2	1.0	2.5	1
2	1.0	1.8	0.5	1
3(非目标)	0.6	1.4	1.6	1

实验时，每一类别中选择 1000 点作为训练样本，1000 点作为测试样本。对比实验结果如表 2 所示。

表 2　对比实验结果

方法	正确分类率(%)	正确拒识率(%)
近邻门限	80.1	88.1
本文方法	80.0	89.6

(2)手写数字分类实验　实验数据来源于 MNIST 手写数字数据库，仿真实验以 “3”类、“5”类作为已知类，对随机抽取的 1000 个数字(包括非训练类数字 “8”)进行识别。实验时，随机抽取属于不同类的 1000 个数字，统计如表 3 所示，识别结果见表 4 所示。

表 3　待识别数字统计

待识别数字	已知类别		未知类
	3	5	8
数目	320	380	300

表 4 不同方法识别结果统计

待识别数字		随机抽取数	正确识别数	错误识别数	拒识数
近邻门限法	“3”类	320	280	10	30
	“5”类	380	370	0	10
	“8”类	300	—	10	290
本文方法	“3”类	320	292	8	20
	“5”类	380	373	0	7
	“8”类	300	—	6	294

将两种不同方法对已知类、非训练类的正确识别率、错误识别率以及拒识率总结于表 5 中。

表 5 已知类数字识别结果

方法	待识别数字	正确识别率%	错误识别率%	拒识率 %
近邻门限法	已知类	92.86	1.43	5.71
	非训练类	—	3.33	96.67
本文方法	已知类	95	1	2.5
	非训练类	—	1.8	98.2

从表 5 中的数据可以看出：本文方法在对已知类识别时，可以达到 95%的正确识别率，而对未知类识别时，可以达到 98.2%的正确拒绝率。这说明本文提出的方法是有效可行的。

5 结论

依据划分与认识相结合的思路，本文研究了基于近邻描述的可拒绝模式识别模型。针对认识层模型，构造了结构风险最小化的自组织多区域多球覆盖模型；针对区分层，利用近邻综合决策构造了基于置信变换的后验估计区分模型。最后的实验结果表明该方法可以对非训练类例外模式进行有效拒绝，并且对已知类的正确识别产生的影响比较小。虽然可拒绝模式识别具有广泛的应用背景，但是目前的工作还很不够。如何将区分层与认识层很好的融合起来，如何构造高维稀疏空间的覆盖模型等都是值得进一步研究的问题。

参 考 文 献

[1] 王守觉. 仿生模式识别——一种模式识别的新模型的理论与方法[J]. 电子学报, 2002, 30(10): 1418–1421.
Wang Shou-jue. Bionic (topological) pattern recognition——A new model of pattern recognition theory and its applications. *Acta Electronica Sinica*, 2002, 30(10): 1417–1420.

[2] 吴涛, 张铃, 张燕平. 机器学习中的核覆盖算法[J]. 计算机学报, 2005, 28(8): 1295–1301.
Wu Tao, Zhang Ling, and Zhang Yan-ping. Kernel covering algorithm for machine learning. *Chinese Journal of Computers*, 2005, 28(8): 1295–1301.

[3] 赵莹, 高隽, 汪荣贵, 胡静. 一种新的广义近邻方法研究[J]. 电子学报, 2004, 32(F12): 196–198.
Zhao Ying, Gao Jun, Wang Rong-gui, and Hu Jing. Extended nearest neighbor method based on bionic pattern recognition. *Acta Electronica Sinica*, 2004, 32(F12): 196–198.

[4] David M J Tax and Robert P W Duin. Support vector data description [J]. *Machine Learning*, 2004, 54(1): 45–66.

[5] Lee Ki Young, Kim Dae-Won, Lee Kwang H, and Lee Doheon. Density-induced support vector data description. *IEEE Trans. on Neural Networks*, 2007, 18(1): 284–289.

[6] Yen Chen-wen, Young Chieh-Neng, and Nagurka Mark L. A false acceptance error controlling method for hyperspherical classifiers. *Neurocomputing*, 2004, 57(1-4): 295–312.

[7] Amit B, Philippe B, and Chris D. A support vector method for anomaly detection in hyperspectral imagery. *IEEE Trans. on Geoscience and Remote Sensing*, 2006, 44(8): 2282–2291.

[8] Rizvi S A, Saasawi T N, and Nasrabadi N M. A clutter rejection technique for FLIR imagery using region based principal component analysis [J]. *Pattern Recognition*, 2000, 33(11): 1931–1933.

[9] Yang G Z and Huang T S. Human face detection in a complex background [J]. *Pattern Recognition*, 1994, 27(1): 58–63.

[10] Cabrera J B D. On the impact of fusion strategies on classification errors for large ensembles of classifiers. *Pattern Recognition*, 2006, 39(11): 1963–1978.

[11] 娄震, 金忠, 杨静宇. 基于类条件置信变换的后验概率估计方法[J]. 计算机学报, 2005, 28(1): 18–24.
Lou Zhen, Jin Zhong, and Yang Jing-yu. Novel approach to estimate posterior probabilities by class-conditional confidence transformations. *Chinese Journal of Computers*, 2005, 28(1): 18–24.

胡正平： 男，1970 年生，博士，副教授，研究方向为统计学习理论与模式识别、医学图像处理.

贾千文： 女，1984 年生，硕士生，研究方向为统计学习理论与模式识别.