文章编号: 1001-0920(2009)04-0547-04

基于局部均值与类均值的近邻分类

曾 勇, 杨煜普, 赵 亮
(上海交通大学 自动化系, 上海 200240)

摘 要: k-近邻分类是一种流行且成功的非参数分类方法,但其分类性能由于离群点的存在而受到损害. 为克服离群点对分类性能的不利影响,提出了一个 k-近邻分类的变形和一个基于局部均值向量与类均值向量的近邻分类方法. 该方法利用了未分类样本在每个训练类中 k 个近邻的局部均值的信息和整体均值的知识,不仅能够克服离群点对分类性能的影响,而且取得了比传统的 k-近邻分类一致好的分类性能.

关键词: k-近邻分类; 局部均值; 类均值; 交叉验证
中图分类号: TP181 **文献标识码**: A

Nearest neighbour classification based on local mean and class mean

ZENG Yong, YANG Yu-pu, ZHAO Liang
(Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China. Correspondent: ZENG Yong, E-mail: zeng_yong@sjtu.edu.cn)

Abstract: The k-nearest neighbour classification is a very popular and successful nonparametric classification method, but its classification performance usually suffers from the existing outliers. To overcome the adverse effect of the existing outliers on classification performance, a variant of the k-nearest neighbour classification and a nearest neighbour classification method based on the local mean and class mean are proposed. The information of the local mean of the k nearest neighbours of the unclassified sample in each class and the knowledge of the ensemble mean are taken into account in the classification method. The proposed classification method overcomes the influence of the existing outliers and achieves a uniformly good classification performance compared with the traditional k-nearest neighbour classification.

Key words: k-nearest neighbour classification; Local mean; Class mean; Cross-validation

1 引 言

在模式识别领域里,通常得不到有关问题概率结构的全部知识,这就排除了利用最优的贝叶斯规则进行分类. 对于对象的内在概率分布较难获得的分类问题,利用非参数分类方法进行分类是必要的,其中 k-近邻分类应用得较为广泛. 根据该分类方法,将一个未分类样本分类为与它最接近的 k 个近邻中出现最多的那个类别. Cover 等[1]已经证明,当训练样本数与近邻数均趋于无穷且近邻数与训练样本数的比值趋于零时,k-近邻的分类误差率逼近最优的贝叶斯分类误差率[1]. k-近邻分类作为一种经典的非参数分类方法,已在垃圾邮件识别[2]、人脸识别[3]、飞机目标自动分类[4]、文本检索[5]等诸多领域得到了广泛的应用.

当可得的训练样本数与数据的内在维数之比较大时,k-近邻分类通常能取得较高的分类精度. 然而,在大多数情形下,可得的训练样本数通常较小,这就导致 k-近邻分类的分类精度显著降低,这也可以解释为什么出现了许多 k-近邻分类的变形来改善 k-近邻分类在小样本时的分类性能. 此外,非参数分类的分类性能通常由于离群点的存在而受到严重削弱[6],k-近邻分类当然也不例外.

Mitani 等[7]提出了一种基于局部均值的非参数分类方法(LMC),该方法不仅能克服离群点对分类性能的影响,而且在小样本情形下能够取得较好的分类性能. 如果选择未分类样本在每类训练样本集里的近邻数为 1,则该分类方法等价于最近邻分类(1-近邻分类);如果选择近邻数等于对应类里的训练样本数时,则等价于欧几里得距离分类[7].

对于模式分类问题,有两个重要的统计量:样

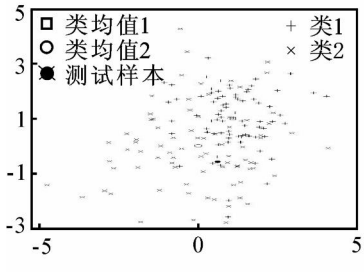
收稿日期: 2008-02-20; **修回日期**: 2008-09-12.
基金项目: 国家 973 计划项目(2004CB720703).
作者简介: 曾勇(1968—),男,四川邻水人,博士生,从事模式识别、神经网络的研究; 杨煜普(1957—),男,西安人,教授,博士生导师,从事智能控制、智能信息处理等研究.

本均值与样本方差. 与类可分离性有关的矩阵,如 Bhattacharyya 距离、离散矩阵以及 Fisher 判别式都是根据不同类里的均值与方差矩阵定义的. Mitani 等提出的基于局部均值的非参数分类只是利用了每类训练样本集里未分类样本的几个近邻的局部均值进行分类,而与类可分离性密切相关的类均值知识并未利用. 基于在类均值互不相同的情形下,可不可以既利用未分类样本在每类里的近邻的局部均值信息、又利用类均值的整体知识进行分类的想法,本文提出一个 k- 近邻分类的变形和一个基于局部均值与类均值的近邻分类方法.

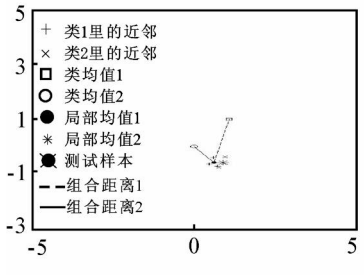
2 准备工作

为叙述方便,以下将本文提出的基于局部均值与类均值的近邻分类方法称为 NNCM(Narest neighbour classification based on local mean and class mean),将 Mitani 等提出的基于局部均值的非参数分类方法称为 LMC (A local mean-based nonparametric classification),将传统的 k- 近邻分类称为 KNN(K-nearest neighbour classification).

用图 1 来说明类均值向量对分类结果的影响,取测试样本在每个训练样本集里的近邻数为 3. 从图 1(a) 可以看出,测试样本位于类 1 的尾部且处于类 2 样本密集的区域,将测试样本分类为类 2 较为合理. 从图 1(b) 可以看出,测试样本与其在类 1 里的局部均值向量重合,对于近邻数为 3 的情形,无论用 KNN 还是用 LMC 进行分类,测试样本的标号都会被判为类 1. 正是由于类均值向量的影响,测试样本到类 1 均值点的距离与到类 1 里局部均值点的距离之和大于相应的到类2对应距离之和,所以将测试样本判为类 2,这里取距离权值为 1. 从图 1(a) 给出的样本分布来看,这样的分类更为合理,也避免了离群点对分类结果的影响.



(a) 样本分布



(b) 类均值对分类性能的影响

图 1 类均值对分类性能的影响

3 基于局部均值与类均值的近邻分类

对于 N 个可得的训练样本,令 $N_1, N_2, \cdots, N_M$ 表示对应于类 $\omega_1, \omega_2, \cdots, \omega_M$ 的训练样本数. 令 $x_j^{(1)}, \cdots, x_j^{(r)}$ 表示测试样本在类 ω_j 里的 r 个近邻,且 $X_j = \{x_j^i \mid i = 1,2,\cdots,N_j\}$ 表示属于类 ω_j 的训练样本集, μ_j 为训练样本集在类 ω_j 的均值向量,有

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_j^i. \tag{1}$$

NNCM 描述如下:

1) 利用得到的测试样本在类 ω_j 的 r 个近邻计算测试样本在类 ω_j 里的局部均值向量

$$y_j = \frac{1}{r}\sum_{i=1}^{r} x_j^{(i)}; \tag{2}$$

2) 计算测试样本与局部均值向量 y_j 在类 ω_j 的距离

$$d_j^l = \sqrt{(x - y_j)^{\mathrm{T}}(x - y_j)}; \tag{3}$$

3) 计算测试样本与类均值向量 μ_j 在类 ω_j 的距离

$$d_j^c = \sqrt{(x - \mu_j)^{\mathrm{T}}(x - \mu_j)}; \tag{4}$$

4) 计算组合距离

$$d_j = d_j^l + w * d_j^c; \tag{5}$$

5) 利用最近邻分类方法进行分类,若测试样本分类为 ω_c 并满足以下条件:

$$d_c = \arg\min\{d_j\},\ j = 1,2,\cdots,M. \tag{6}$$

在式(6) 中如果有多个 d_j (在多类分类情形里) 相等,则在对应 d_j 的几个类里随机选择一个类别号分配给测试样本. 这里 w 为距离加权系数,它反映了类均值向量对分类结果的影响程度,其值愈大,对分类结果影响愈大. 加权系数 w 的取值范围从 0 到 1,本文 w 根据下式取值:

$$w = 1.25^{-(i-1)} \text{ or } w = 0,\ i = 1,2,\cdots,41. \tag{7}$$

如果 $w = 0$,则提出的 NNCM 等价于 LMC. 对于 NNCM,关键问题是发现较优的近邻数 r^* 与加权系数 w^*. 这里采用 m 重交叉验证方法[8] 来获得较优的参数 r^* 与 w^*.

分两个阶段用交叉验证方法来获得较优的参数 r^* 与 w^*. 首先在固定近邻数 r 的情况下得到较优的加权系数 w_i^*,具体步骤如下:

Step1: 将可得的训练样本集分为训练集与测试集两个部分,每次的测试集不同.

Step2: 固定参数 r 与 w,利用 NNCM 对测试集进行分类,得到相应的误差 $e_i^{(r)}(r,w)$.

Step3: 重复 Step1 与 Step2 m 次,得到

$$E_{cv}^{(r)}(r,w) = \frac{1}{m}\sum_{i=1}^{m} e_i^{(r)}(r,w). \quad (8)$$

训练次数 m 对 $E_{cv}^{(r)}(r,w)$ 有影响,如果可得的训练样本集较小,则训练次数 m 等于训练样本数 N,否则 m 小于 N.

Step4: 重复 Step1 ～ Step3 共 42 次,每次距离加权系数 w 根据式(7) 取不同的值,得到局部最小分类误差为

$$\mathrm{Error}_l^*(r,w_l^*) = \arg\min\{E_{cv}^{(r)}(r,w)\},$$
$$w = 0, 1.25^{-41}, \cdots, 1.25^{-1}, 1. \quad (9)$$

第2阶段,根据下式得到较优的参数 r^* 与 w^*:

$$[r^*, w^*] = \arg\min\{\mathrm{Error}_l^*(r,w_l^*)\},$$
$$r = 1,2,\cdots,N_{\min},\ N_{\min} = \min(N_1,\cdots,N_M), \quad (10)$$

其中 $N_1, N_2, \cdots, N_M$ 是对应于类 $\omega_1, \omega_2, \cdots, \omega_M$ 的训练样本数.

4　实验结果

用4个数据集对NNCM的分类性能进行评估. Mitani 已经将 LMC 与最近邻分类、k- 近邻分类、Parzen 分类[9] 以及BP神经网络分类[10] 等4种分类方法按照平均误差率对 5 种分类方法的分类性能进行了比较,其结论如下:无视训练样本规模的大小与维数,LMC 的分类性能通常都优于其他 4 种分类方法. NNCM 在本质上与 LMC 相似,是对 LMC 的改进,因而具有 LMC 类似的分类性能. 在这里不再把 NNCM 与最近邻分类、k- 近邻分类、Parzen 分类以及 BP 神经网络分类等 4 种分类方法进行比较,只将 NNCM 与 LMC 以及传统的 k- 近邻分类进行比较,采用的距离度量是欧几里得距离.

4.1　实验数据

假定所用数据集里样本类概率的先验概率相同,4 个所用的数据集与 Mitani 所采用的数据集相同. 该实验中,测试样本都为 2000 个. 对每个数据集进行 100 次实验,每次所用数据由计算机重新产生. 对于每个数据集,将 100 次实验结果得到的分类误差的平均值作为该分类方法的分类误差,并给出相应的 95 % 置信区间.

所采用的 4 个数据集分别是 *I-Λ*, *I-4I*, *I-I*[6] 与 Ness[11]. 在这些数据集里,μ_i 与 Σ_i 分别表示样本类 ω_i 的均值向量与协方差矩阵. 重新描述这 4 个数据集如下:

1) *I-Λ* 数据由 8 维的高斯数据组成,即

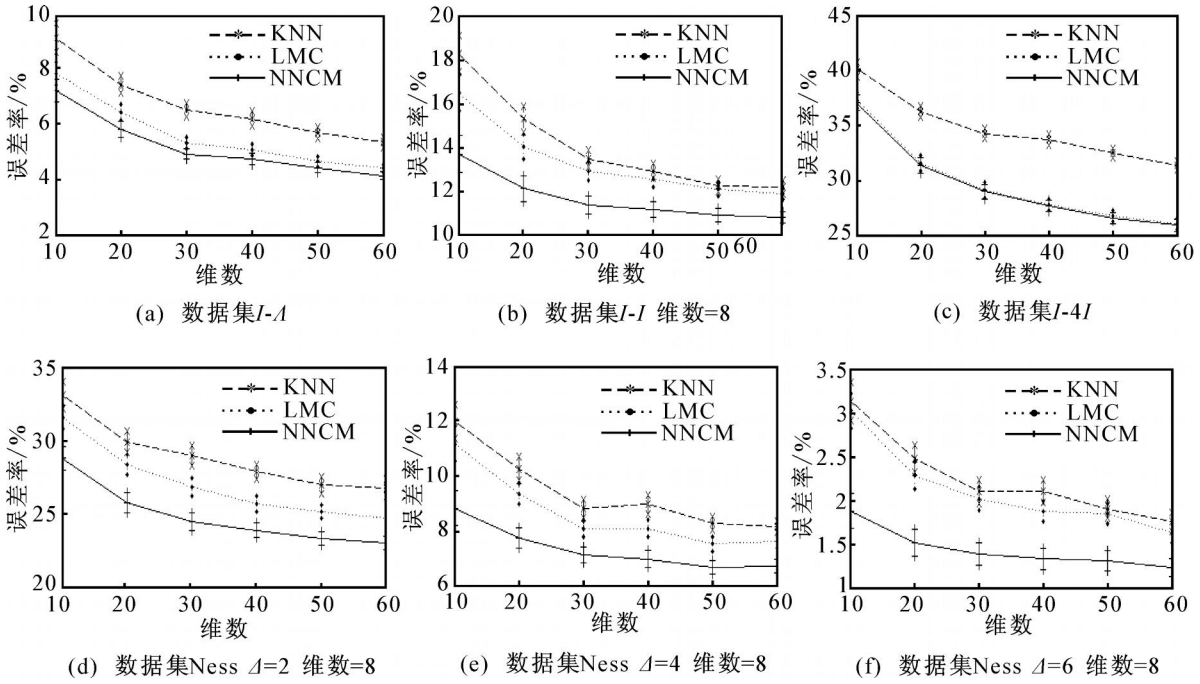
$$\mu_1 = \mathbf{0},$$
$$\mu_2 = [3.86, 3.10, 0.84, 1.64, 1.08, 0.26, 0.01]^{\mathrm{T}},$$
$$\Sigma_1 = \mathrm{I}_8,$$
$$\Sigma_2 = \mathrm{diag}[8.41, 12.06, 0.12, 0.22, 1.49, 1.77, 0.35, 2.73],$$

其中 I_i 与 diag[ ] 分别表示单位矩阵与对角矩阵.

2) *I-4I* 数据同样由 8 维的高斯数据组成,即

$$\mu_1 = \mu_2 = \mathbf{0},$$
$$\Sigma_1 = \mathrm{I}_8,$$
$$\Sigma_2 = 4\mathrm{I}_8.$$

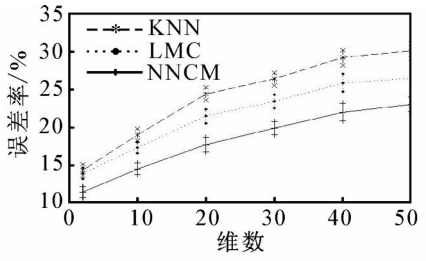
3) *I-I* 数据由 p 维的高斯数据构成,维数 p 可以



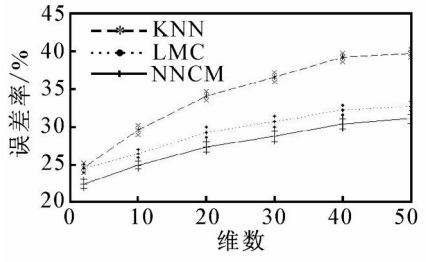
(a) 数据集*I-Λ*　(b) 数据集*I-I* 维数=8　(c) 数据集*I-4I*

(d) 数据集Ness Δ=2 维数=8　(e) 数据集Ness Δ=4 维数=8　(f) 数据集Ness Δ=6 维数=8

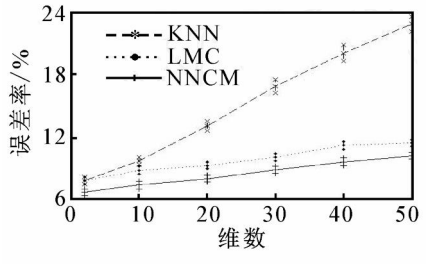
图 2　训练样本大小对分类性能的影响



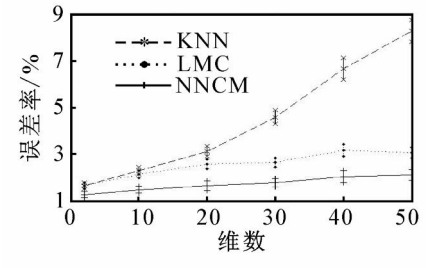
(a) 数据集*I-I* 训练样本数量=30



(b) 数据集Ness Δ=2 训练样本数量=30



(c) 数据集Ness Δ=4 训练样本数量=30



(d) 数据集Ness Δ=6 训练样本数量=30

图 3 数据特征维数对分类性能的影响

变化,即

$$\mu_1 = \mathbf{0},$$

$$\mu_2 = [2.56, 0, \cdots, 0]^{\mathrm{T}},$$

$$\Sigma_1 = \Sigma_2 = \mathrm{I}_p.$$

4) Ness 数据由 p 维的高斯数据构成,即

$$\mu_1 = \mathbf{0},$$

$$\mu_2 = [\Delta/2, 0, \cdots, 0, \Delta/2]^{\mathrm{T}},$$

$$\Sigma_1 = \mathrm{I}_p,$$

$$\Sigma_2 = \begin{bmatrix} \mathrm{I}_{p/2} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathrm{I}_{p/2} \end{bmatrix}.$$

在实验中设置 Δ 分别为 2,4,6 且从 2～50 变化维数 p.

4.2 实验结果

对 *I-*Δ, *I-*4*I*, *I-I* 与 Ness 数据集,从两个方面对分类性能进行评估:1) 训练样本集大小对分类性能的影响;2) 特征维数对分类性能的影响. 图2和图3 给出了 4 个数据集上 KNN,LMC 以及 NNCM 三种分类方法分类性能的比较结果. 从图 2 和图 3 可以看出,NNCM 与 KNN 相比具有一致好的分类性能;与 LMC 相比,在大多数情形下,NNCM 的分类性能也优于 LMC. 同时也注意到,在类均值相同的情况下,NNCM 的分类性能与 LMC 并没有多大改善.

5 结 论

本文提出一个 k-近邻分类的变形和基于局部均值与类均值的近邻分类方法. 该分类方法能取得比 k-近邻分类一致好的分类性能. 对两个重要的统计量:样本均值与样本方差,只用了样本均值这一统计量,对于类均值互同或类均值不同但方差较大的情形,如何同时利用这两个统计量进行分类,是今后关注和研究的方向.

参考文献(References)

[1] Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Trans on Information Theory, 1967, 13(1): 21-27.

[2] 李斌, 李义兵, 何红波. 基于 LZ 复杂性相似度的垃圾邮件识别[J]. 计算机工程与应用, 2007, 43(29): 176-178.
(Li B, Li Y B, He H B. LZ complexity similarity based spam detection [J]. Computer Engineering and Applications, 2007, 43(29): 176-178.)

[3] 贺云辉, 赵力, 邹采荣. 基于核的最近邻特征重心分类器及人脸识别应用[J]. 电路与系统学报, 2007, 12(2): 5-10.
(He Y H, Zhao L, Zou C R. Kernel-based nearest neighbor feature centroid classifiers for face recognition [J]. J of Circuits and Systems, 2007, 12(2): 5-10.)

[4] 丁建江, 张贤达. 基于调制特征的飞机目标自动分类[J]. 清华大学学报, 2003, 43(7): 887-890.
(Ding J J, Zhang X D. Automatic classification of aircraft based on modulation features[J]. J of Tsinghua University, 2003, 43(7): 887-890.)

[5] Chen C Y, Chang C C, Lee R C T. A near pattern-matching scheme based upon principal component analysis[J]. Pattern Recognition Letters, 1995, 16(4): 339-345.

[6] Fukunaga K. Introduction to statistical pattern recognition[M]. 2nd ed. San Diego: Academic Press, 1990.

(下转第 556 页)

行训练,而 90 % 用于测试,即只有 30 组样本用于训练,而本实验中模糊评判模型设定的未知数就达到 33 个,这样训练效果不理想也是可以理解的;而使用测试集验证的方式,训练样本达到 300 组,训练后的评估精度大幅提高,达到了应用可接受标准,特别是 EDA 进化后,评估精度能够随列车运行数据的不断丰富稳中有升,能够胜任保障列车运行安全的任务.

通过比较 4.1 节与 4.2,4.3 节的 EDA 进化结果可以发现,由于 4.1 节编码时不考虑变量联系,而后两节将关联分解,导致相同参数情况下后者效果更优,说明概率模型与研究问题变量关系匹配的重要性.在避免使用复杂概率模型的情况下,本文使用将变量关系分解的形式编码,得到了很好的效果.

5 结 论

本文主要研究了基于进化学习的磁浮列车系统故障等级评判方法.针对传统模糊综合评判方法模型参数难以确定和优化的缺点,提出了基于分布估计算法的模糊参数优化方法,通过使用 EDA 对评判模型进行逼近建模,实现了模糊评估模型参数自动学习和优化.性能评测显示,基于 EDA 的故障综合评判方法能够较快收敛到最优结果,并有很高的评判精度,效果明显优于遗传算法和其他机器学习算法,能够胜任磁浮列车系统故障评级的任务.

参考文献(References)

[1] 周树德, 孙增圻. 分布估计算法综述[J]. 自动化学报, 2007, 33(2): 113-124.
(Zhou S D, Sun Z Q. A survey on estimation of distribution algorithms [J]. Acta Automatica Sinica, 2007, 33(2): 113-124.)

[2] 龙志强, 吕治国. 基于模糊综合评估的磁浮列车故障诊断系统[J]. 信息与控制. 2004, 33(2): 227-230.
(Long Z Q, Lv Z G. The fault diagnostic system of maglev train based on fuzzy comprehensive evaluation [J]. Information and Control. 2004, 33(2): 227-230.)

[3] 黄文虎, 夏松波, 刘瑞岩. 设备故障诊断原理、技术及应用[M]. 北京: 科学出版社, 1996.
(Huang W H, Xia S B, Liu R Y. Theory technology and application on equipment fault diagnosis [M]. Beijing: Science Press, 1996.)

[4] Larranaga P, Etxeberria R, Lozano J A, et al. Optimization by learning and simulation of bayesian and gaussian networks[M]. Spain: University of the Basque Country, 1999.

[5] Larranaga P, Etxeberria R, Lozano J A, et al. Optimization in continuous domains by learning and simulation of gaussian networks[C]. Proc of the 2000 Genetic and Evolutionary Computation Conference Workshop Program. Las Vegas: Nevada, 2000: 201-204.

[6] Sebag M, Ducoulombier A. Extending population based incremental learning to continuous search spaces [C]. Proc of the 5th Conf on Parallel Problem Solving from Nature-PPSN V. Springer-Verlag, 1998: 418-427.

[7] MÄuhlenbein H. The equation for response to selection and its use for prediction [J]. Evolutionary Computation, 1997, 5(3): 303-346.

[8] Shapiro J L. Drift and scaling in estimation of distribution algorithms[J]. Evolutionary Computation, 2005, 13(1): 99-123.

[9] Hohfeld M, Rudolph G. Towards a theory of population based incremental learning[C]. Proc of the 4th Int Conf on Evolutionary Computation. IEEE, 1997: 1-5.

[10] Cristina G, Lozano J A, Larranaga P. Analyzing the PBIL algorithm by means of discrete dynamical systems [J]. Complex Systems, 2001, 12(4): 465-479.

(上接第 550 页)

[7] Mitani Y, Hamamoto Y. A local mean-based nonparametric classifier [J]. Pattern Recognition Letters, 2006, 27(10): 1151-1159.

[8] Duda R O, Hart P E, Stork D G. Pattern classification [M]. 2nd ed. New York: John Wiley Sons, 2001.

[9] Jain A K, Ramaswami M D. Classifier design with Parzen windows[z]. Amsterdam: Elsevier, 1988.

[10] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation [z]. Cambridge: MIT Press, 1986.

[11] Ness J V. On the dominance of nonparametric Bayes rule discriminant algorithms in high dimensions [J]. Pattern Recognition, 1980, 12 (3): 355-368.