第 21 卷 第 10 期 控 制 与 决 策 2006 年 10 月
Vol. 21 No. 10 Control and Decision Oct. 2006

文章编号: 1001-0920(2006)10-1103-06

基于共享最近邻聚类和模糊集理论的分类器

李订芳[a], 胡文超[a], 何炎祥[b]
(武汉大学 a. 数学与统计学院, b. 计算机学院, 武汉 430072)

摘 要: 提出一种基于共享最近邻聚类和模糊集理论的分类器. 首先, 在提出与核点密切相关的核半径概念的基础上, 应用共享最近邻聚类得到正常类空间的部分核点和核半径, 建立求解正常类空间补充核点的多目标优化模型, 从而获得刻画正常类空间的全部核点和核半径. 然后, 将模糊集理论引入正常类的类属划分中, 利用核点和核半径定义正常类的隶属度函数, 建立基于隶属度函数的分类函数或分类器. 实验表明, 该分类器能处理包含噪音、孤立点和不规则子类的高维数据集的分类问题.

关键词: 分类器; 共享最近邻聚类; 模糊集; 遗传算法; 优化模型
中图分类号: TP391.4 **文献标识码**: A

Classifier Based on Shared Nearest Neighbor Clustering and Fuzzy Set Theory

LI Ding-fang[a], *HU Wen-chao*[a], *HE Yan-xiang*[b]
(a. School of Mathematics and Statistic, b. School of Computer Sciences, Wuhan University, Wuhan 430072, China. Correspondent: LI Ding-fang, E-mail: dfli@whu.edu.cn)

Abstract: A classifier based on shared nearest neighbor clustering and fuzzy set theory (SNNFT) is proposed. The concept of core radii closely related with core point is defined and a portion of core points and core radii are obtained by applying shared neighbor clustering. The multi-objective optimization model of complementary core points is established. Consequently all core points and core radii are obtained to depict normal class space. By introducing fuzzy set theory to partition normal class space, classification function or classifier based on membership function of normal class defined using core points and core radii are constructed. Experiments show that SNNFT can cope with classification problem with high dimension dataset which contains noise, outliers and irregular sub clusters.

Key words: Classifier; Shared nearest neighbor clustering; Fuzzy set; Genetic algorithm; Optimization model

1 引 言

分类器是一种能够把数据集中的样本映射到给定类别的分类函数或分类模型, 已被广泛应用于数据挖掘、文字识别、文本分类、语音识别、基于内容的多媒体数据库检索、图像处理、自然语言理解等领域[1~4].

在科学实验和生产管理工作中, 由于人力、财力和时间等因素, 人们很难收集所有的样本并对收集的训练样本指定类别, 通常只能获得属于某一个子类的训练样本[5]. 另外, 医疗诊断、信息安全等领域存在大量包含有噪音、孤立点和不规则子类的高维数据, 构造能有效处理这类数据的分类器是这些领域的重要课题. 构造分类器的过程分为训练和测试两个阶段, 训练阶段主要建立一个刻画训练集的模型, 测试阶段则利用模型对测试集进行分类. 目前分类器的构造方法主要有统计方法、机器学习方法、遗传优化方法和神经网络方法等. 典型的分类器包括 CART, C4.5, Nearest neighbor, 贝叶斯分类和 SVM[6,7]. 其中基于聚类的分类器主要有 K-means[8]和 FCC[9], 它们的基本思想是对训练集(正

收稿日期: 2005-08-01; **修回日期**: 2005-10-24
基金项目: 国家自然科学基金重大研究计划项目(90104005); 国家重大基础研究前期研究专项(2003CCA 00200).
作者简介: 李订芳(1966—), 男, 湖南平江人, 副教授, 博士后, 从事数据挖掘、智能信息处理等研究; 胡文超(1982—), 男, 湖北沔阳人, 硕士, 从事数据挖掘、机器学习等研究.

常类样本)聚类得到子类,通过子类的信息(如中心和半径)来刻画正常类空间,并由此定义分类函数或分类器[10]. Y-means 和 FCC 中采用 K-means 聚类算法或其变种训练分类器,由于 K-means 聚类算法存在不能有效处理包含有噪音、孤立点和不规则子类数据集的缺陷[11],由其训练的分类器很难正确地刻画训练集

共享最近邻聚类(SNN)是 Levent Ertoz[12]提出的一种整合了多种聚类思想的聚类方法,可以处理包含噪音、孤立点以及任意形状、大小、密度的子类的高维数据的聚类问题. 本文针对医疗诊断、信息安全等领域大量存在的二分问题,提出一种基于共享最近邻聚类和模糊集理论的分类器(SNNFT). 首先提出与核点密切相关的核半径概念,然后应用共享最近邻聚类方法 SNN 得到正常类空间的部分核点和核半径. 针对由 SNN 得到的核点不能完整刻画正常类空间的问题,建立了求解正常类空间补充核点的多目标优化模型,并采用基于小生境技术的遗传算法求解得到补充核点,从而获得刻画正常类空间的全部核点和核半径. 再将模糊集理论引入正常类类属划分中,利用核点和核半径定义了正常类的隶属度函数. 最后建立了基于隶属度函数的分类器. 多个数据集上的实验表明,SNNFT 是一种有效的基于聚类的分类器,能处理包含噪音、孤立点和不规则子类高维数据集的分类问题

2 SNN 聚类

聚类是一种重要的数据挖掘手段,它根据数据之间的"相似程度"将数据划分成不同的数据集合,使得数据集合内部对象之间相似度变大,同时数据集合之间的差别增大[13]. 传统的聚类方法大部分直接利用距离定义相似性,如两个对象之间的距离越近,相似性就越高. 由于数据在高维空间中是稀疏的,点对之间的距离或相似性度量变得趋于一致,使得聚类变得困难. 因此,在高维数据集中,直接用距离定义相似性度量很难刻画一个对象与另一个对象是否相似[14]. 但并不是距离公式选取的问题,而是直接用距离来定义相似性度量不合适. 2003 年,Levent Ertoz 等提出的 SNN 方法解决了高维数据的聚类问题. SNN 的核心思想是根据共享的最近邻来定义两个对象之间的相似性,通过引入 CURE[15]中代表点的思想定义核点,能处理包含有任意形状、大小子类时数据的聚类问题;同时又引入 DBSCAN[16]中密度的思想,使 SNN 能实现有噪音、孤立点和不同密度子类时数据的聚类

设 Ω 是 n 维样本空间,$p, q \in \Omega$,N 为训练集样本数. 点 p 的最近 k 邻接列表表示与点 p 距离最近的 k 个点组成的集合.

定义 1(相似性度量) 点 p 和 q 之间的相似性度量或连接强度定义为 p 和 q 的最近 k 邻接列表中相同点的数目,即

$$\text{Similarity}(p, q) = \text{size}(nn[p] \cap nn[q]), \quad (1)$$

其中:$nn[p]$ 和 $nn[q]$ 分别是 p 和 q 的最近 k 邻接列表,$\text{size}(A)$ 是集合 A 的大小.

定义 2(共享最近邻图) $\langle V, E\rangle$ 表示共享最近邻图. $\forall u, v \in \Omega, u, v$ 之间有连接当且仅当 $u \in nn[v]$ 且 $v \in nn[u]$. 连接强度由定义 1 中式(1)给出.

定义 3(密度) 点 p 的密度为 p 的最近 k 邻接列表中与 p 相似的点的数目,即

$$\text{Density}(p) = \text{count}(\text{Similarity}(p, q) \geq \text{strong}), \quad (2)$$

其中:q 在 p 的最近 k 邻接列表中,strong 为判断两个点是否相似的阈值,即两个点相似的条件是它们共享了 strong 个或更多的最近邻.

定义 4(核点和骨架) 核点(代表点)为高密度的点,骨架为所有核点构成的集合. 给定核点在整个训练样本中的比例 topic,则核点(代表点)为密度最大的 $N * \text{topic}$ 个点. 有

$$\text{Skeleton} = \{\text{kernel} \mid \text{Density}(\text{kernel}) \geq \text{topic\_threshold}\}, \quad (3)$$

其中:Skeleton 表示骨架,kernel 表示核点,topic_threshold 为密度排在第 $N * \text{topic}$ 位的样本点的密度值. 记第 i 个核点为 k_i,如果核点有 s 个,则骨架还可表示为 $\text{Skeleton} = \{k_i \mid i = 1, 2, \cdots, s\}$. 给定合并阈值 merge_threshold,如果两个核点之间的相似性度量大于 merge_threshold,则将它们合并为一个聚类簇.

定义 5(噪音) 噪音为所有不与任何一个核点相似的非核点,即

$$\text{Noise} = \{\text{non\_kernel} \mid \text{Similarity}(\text{non\_kernel}, k_i) < \text{strong}\}, \quad (4)$$

其中:Noise 表示噪音,non_kernel 表示非核点,$1 \leq i \leq s$. 在实际应用中,噪音也定义为低密度的点,即

$$\text{Noise} = \{\text{non\_kernel} \mid \text{Density}(\text{non\_kernel}) < \text{noise\_threshold}\}. \quad (5)$$

给定噪音在整个训练样本中的比例 noise,则 noise_threshold 为密度排在第 $N * (1 - \text{noise})$ 位的样本点的密度值.

SNN 对样本进行聚类,得到由核点表征的聚类簇.

3 分类器构造

SNN 聚类在消除或降低噪音、孤立点和不规则

子类的影响以及提高聚类性能方面具有优越性, 但仅由 SNN 聚类获得的是样本空间的部分核点, 为使所构建的分类器具有完全的分类能力, 必须补充刻画样本空间的核点. 下面首先给出核半径的概念, 由 SNN 得到部分核点和核半径, 然后建立求解补充核点的多目标优化模型, 并结合模糊集理论构造分类器. 分类器构造的流程如图 1 所示.

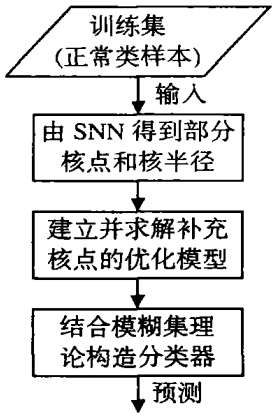

图 1　分类器构造流程

3.1　核点和核半径计算

首先定义与核点密切相关的核半径概念.

定义 6 (核半径)　核点 k_i 的核半径 r_i 为 k_i 的最近邻接列表中与 k_i 的相似性大于合并阈值 merge-thesho ld 的所有点与 k_i 的最大距离(如欧氏距离), 即

$$r_i = \max\{\mathrm{dist}(k_i, p) \mid \mathrm{Similarity}(k_i, p) > \mathrm{merge\_thesho ld}\}, \quad (6)$$

其中: $p \in nn[k_i]$, $nn[k_i]$ 表示 k_i 的最近邻接列表, $\mathrm{dist}(k_i, p)$ 表示 p 与 k_i 的距离(如欧氏距离).

由核半径的定义, 训练集中由核点 k_i 表征的子类可表示为 $\{x \mid \mathrm{dist}(k_i, x) \leq r_i\}$.

由 SNN 计算核点和核半径的算法如下:

算法 1

1) 计算点对之间的欧氏距离, 构造距离矩阵 $D = (a_{i,j})_{n\times n}$, 其中 x_i, y_j 为样本点,

$$a_{i,j} = ((x_{i,1} - y_{j,1})^2 + (x_{i,2} - y_{j,2})^2 + \cdots + (x_{i,n} - y_{j,n})^2)^{1/2},$$

由 D 构造最近邻接列表矩阵 nn;

2) 由定义 1 和 nn 构造相似性矩阵 SI;

3) 由 SI 和定义 2 构造共享最近邻图 SH;

4) 由 SH 和式(2) 计算每个样本点的密度;

5) 设定阈值 strong, topic, noise, merge-thesho ld, 由定义 4～6 得到核点和核半径 $\{(k_1, r_1), (k_2, r_2), \cdots, (k_s, r_s)\}$, 并去除训练集中的噪音.

3.2　补充核点的优化模型

由 SNN 计算得到的核点表征与核点密切相关的点组成的聚类簇. 然而, 正常类样本中存在与得到的核点均不相关的部分点, 即正常类样本中有一部分不包含在任何聚类簇中的点. 所以, 核点集(骨架)只能部分地刻画正常类空间, 达不到完整刻画正常类空间的目的, 需要补充核点. 如图 2 所示, 超球体的球心为由 SNN 得到的核点, x 为需要补充的核点.

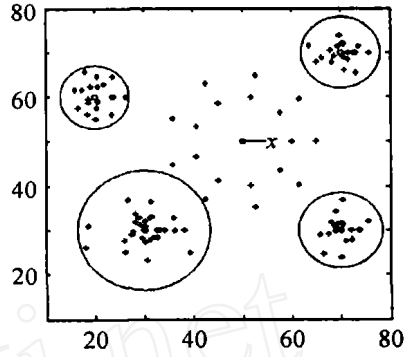

图 2　补充的核点

设 T 表示训练集, A 表示由 SNN 得到的核点所表征的聚类簇的并, $U = T - A$ 表示不包含在任何聚类簇中的点. 由定义 3 可知, 核点 $\{k_1, k_2, \cdots, k_s\}$ 是 A 中高密度的点, 则补充的核点是 U 中高密度的点. 一方面, 补充的核点的密度应尽可能大; 另一方面, 与由 SNN 得到的核点的重叠尽可能小.

定义补充核点 k^* 与已有核点 k_i 之间的重叠为

$$\mathrm{Overlapping}(k^*, k_i) = \exp\left(\frac{-\mathrm{dist}(k^*, k_i)^2}{2((r^* + r_i)/2)^2}\right), \quad i = 1, 2, \cdots, s \quad (7)$$

当核点之间的距离为 0 时, 重叠最大, 最大值为 1; 当核点之间的距离大于两个核点对应的核半径之和时, 重叠接近于 0.

由式(2) 和(7), 可建立如下求解补充核点的多目标优化模型:

Maximize:

$$\mathrm{Covering}(k^*) = \frac{\mathrm{Density}(k^*)}{k}, \ r^* > 0; \quad (8)$$

Minimize:

$$\mathrm{Overlapping}(k^*) = \frac{\sum_{i=1}^{s}\mathrm{Overlapping}(k^*, k_i)}{s}, \quad r^* > 0. \quad (9)$$

引入目标函数(8) 相对于目标函数(9) 的重要性程度项 λ, 可将该多目标优化问题(8), (9) 转化为如下单目标优化问题:

Maximize:

$$f(k^*) = \lambda \frac{\mathrm{Density}(k^*)}{k} - (1-\lambda)\frac{\sum_{i=1}^{s}\mathrm{Overlapping}(k^*, k_i)}{s}, \quad (10)$$

其中: $r^* > 0$, $r_i (i = 1, 2, \cdots, s)$ 为核点 $k_i (i = 1, 2, \cdots, s)$ 的核半径, k 为邻接列表大小, r^* 为补充核点 k^* 的核半径, r^* 可通过构造 $S \cup k^*$ 的共享最近邻图得到, 而 $S \cup k^*$ 的共享最近邻图可以在 S 的共享最近邻图的基础上构造. 由上述分析可知, 该目标函数的极值点就是需要补充的核点.

3.3 优化模型求解的遗传算法

优化问题(10)中目标函数的解析性太差, 而传统优化方法一般需要目标函数导数的信息. 遗传算法在搜索过程中仅使用评价函数值来评估个体或解的优劣, 并作为以后遗传操作的依据. 但标准遗传算法求解多峰值函数的优化问题时, 经常只能找到个别的最优解, 甚至得到的是局部最优解. 为了能找出全部的最优解, 本文采用 De Jong 在 1975 年提出的基于小生境技术的遗传算法求解优化问题(10). 该算法可以保持解的多样性, 同时具有很高的全局寻优能力和收敛速度.

遗传算法中个体的染色体包含了补充核点 k^* 和核半径 r^* 的信息, 其中 k^* 为 n 维向量. 初始种群的个体随机产生, 交叉操作作用在 k^* 的每一维上, 变异操作作用在 k^* 的每一位上. 设 t 时刻的种群为 $P(t)$, 经过交叉操作的种群记为 $P_c(t)$, 经过变异操作的种群记为 $P_m(t)$. 为保证适应度函数值非负, 定义适应度函数为目标函数(10)加上 1, 另外限制搜索范围为 $r^* > 0$, 即有

$$F = \begin{cases} \lambda \dfrac{\mathrm{Density}(k^*)}{k} - (1-\lambda)\dfrac{\sum_{i=1}^{s}\mathrm{Overlapping}(k^*, k_i)}{s} + 1, \\ \quad r^* > 0; \\ 0, \text{否则.} \end{cases} \quad (11)$$

基于小生境技术的遗传算法的求解补充核点的算法如下:

算法 2

1) 通过初始化过程, 随机产生第一代群体 $P(t)$;

2) 由式(11)计算 $P(t)$ 的适应度值;

3) 对群体进行交叉、变异操作;

4) 计算交叉、变异后群体的适应度值;

5) 对群体进行小生境处理, 调整其适应度值以保持群体多样性;

6) 采用轮盘赌法和最优个体保存策略选择生成下一代群体;

7) 重复 3)～6), 直到迭代次数超过预先设定值.

记由算法 2 得到的最佳个体(补充核点)为 $\{k_{s+1}, k_{s+2}, \cdots, k_t\}$, $r_i (i = s+1, s+2, \cdots, t)$ 为对应的核半径. 于是正常类空间的全部核点和对应的核半径分别为 $\{k_1, k_2, \cdots, k_s, k_{s+1}, \cdots, k_t\}$ 与 $\{r_1, r_2, \cdots, r_s, r_{s+1}, \cdots, r_t\}$.

3.4 分类器函数

得到全部核点和对应的核半径后, 可由清晰规则来判断新样本是否属于由核点表征的聚类簇, 即新样本属于聚类簇当且仅当新样本与核点的距离不大于核半径. 清晰规则把新样本严格划分到某个聚类簇中, 具有非此即彼的性质. 而在医疗诊断和信息安全等领域, 如健康和患病、正常和异常(入侵)等之间并没有严格的边界, 它们在性态和类属方面存在着中介性, 需要一定的平滑过渡, 即适合进行软划分. 模糊集理论通过定义集合的隶属度为软划分提供了理论工具. 定义核点 k_i 和核半径 r_i 所代表子类的隶属成员函数为

$$\mu_i(x) = \exp\left(-\frac{\mathrm{dist}(x, k_i)^2}{2r_i^2}\right), \quad i = 1, 2, \cdots, t \quad (12)$$

给定新样本 x, 由式(12)计算 x 对每个子类的隶属度, 则 x 属于聚类簇 i^* 当且仅当 $i^* = \arg\max_{1 \le i \le t} \mu_i(x)$. 在二分问题中, 只需判断样本是否属于正常类, 而不必判断样本属于哪个子类, 即只需给出正常类的隶属成员函数. 定义正常类为由所有核点和核半径表征的聚类簇的并, 则正常类的隶属成员函数为所有子类隶属成员函数的最大值, 即

$$\mu_{\mathrm{normal}}(x) = \max\{\mu_i(x) \mid \forall\, i = 1, 2, \cdots, t\}. \quad (13)$$

为了最后判定 x 是正常类还是异常类, 设定正常类阈值 $\theta \in [0, 1]$, 则分类器函数为

$$\mathrm{Classifier}(x) = \begin{cases} \mathrm{normal}, & \mu_{\mathrm{normal}}(x) \ge \theta, \\ \mathrm{abnormal}, & \mu_{\mathrm{normal}}(x) < \theta \end{cases} \quad (14)$$

即如果 $\mu_{\mathrm{normal}}(x) \ge \theta$, 则判定 x 属于正常类, 否则 x 属于异常类.

3.5 时间复杂性分析

SNNFT 的时间复杂性由构建共享最近邻图的时间和遗传算法求解补充核点需的时间决定.

构建共享最近邻图的时间依赖于给定数据集的维数. 对低维数据集而言, 可以通过 k-d 树快速计算

和查找最近的 k 个邻居, 如果有 n 个数据项, 则其时间复杂性为 $O(n\log n)$. 对高维数据集而言, 没有有效的方法来构建共享最近邻图, 而只能通过计算所有点对之间的距离来构建共享最近邻图, 此时时间复杂性为 $O(n^2)$. 遗传算法是启发式算法, 其时间复杂性由遗传代数和群体规模决定, 与数据集的规模无关, 因此时间复杂性为 $O(1)$. 综上所述, SNNFT 的时间复杂性为 $O(n^2)$.

4　实验研究

刻画分类器性能的指标主要是检测率和误警率. 其中检测率为被正确判定为异常的异常类样本占所有异常类样本的百分数, 用 DR 表示; 误警率为被错误判定为异常的正常类样本占所有正常类样本的百分数, 用 FA 表示. 记被正确判定为异常的异常类样本数目为 TP, 被错误判定为正常的异常类样本数目为 FN, 被错误判定为异常的正常类样本数目为 FP, 被正确判定为正常的正常类样本数目为 TN, 则有

$$DR = \frac{TP}{TP + FN}, FA = \frac{FP}{TN + FP}.$$

变动正常类阈值 θ, 可得到坐标点 $(FP, TP)_\theta$, 将这些坐标点画在笛卡尔坐标系中就得到了 ROC (Receiver Operating Characteristic) 图, ROC 图可以用来评价分类器的性能[17]. 如果 m 分类器的 ROC 图在 n 分类器的 ROC 图上方, 则认为 m 分类器比 n 分类器好. 记只用 SNN 得到部分核点所构造的分类器为 B.

4.1　补充核点的必要性

选取医疗诊断领域的 Liver Disorders 和 Pima Indians Diabetes 数据集来说明补充核点的必要性. 实验中, 从 Liver Disorders 数据集中随机挑选 140 个正常类记录组成训练集, 其余的 205 个记录构成测试集. 从 Pima Indians Diabetes 数据集中随机挑选 250 个正常类记录组成训练集, 其余的 518 个记录构成测试集. 图 3, 图 4 分别是两个数据集对应的 ROC 图. 由图 3 和图 4 可知分类器 B 的检测率随着

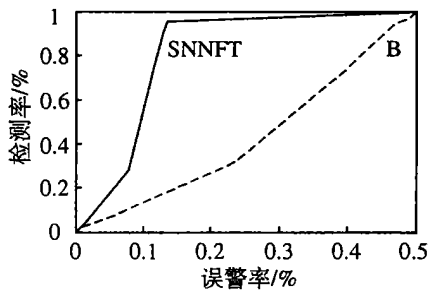

图 3　Liver Disorders 的 ROC 图

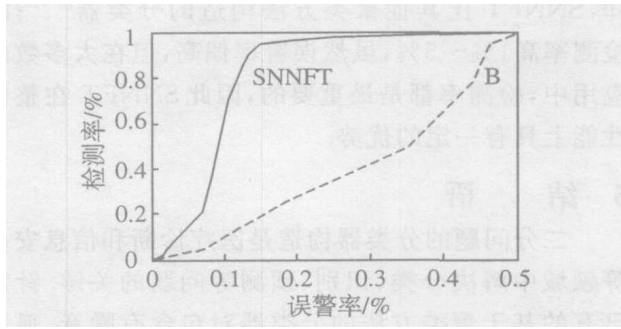

图 4　Pima Indians Diabetes 的 ROC 图

误警率的增加上升很慢, 分类器的性能较差; 而分类器 SNNFT 比 B 的性能则有很大的提高. 这表明只用 SNN 得到部分核点构造的分类器 B 不能完整地刻画正常类空间, 从而导致分类器 B 的误警率较高, 需要补充核点.

4.2　对噪音的不敏感性

使用合成数据集来说明 SNNFT 对于噪音数据的不敏感性. 其中合成数据由 1451 个正常类样本, 2 179 个异常类样本以及 1 000 个噪音样本组成. 分别选取正常类样本和 0%, 10%, 20% 的噪音样本作为训练集进行实验, 图 5 是不同比例噪音训练集对应的 ROC 图. 从图中可知, SNNFT 在 3 个训练集上的性能接近, 对于同样的误警率, 检测率的变化不超过 5%. 因此 SNNFT 对噪音是不敏感的.

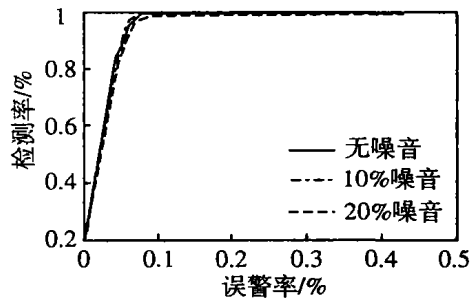

图 5　不同比例噪音训练集对应的 ROC 图

4.3　与部分基于聚类的分类器的性能比较

选取信息安全领域的 KDDCup'99 数据集来比较 SNNFT 与其他基于聚类的分类器的性能. 从 10% 的 KDDCup'99 数据集中随机选取 10 000 条正常类记录作为训练集, 55 535 条记录作为测试集, 属性全部选取连续属性. 表 1 是基于不同聚类算法的分类器在 KDDCup'99 数据集上的表现. 由表可

表 1　基于不同聚类算法的分类器在 KDDCup'99 数据集上的表现

聚类算法	检测率/%	误警率/%
Y-means[8]	89.89	1.00
FCC[9]	98.1	1.6
FAD+without PCA[10]	94.09	7.84
SNNFT	99.15	5.02

知, SNNFT 比其他聚类方法构造的分类器[8~10]的检测率高 1%~5%, 虽然误警率偏高, 但在大多数的应用中, 检测率都是最重要的, 因此 SNNFT 在整体性能上具有一定的优势.

5 结 语

二分问题的分类器构造是医疗诊断和信息安全等领域中解决分类、识别、预测等问题的关键. 针对已有的基于聚类方法的分类器对包含有噪音、孤立点和任意形状、大小和密度子类的高维数据集适应性差的问题, 本文构造了一种基于共享最近邻聚类和模糊集理论的分类器. 在保持共享最近邻聚类对数据质量与分布不敏感的优点的基础上, 通过提出与核点密切相关的核半径概念, 建立补充核点的优化模型, 得到了完整刻画正常类空间的全部核点和核半径, 并进一步结合模糊集理论构造了分类器. 实验表明, 直接利用共享邻聚类方法构造的分类器的性能无法满足实际需要, 而通过建立补充核点(共享最近邻聚类无法求出的核点)的优化模型, 结合模糊集理论构造的分类器 SNNFT 不仅对于噪音数据具有不敏感性, 而且其性能也有极大的提高, 超过 99% 的检测和小于 6% 的误警率说明 SNNFT 是一种有效的分类器.

参考文献(References)

[1] Richard O D, Peter E H, David G S. *Pattern Classification* [M]. 2nd ed. New York: Wiley, 2001: 6-10.

[2] Dietrich Paulus, Joachim Hornegger. *Applied Pattern Recognition* [M]. 2nd ed. Vieweg: Braunschweig, 1998: 1-10.

[3] 万红梅, 金连文, 尹俊勋, 等. 结合距离分类器的神经网络手写体汉字识别[J]. *计算机工程与应用*, 2004, 40(11): 55-57.
(Wan H M, Jin L W, Yin J X, et al. A Neural Network Method Combined with Distance-classifier for Handwritten Chinese Character Recognition [J]. *Computer Engineering and Application*, 2004, 40(11): 55-57.)

[4] 李荣陆, 胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法[J]. *计算机研究与发展*, 2004, 41(4): 539-545.
(Li R L, Hu Y F. A Density-based Method for Reducing the Amount of Training data in KNN Test Classification [J]. *Computer Research and Development*, 2004, 41(4): 539-545.)

[5] Byeungwoo J, David A L. Partially Supervised Classification Using Weighted Unsupervised Clustering [J]. *IEEE Trans on Geoscience and Remote Sensing*, 1999, 37(2): 1073-1079.

[6] David Hand, Heikki Mannila, Padhraic Smyth. *数据挖掘原理*[M]. 北京: 机械工业出版社, 2003: 209-232.
(David Hand, Heikki Mannila, Padhraic Smyth. *Principles of Data Mining* [M]. Beijing: China Machine Press, 2003: 209-232.)

[7] Sanjeev R K, Gabor L, Santosh S V. Learning Pattern Classification — A Survey [J]. *IEEE Trans on Information Theory*, 1998, 44(6): 2178-2206.

[8] Guan Y, Ghorbani A, Belacel N. Y-means: A Clustering Method for Intrusion Detection [A]. *Proc of Canadian Conf on Electrical and Computer Engineering* [C]. Quebec, 2003: 1083-1086.

[9] Qiang Wang, Vasileios Megalooikonomou. A Clustering Algorithm for Intrusion Detection [A]. *SPIE Conf on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security* [C]. Orlando, 2005: 31-38.

[10] Elizabeth Leon, Olfa Nasraoui, Jonatan Gomez. Anomaly Detection Based on Unsupervised Niche Clustering with Application to Network Intrusion Detection [A]. *Proc of IEEE Conf on Evolutionary Computation* [C]. Portland, 2004: 502-508.

[11] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review [J]. *ACM Computing Surveys*, 1999, 31(3): 264-323.

[12] Levent Ertoz, Michael Steinbach, Vipin Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data [A]. *Proc of the 3rd SIAM Int Conf on Data Mining* [C]. California, 2003: 47-58.

[13] Margaret H D. *Data Mining Introductory and Advanced Topics* [M]. Beijing: Tsinghua University Press, 2003.

[14] Jarvis R A, Patrick E A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors [J]. *IEEE Trans on Computers*, 1973, C-22(11): 1025-1034.

[15] Guha S, Rastogi R, Shim K. Cure: An Efficient Clustering Algorithm for Large Databases [A]. *Proc of the ACM SIGMOD Int Conf on Management of Data* [C]. Seattle, 1998: 73-84.

[16] Jorg Sander, Martin Ester, Hans-Peter Kriegel, et al. Density-based Clustering in Spatial Databases: The Algorithm Gdbscan and Its Applications [J]. *Data Mining and Knowledge Discovery*, 1998, 2(2): 169-194.

[17] Foster Provost, Tom Fawcett. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions [A]. *Proc of 3rd Int Conf on Knowledge Discovery and Data Mining* [C]. California, 1997: 43-48.