基于计算几何的非线性可视化分类器设计

张 涛[1,2],洪文学[2]
(1.燕山大学信息科学与工程学院,河北秦皇岛 066004;2.燕山大学电气工程学院,河北秦皇岛 066004)

摘 要: 分类界面的计算是分类器设计的基本问题之一.本文以训练样本的空间表示为出发点设计了基于计算几何的区域主动生长的类界面求取方法.该方法首先对表示空间进行空间量化,并将量化后的点集按照信息表示分为基点与非基点,通过对基点区域的主动生长,使得整个表示空间任意区域均可进行类别表示,从而完成分类界面的计算过程.在具体分类器设计中,利用散点图的组合特性,将低维数据映射到多个可视空间,形成可视化组合分类器.对 UCI 数据集的分类实验表明,该分类器不但具有良好的可视化特性,而且分类性能已经达到或超过主流分类器的水平.

关键词: 计算几何; 主动生长; 分类界面; 可视化; 组合分类器
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2011) 01-0053-06

A Nonlinear Visual Classifier Based on Computational Geometry

ZHANG Tao[1,2], HONG Wen-xue[2]
(1. *College of Information Engineering; Yanshan University; Qinhuangdao; Hebei* 066004; *China*;
2. *College of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei* 066004; *China*)

Abstract: One of the basic issues in pattern recognition is to calculate the boundary between different categories. In this paper, we propose a novel method for that based on computational geometry named active expansion. At first, we quantize the description space. And then term the set as base and non-base points according their distribution, by active expanding for base points, any point in the whole space could express the category information and the boundary is obtained. Using this method, we design the scatter classifier which incorporates the active expansion with combining feature attribute of scatter plot, that mapping the data from low dimension to high dimension and conforming a visual combing classifier. The experiments against UCI datasets show that performance of the novel classifier has been equivalent to the popular classifiers, and outweigh in some dataset.

Key words: computational geometry; active expansion; boundary; visualization; combining classifier

1 引言

分类器设计是模式识别中的一个重要问题[1],而设计分类器的最终目的是在多维空间中找到可以将不同类别区分开的分类界面.经典的分类界面计算方法[1~5]有:线性的、曲线的、非参数估计与支持向量等方法.

对于线性可分数据,以线性判别分析(Linear Discriminant Analysis,LDA)为代表的线性分类器最为有效,其试图在高维空间中利用线性分类界面将不同类别分开.但是在涉及到生物信息学等复杂的多元数据集中,不同类别的数据点之间相互交叉,此时线性分类器无法很好的工作.因此需要使用基于核方法的比如 QDA (Quadratic Discriminant Analysis) 及 SVM (Support Vector Machine) 等非线性分类器.核方法的根本思想为 Cover 定理,将数据映射到高维空间并利用线性分类器进行分类.另一种常用的分类方法基于测量的思想,比如 k 近邻(k-Nearest Neighborhood, kNN),其通过判断未知样本与训练样本之间的距离获得类别标签,在本质上属于 voronoi 图的在模式识别中的应用[6].显然,线性分类器物理意义简单,但是不适用于样本混叠的数据分布;核方法在实践中取得了较好的分类结果,但其对于多类问题的分类比较麻烦,且难以实现分类过程可视化;kNN 方法具有良好的解释特性,但其在分类阶段计算复杂,不利于实时的信息处理,且分类精度受近邻数目影响较大.

对于分类器的设计,其实质是在特定的空间内通过对采样点的分析获得原始类别区域的过程,该过程可以理解为在特定空间内求几何点到几何区域的反问题.因此,本文提出利用计算几何进行分类界面的计算.同时,受到已知信息不充分的限制,反问题的求解往往非常困难,应保留充分的人机交互能力,以便于对成功分类

收稿日期:2009-10-19;修回日期:2010-01-26;
基金项目:国家自然科学基金(No.60904100)

数据的知识发现和对不成功分类数据的知识集成.因此,本文所提算法要求具有良好的解释性和可视化特性.基于此,本文提出一种基于计算几何思想的,物理意义简单、便于理解、分类效果良好且分类实时性较好的新的分类界面生成方法.该方法基于训练样本数据在特定空间的直观分布,通过主动生长获得一种新的非线性分类界面.本文利用该方法进一步设计了散点图分类器,对不同特征组合获得多个分类结果,再通过交互式或组合分类器[7]进行决策级融合,获得最终的分类结果.

2 基于计算几何的主动生长分类界面形成

在有监督分类中,分类界面通过对已知类别的训练样本进行分析与计算获得.设有 m 个训练样本 $X=\{x_1,x_2,\cdots,x_m\}$,每个样本具有 d 个属性,即 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}\in R^d$.设训练样本的类别集合为 $C=\{c_1,c_2,\cdots,c_k\}$,一般情况下,$k\leqslant m$.每个训练样本所对应的类别标签 $L(x_i)\in C$.通常,每个训练样本对应于表示空间的一个点.点与空间的关系表示的是类别样本在表示空间中的几何分布,本文即以表示样本的点为分类依据,利用计算几何的思想,提出主动生长的分类界面计算方法.为了描述简单,本文先分析一维数据,即 $d=1$ 时的情况,再进行高维扩展.

2.1 当 $d=1$ 时

2.1.1 空间量化

为了减少由于训练样本中离群点所引起的过学习问题并降低计算复杂度,本文引入量化方法对数据进行量化.对于样本集合 X 中某一属性集合 $X_j=\{x_{ij}\mid i=1,2,\cdots m\}$,其分布范围为$[\min X_j,\max X_j]$,其中

$$\max X_j=\{x_{ij}\mid i=\arg\max_a x_{aj}\} \tag{1}$$

$$\min X_j=\{x_{ij}\mid i=\arg\min_a x_{aj}\} \tag{2}$$

将该分布范围进行 N 阶的线性量化,则量化阶为 $\tau=\dfrac{\max X_j-\min X_j}{N}$.量化后的数据空间集合为

$$V=\{v_n\mid n=1,2,\cdots,N\} \tag{3}$$

其中,$v_n=\{x_{ij}\mid x_{ij}\in[(n-1)\tau,n\tau]\}$.本节仅考虑 1 维的情况,因此 $j=1$,以下以 x_i 代替 x_{ij}.

若当 $v_n\neq\Phi$,即

$$\sum_{j=1}^{k}\#\{x_i\mid x_i\in v_n\}>0 \tag{4}$$

其中,$\#\{\cdot\}$表示计算符合条件的样本个数.此时通过 v_n 可以用于表示样本在该量化区间内的概率分布.当 $v_n\neq\Phi$ 时,称 v_n 为基点(base point),基点的集合为 $B=\{v_n\mid v_n\neq\Phi\}$.对于任一基点 v_n,该区域内样本为 c_i 类的概率为

$$p(c_j\mid v_n)=\frac{\#\{i\mid x_i\in v_n,L(x_i)=c_j\}}{\sum_{p=1}^{k}\#\{i\mid x_i\in v_n,L(x_i)=c_p\}} \tag{5}$$

令 $p_n(c_j)=p(c_j\mid v_n)$,$p_n=\{p_n(c_j)\mid j=1,2,\cdots,k\}$.所有 v_n 对应的类别概率分布集合为 $P=\{p_n\mid n=1,2,\cdots,N\}$.

由集合 B 的定义可知,B 为 V 的子集.因此 V 与 B 的差集 $\bar{B}=V\setminus B=\{v_n\mid v_n\in V,v_n\notin B\}=\{v_n\mid v_n=\Phi\}$.在 $\bar{B}$ 中,$v_n=\Phi$,表示该点当前无类别信息,称为非基点(non-base point).非基点需要通过对基点的主动生长确定其类别分布概率.

2.1.2 基点的主动生长

由于非基点无法表示当前区域类别分布信息,需要对基点信息进行主动生长.在单方向上,生长方法依基点 v_n 与其邻域 v_{n+1}和次邻域 v_{n+2}的分布关系而定.当 v_n 为基点时,三者之间的可能关系如图 1 所示.图中,以实心点表示基点,空心点表示非基点.

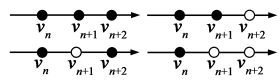

图1 基点v_n与其邻域、次邻域关系

由图 1 可知,对于基点及其邻域、次邻域的关系,共分为四种情况.而根据邻域点的类别,可分为邻域为基点与邻域为非基点两大类.

①当邻域为基点时,$v_{n+1}\in B$,该点可以表示对应的类别分布.因此无需对其进行处理.

②当邻域为非基点时,$v_{n+1}\in\bar{B}$,此时需要根据 v_n 与v_{n+2}的关系对该点进行处理.当 $v_{n+2}\in\bar{B}$ 时,v_{n+1}处的信息完全由 v_n 生长得到,因此

$$p_{n+1}(c_j)=p_n(c_j) \tag{6}$$

而当 $v_{n+2}\in B$ 时,v_{n+1}处的信息需要对 v_n 与v_{n+2}进行信息的融合,有

$$p_{n+1}(c_j)=\frac{p_n(c_j)+p_{n+2}(c_j)}{\sum_{j=1}^{k}[p_n(c_j)+p_{n+2}(c_j)]} \tag{7}$$

显然,当 v_n 为基点时,无论其邻域 v_{n+1}初始是否为基点,在一次生长后 v_{n+1}必然具有表示类别分布的功能,即成为了新的基点.通过主动生长,使得 $B\to V$,而$\bar{B}\to\Phi$.若干次生长后,必然会达到 $\bar{B}=\Phi$,此时空间内任意点均可表示类别概率分布信息,从而获得了该表示空间的分类界面.

2.2 主动生长的等效算法

主动生长是指由基点出发,向空间中的非基点生长,并最终令整个空间的点全部成为基点的过程.从另一个角度讲,主动生长是一个令非基点集合由非空到空的过程.由非基点出发进行算法设计,仅需考虑邻域情况而省略了次邻域,因此,在一定程度上减少了计算复杂度.非基点的邻域可能分布如图 2 所示.

由图 2 分析可知,当 v_n 为非基点时,其邻域状况可

以分为三类情况:

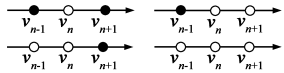

图2　当v_n为非基点时的可能分布情况

①v_{n-1}与 v_{n+1}有一个为非基点,而另一个为基点,即 $p_{n-1}\cdot p_{n+1}=0$ 且 $p_{n-1}+p_{n+1}\neq 0$.该情况下,当前的非基点 v_n 将完全依据基点域进行生长.

②v_{n-1}与 v_{n+1}均为非基点,即 $p_{n-1}\cdot p_{n+1}=0$ 且 $p_{n-1}+p_{n+1}=0$ 此时 v_n 不生长,即 $p_n=0$.

③v_{n-1}与 v_{n+1}均为基点,即 $p_{n-1}\cdot p_{n+1}\neq 0$.此时 v_n 为生长结果实际上是两个邻域融合的过程.

根据三种不同的情况,对当前点的一次生长结果如式(8)所示.

$$p_n=\begin{cases}0, & p_{n-1}\cdot p_{n+1}=0,\ p_{n-1}+p_{n+1}=0\\ p_{n-1}+p_{n+1}, & p_{n-1}\cdot p_{n+1}=0,\ p_{n-1}+p_{n+1}\neq 0\\ \dfrac{1}{2}(p_{n-1}+p_{n+1}), & p_{n-1}\cdot p_{n+1}\neq 0\end{cases} \tag{8}$$

由于当 $p_{n-1}\cdot p_{n+1}=0$ 且 $p_{n-1}+p_{n+1}=0$ 时,$p_{n-1}+p_{n+1}=0$.因此式(8)可简化为

$$p_n=\begin{cases}p_{n-1}+p_{n+1}, & p_{n-1}\cdot p_{n+1}=0\\ \dfrac{1}{2}(p_{n-1}+p_{n+1}), & p_{n-1}\cdot p_{n+1}\neq 0\end{cases} \tag{9}$$

式(9)可进一步简化为

$$p_n=\frac{1}{2}\left|\operatorname{sign}\left|p_{n-1}\cdot p_{n+1}\right|-2\right|\cdot(p_{n-1}+p_{n+1}) \tag{10}$$

其中 sign(·)为符号函数.根据以上分析,主动生长过程可以用伪代码 ActiveExpansion 表示:

```
procedure ActiveExpansion is
begin
    计算非基点集合 B̄
    loop while B̄ ≠ Φ
        对每一个 v_n ∈ B̄
        p_n(c_j) = expansion(p_{n-1}, p_{n+1})
        计算非基点集合 B̄
    end loop
end ActiveExpansion
```

其中,函数 expansion 的伪代码为

```
procedure expansion is
interface p_n(c_j) = expansion(p_{n-1}, p_{n+1})
begin
j = 1
loop
    p(c_j) = 1/2 | sign | p_{n-1}(c_j)·p_{n+1}(c_j) | - 2 |
    ·(p_{n-1}(c_j) + p_{n+1}(c_j))
    j = j + 1
    exit when j = k
    p_n(c_j) = p(c_j) / Σ_{j=1}^{k} p_n(c_j)
end expansion
```

2.3　高维情况

对于 $d>1$ 的数据,可以在高维空间中对各维独立生长获得高维下的边界.设第 l 维数据的第 n 个数据为基点,其生长后的邻域为 $p_{n+1}^{l}=T_l(v_{n+1})$,则综合各维生长,其结果为

$$p_{n+1}^{(1,2,\cdots,d)}=\frac{1}{d}\sum_{l=1}^{d}p_{n+1}^{l}=\frac{1}{d}\sum_{l=1}^{d}T_l(v_{n+1}) \tag{11}$$

显然,对于 $d>3$ 的高维空间,虽然在数学可以进行生长,但无法进行直观的可视化,进而无法完成交互式的分类过程.虽然 3 维空间对于人类视觉可视,但考虑到现阶段的显示器件均为 2 维显示,为了保证分类过程的可视性与可交互性,在此不对 3 维的空间分割进行讨论.因此可将 d 维空间分割为 $n_d=\binom{d}{2}$ 个二维子空间集合.将未知类别的 d 维数据在该集合空间上投影,可以在不同的二维空间分别进行分类结果的判断;若将各分类结果进行融合,可以达到组合分类的目的.

3　分类器实现

上一节描述了基于主动生长思想的分类界面计算过程,该思想可以应用于任意数据分布空间.在各种数据描述与表示中,多元图表示方法中的散点图表示最为直观并易于理解[8],因此本文以散点图为数据表示空间,通过对散点图的生长获得分类性能,形成散点图分类器.其具体的实现过程如图 3 所示[9].

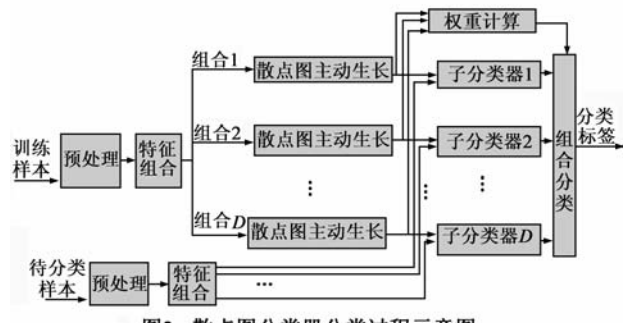

图3　散点图分类器分类过程示意图

3.1　预处理

在量化阶段,首先要得到不同样本间同一属性的最大值与最小值.但通常情况下,对于 $j\neq k$ 有 $[\min X_j,\max X_j]\neq[\min X_k,\max X_k]$.为了简化后期主动生长阶段的范围,可以对数据进行归一化预处理.这样的处理方式符合的 Duin 提出的数据相对描述与概念描述[10].归一化处理后

$$Norm(x_{ij})=\frac{x_{ij}-\min X_j}{\max X_j-\min X_j}\in[0,1] \tag{12}$$

在预处理过程,为了改善样本在空间中的分布特性,还可以进行非线性优化.其优化规则与分析参见文献[11].由于分类器的设计与优化并非本文的重点内容,

在此仅利用较为简单的多项式方法进行优化,优化后

$$x_{ij}^{*} = Opt((Norm)x_{ij}) = (Norm(x_{ij}))^{a} \quad (13)$$

3.2 特征组合与主动生长

为了对数据进行分类,需要将其在特定的空间进行表示,好的表示是良好分类的基础[10].基于表示的直观性,本文采用二维散点图进行特征表示.二维散点图是数据可视化中最常用的方法之一.它可以显示两个变量之间的关系.每个数据样本对应一个点或者标记,其位置坐标由两个变量的值决定.通过散点图可以观察和理解聚类,离群点,趋势以及相关等数据结构信息.散点图还有一个优点就是可以同时表达更多的样本.二维数据的平面散点图实际上就是以二维数据变量为坐标在平面直角坐标系中描点表示.对于完成列归一化的数据矩阵$[x_{ij}^{*}]$,其对应的散点图坐标为

$$\begin{cases} x = x_{ij}^{*} \\ y = x_{ik}^{*} \end{cases} \quad (14)$$

通过散点图对特征进行组合,对原始数据进行了升维操作.将 d 个属性的数据组成 n_d 个特征对,每个特征对在一个二维空间中进行生长.由于二维空间具有可视化的特点,因此整个生长过程是可视的.特征组合与生长过程如图4所示.图中,RGB三基色分别表示三个不同的类别,而混合色表示概率不为1的不确定点.

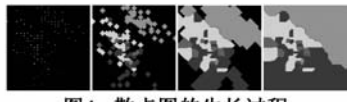

图4 散点图的生长过程

3.3 分类器的形成与分类过程

d 维空间形成的分类界面无法进行直观的观察,因此可将其分割为 n_d 个二维空间,形成 n_d 个基分类器.在对未知样本分类过程中,可以通过基分类器组合来获得分类性能.该分类过程既可以通过直观的交互式完成,也可以通过分类器自动完成.

在散点图分类器中,主动生长之后的空间颜色可以表示当前映射点的类别信息.因此对于任意未知类别的样本,仅通过对比其散点图坐标在对应分类空间上的颜色即可判断其所属类别及概率,可以通过人类视觉对类别进行直接的判断,而无需后续的模式识别过程.例如 Iris 数据集 petal length 与 petal width 特征形成子分类空间如图5所示,对于未知样本 s1,其对应颜色为蓝色,因此可判断为 virginica 类(蓝色表示);而对于未知样本 s2,其对应的颜色为青色偏绿,属于绿色与蓝色按 2∶1 的比例混合得到,因此其属于 versicolor 类(绿色表示)的概率为67.7%,而属于 virginica 类的概率为 33.3%.交互的分类过程可以很好的集成人类的识别知识,并有利于新知识的发现.

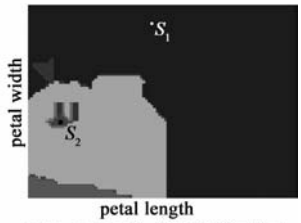

图5 Iris子分类器分类空间

另外,该分类器也可以利用自动分类方法完成.设未知样本为 $u = \{u_1, u_2, \cdots, u_d\}$,则其属于 c_j 的概率为

$$p(c_j \mid u) = \sum_{\substack{i1\in[1,d] \\ i2\in(i1,d]}} \alpha_{(i1,i2)} p(c_j \mid u_{i1}, u_{i2}) \quad (15)$$

其中,$\alpha_{(i1,i2)}$ 表示在($i1$, $i2$)属性组合情况下形成的基分类器在组合分类过程中的加权.本文的加权方法采用混色比例加权方法,即对于某个特定的空间,根据生长后空间内基色与混合色的比例确定权重.对于某个特定空间,其权重为:

$$\alpha = \frac{\#\{p_n(c_j) = 1 \mid n \in [1, N^2], j \in [1, k]\}}{N^2 - \#\{p_n(c_j) = 1 \mid n \in [1, N^2], j \in [1, k]\}} \quad (16)$$

最终的分类标签为

$$c = \arg\max_{j}(c_j \mid p(c_j \mid u)) \quad (17)$$

4 实验设计与结果分析

4.1 分类实验

为了测试散点图分类器的性能,本文利用 UCI 数据库(http://archive.ics.uci.edu/ml/)中的多个数据集对分类器进行测试.所选择的数据集及其属性如表1所示.其中,前两个数据集(Pima-Indians-diabetes 与 liver-disorder)是简单的二分类问题;接下来的三个数据集(Iris,wine 和 glass)可以用于测试针对中等分类复杂度下的多类别数据的分类性能,也是对分类器进行性能测试中最常用的数据集.其中,Iris 与 wine 数据集为模式分类的常用测试集,分别测试低维特征与高维特征情况下的多类分类器的分类性能,而 glass 数据集中,特征维数相对较高,且类别数目较高,可以表现多类分类性能;而最后两个数据集(ionosphere 与 breast cancer)数据维数均超过 30 维,数据维数相对较高,用于评价高维情况下的分类情况.实验中用到的这些数据集来自物理科学与生命科学领域,具体的应用包括了疾病诊断、产品分类、物种识别、气象分析等,且均为实际测量的实验数据,含有一定的测量误差,因此可以在一定程度上代表分类器在实际应用中的分类性能[12,13].

表1 实验用到的各数据集属性

数据集	所属领域	特征数	类别数	样本数
Pima-Indians-diabetes	生命科学	8	2	768
liver-disorder	生命科学	6	2	345
Iris	生命科学	4	3	150
wine	物理科学	13	3	178
glass	物理科学	9	6	214
ionosphere	物理科学	34	2	351
breast cancer	生命科学	30	2	569

在实验过程中,为了确保分类性能更为客观,并避

免训练集和测试集的依赖,分类器精度的估计采用留一法交叉验证(leave one out cross validation,LOOCV).留一法是指设数据集共有 N 个样本,使用($N-1$)个样本设计分类器,并估计剩余的一个样本;对于训练集重复 N 次.这种估计虽然计算量大,但是无偏的.

为了获得分类性能的直观认识,本文利用 LDA,QDA,kNN,parzen 窗,SVM 等经典分类器[1]作为对比测试方法.其中,LDA 为典型的基于参数估计的线性分类器,而 QDA 为典型的基于参数估计的非线性分类器;kNN 和 parzen 窗方法为总体分布的非参数估计方法,其中,kNN 为基于测量的分类方法,而 parzen 窗分类器则是利用了对空间的窗口分析.SVM 作为统计学习理论和核方法的典型代表,得到了广泛的认可.在本实验中,为了保证测试结果的客观性,参考分类器均采用 PRTools 中的软件包完成.实验结果如表 2 所示,对于某一数据集,最佳的分类精度用粗体标出.

4.2　结果分析与讨论

以上实验结果表明,对于低维度二分类问题,本文构造的散点图分类器在分类精度上并不具有优势,其分类性能与经典分类器相当.对于中等复杂度的分类问题,散点图分类器的分类精度则已经那个达到甚至超过主流分类器水平;而对于高维分类问题,散点图分类器的分类精度全面超越了目前的主流分类器,达到了较好的分类效果.从对 7 个 UCI 数据的综合实验结果来看,散点图分类器在 Iris、glass 等 4 个数据集的分类精度达到了最高水平,而其他几个数据集的分类精度也与经典分类器相当.因此,该实验充分证明了基于主动生长思想的散点图分类器具有良好的泛化能力,已经达到了主流分类器的水平,也说明了主动生长思想所形成的分类界面的分类性能.

表 2　不同分类器分类性能对比

数据集	散点图分类器	LDA	QDA	kNN (k = 1)	Parzen 窗	SVM
Pima-Indians-diabetes	75.13	**77.47**	73.96	76.17	75.26	77.08
liver-disorder	65.80	69.86	59.42	68.41	63.19	**70.15**
Iris	**98.67**	98.00	97.33	96.67	96.67	96.00
wine	98.31	98.88	**99.44**	76.97	71.35	96.07
glass	**76.17**	64.49	57.01	73.36	67.29	56.08
ionosphere	**90.32**	88.50	85.94	83.71	86.26	88.82
breast cancer	**96.13**	95.78	95.61	92.09	92.44	95.43

借助 PRTools 软件包,图 6 列出了不同分类器针对 Iris 数据集中 sepal length 和 sepal width 特征的分类界面.借此可以分析散点图分类器获得较好分类性能的原因.

首先,由于散点图分类器采用了主动生长方式计算分类界面,分界线由所有样本点主动生长获得,将空间划分为多个区域,不要求相同类别区域连通,且允许模糊区域存在.因此最终形成的分类界面与 LDA、QDA 等指定函数形式的分类界面不同,所获得的分类界面为非线性自由界面,不受函数形式的约束.其次,散点图分类器在分类过程中通过特征组合进行升维,且升维后仍进行非线性分类,因此对于复杂分类问题,其性能优于单纯的非线性分类 QDA 与高维线性分类的 SVM.第三,与非参数估计分类方法相比,kNN 在分类前要指定 k 的大小,但未知数据可能落在任意区域,其近邻数目不确定,因此会出现偏差.散点图分类器则通过区域生长使得空间任意坐标下的类别分布取决于周边所有样本的类别分布,因此可以理解为具有自适应的 k 值选择,并且通过量化来减少噪声数据的影响.第四,散点图分类器利用了组合分类方法,最大程度上利用了训练样本提供的信息,也有助于分类精度的提高.

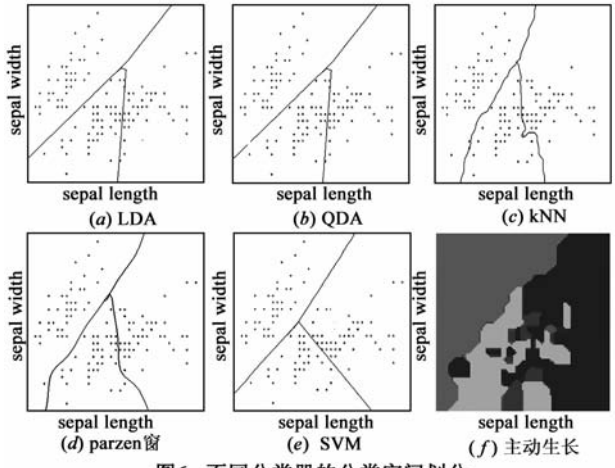

图6　不同分类器的分类空间划分

另外,散点图分类器具有良好的可视化特性和可解释特性,为利用该分类器进行知识发现打下了良好的基础.

5　结论

针对表示空间中分类界面的形成,本文提出一种新的分类界面计算方法.该方法通过对训练样本在表示空间进行主动区域生长获得数据空间的划分,获得非线性的模糊分类表示,提高分类过程的可解释性和可交互性.通过 UCI 数据集的实验分析表明,利用该方法形成的散点图分类器在分类精度上与经典分类器达到了同一级别,具有很好的学术价值和应用前景.

但目前该分类器也存在这一定的问题,比如训练阶段计算复杂度偏高,对低维度二分类数据分类不理想等,这些问题将在今后的研究中改进.如何找到最适合于主动生长的数据表示方法,也是进一步研究的方向.

参考文献:

[1] R O Duda, P E Hart, D G Stork. Pattern Classification[M]. New York: Wiley, 2000.

[2] G J McLachlan. Discriminant Analysis and Statistical Pattern Recognition[M]. New York: Wiley Interscience, 2004.

[3] V Vapnik. Statistical Learning Theory[M]. New York: Wiley Interscience, 1998.

[4] Anil K Jain, Robert P W Duin, Jianchang Mao. Statistical pattern recognition: A review[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4 – 37.

[5] Y Tominaga. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN[J]. Chemometrics and Intelligent Laboratory Systems, 1999, 49(1): 105 – 115.

[6] Maurizio Filippone, Francesco Camastra, Francesco Masulli, Stefano Rovett. Asurvey of kernel and spectral methods for clustering[J]. Pattern Recognition, 2008, 41(1): 176 – 190.

[7] Nikunj C Oza, Kagan Tumer. Classifier ensembles: Select real-world applications[J]. Information Fusion, 2008, 9(1): 4 – 20.

[8] 洪文学,李昕,徐永红. 基于多元图表示原理的信息融合与模式识别技术[M]. 北京: 国防工业出版社, 2008. 95 – 99.
Hong Wenxue, Li Xin, Xu Yonghong. Information Fusion And Pattern Recognition Based On Graphical Representation Theory [M]. Beijing: National defence industry press, 2008. 95 – 99. (in Chinese)

[9] Zhang Tao, Hong Wenxue. A novel visual combining classifier based on a two-dimensional graphical representation of the attribute data[A]. Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery[C]. IEEE Press, 2009. 71 – 75.

[10] Elzbieta PeRkalska, Robert P W Duin, Pavel Paclik. Prototype selection for dissimilarity-based classifiers[J]. Pattern Recognition, 2006, 39(2): 189 – 208.

[11] 张涛,洪文学,宋佳霖,常凤香. 基于非线性变换的图表示优化[J]. 燕山大学学报, 2008, 32(5): 416 – 420.
Zhang Tao, Hong Wenxue, Song Jia-lin, Chang Fengxiang. The adaptation for graphical representation based on nonlinear transformation[J]. Journal of Yanshan University, 2008, 32(5): 416 – 420. (in Chinese)

[12] 李洁,邓一鸣,沈士团. 基于模糊区域分布的分类规则提取及推理算法[J]. 计算机学报, 2008, 31(6): 934 – 941.
Li Jie, Deng Yiming, Shen Shituan. Classification rule extraction based on fuzzy area distribution and classification reasoning algorithm[J]. Chinese Journal of Computers, 2008, 31(6): 934 – 941. (in Chinese)

[13] Enwang Zhou, Alireza Khotanzad. Fuzzy classifier design using genetic algorithms[J]. Pattern Recognition, 2007, 40(12): 3401 – 3414.

作者简介:

张 涛 男, 1979 年 3 月生于河北省唐山市. 2003 年毕业于燕山大学通信与信息系统专业. 现为燕山大学讲师. 主要从事可视化模式识别, 图像处理等领域的研究.
E-mail: zhtao_79@163.com

洪文学 男, 教授、博士生导师. 1953 年 5 月出生于黑龙江依安县. 1983 年毕业于哈工大电测技术与信息处理仪器专业. 现为燕山大学生物医学工程研究所所长, 主要从事可视化信息融合、可视化模式识别与疾病诊断和中医工程学等方面研究工作.
E-mail: hongwx@ysu.edu.cn