第4章 模式识别—
4.4其他分类器

信控学院 蔡利梅

4.4.1组合分类器

（1）基本概念

- 构建一组单独的分类器（个体），整合结果，以获得更好的性能。
- 个体分类器为同一种称为同质，反之称为异质
- 要求：**多样性**，不同个体分类器间的分类结果具有差异性；**准确性**，个体分类器具有较好的分类性能
- 性能评价：
  - 泛化误差：误分类概率
  - 计算复杂度等

（2）Bagging算法

- **Bootstrap Aggregating，多次采样同一数据集得到多组数据，分别进行训练得到若干弱分类器，再通过对弱分类器结果投票得到强分类器**
- **特点：并行**

例：有10个考试成绩，用三个最小距离分类器设计Bagging组合分类器。

序号	1	2	3	4	5
成绩	(10,70)	(20,70)	(30,10)	(40,60)	(60,50)
全通过	否	否	否	否	否

序号	1	2	3	4	5
成绩	(60,80)	(70,90)	(80,70)	(90,80)	(100,60)
全通过	是	是	是	是	是

- 设计一
  - 抽样

序号	1	2	4	1	3	4
成绩	(10,70)	(20,70)	(40,60)	(60,80)	(80,70)	(90,80)
全过	否	否	否	是	是	是

  - 设计最小距离分类器一

$$\mu_1 = (23.3, 66.7)$$

$$\mu_2 = (76.7, 76.7)$$

$$y_1 = \begin{cases} -1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 < 0 \\ 1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 \geq 0 \end{cases}$$

  - 决策

$$y_1 = (-1, -1, -1, -1, 1, 1, 1, 1, 1, 1)$$

设计二

抽样

序号	4	4	2	1	3	4
成绩	（40,60）	（40,60）	（20,70）	（60,80）	（80,70）	（90,80）
全通过	否	否	否	是	是	是

设计最小距离分类器二

$$\mu_1 = (33.3, 63.3)$$
$$\mu_2 = (76.7, 76.7)$$

$$y_2 = \begin{cases} -1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 < 0 \\ 1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 \geq 0 \end{cases}$$

决策 $y_2 = (-1, -1, -1, -1, -1, 1, 1, 1, 1, 1)$

设计三

- **抽样**

序号	2	5	1	3	5	5
成绩	(20,70)	(60,50)	(10,70)	(80,70)	(100,60)	(100,60)
全过	否	否	否	是	是	是

- **设计最小距离分类器三**

$$\mu_1 = (30,63.3)$$

$$\mu_2 = (93.3,63.3)$$

$$y_3 = \begin{cases} -1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 < 0 \\ 1 & \|X-\mu_1\|^2 - \|X-\mu_2\|^2 \geq 0 \end{cases}$$

- **决策** $y_3 = (-1,-1,-1,-1,-1,-1,1,1,1,1)$

□ 投票表决

序号	1	2	3	4	5
成绩	(10,70)	(20,70)	(30,10)	(40,60)	(60,50)
全通过	否	否	否	否	否
y_1	-1	-1	-1	-1	1
y_2	-1	-1	-1	-1	-1
y_3	-1	-1	-1	-1	-1
投票	-1	-1	-1	-1	-1

□ 投票表决

序号	1	2	3	4	5
成绩	(60,80)	(70,90)	(80,70)	(90,80)	(100,60)
全通过	是	是	是	是	是
y_1	1	1	1	1	1
y_2	1	1	1	1	1
y_3	-1	1	1	1	1
投票	1	1	1	1	1

（3）Boosting算法

融合多个分类器进行决策的方法；不是简单地对多个分类器的输出进行投票决策，而是通过一个迭代过程对分类器的输入和输出进行**加权**处理。

- AdaBoost算法 □ 基本思路

◆ 训练样本$\{X_1, X_2, \cdots, X_N\}$，M个弱分类器在样本X上的输出：$g_m(X) \in \{-1, 1\}(m = 1, \cdots, M)$。

◆ 给数据和弱分类器以权值；在迭代中，如果上一轮的计算中某一数据分类正确，其对应**权重相应减少**，**反之增加**；计算弱分类器的分类正确率，正确率提高，**加大该分类器权重**，反之减少。

□ 算法步骤

◆ 初始化　　训练样本$\{X_1, X_2, \cdots, X_N\}$的权重$\beta_i = \frac{1}{N}$

◆ 迭代$(m = 1, \cdots, M)$

分类器目标函数中各样本对应的项进行加权，具体问题具体分析

1）将训练样本加权后构造分类器$f_m(X) \in \{-1, 1\}$；

2）计算分类器的错误率e_m，并确定该分类器在组合分类器中的权重：如果$e_m < 0.5, \alpha_m = \frac{1}{2} ln\left(\frac{1-e_m}{e_m}\right)$

3）修改样本权重$\beta_i = \frac{\beta_i exp\left(-\alpha_m y_i g_m(X_i)\right)}{2\sqrt{e_m(1-e_m)}}$

4）待分类样本X，强分类器输出为$sgn\left(\sum_{m=1}^{M} \alpha_m g_m(X)\right)$

例：有10个考试成绩，弱分类器采用最小距离分类器，采用Adaboost算法设计组合分类器。

序号	1	2	3	4	5
成绩	(10,70)	(50,70)	(30,10)	(40,50)	(70,50)
全通过	否	否	否	否	否

序号	1	2	3	4	5
成绩	(60,60)	(70,90)	(80,70)	(90,80)	(100,60)
全通过	是	是	是	是	是

- 采用最小距离分类器初始化进行分类

$\mu_1 = (40, 50)$ $\quad g_0 = (-1, -1, -1, -1, 1, -1, 1, 1, 1, 1)$

$\mu_2 = (80, 72)$ $\quad$ 正确率：80%

- 初始化样本权值：$\beta_i = \frac{1}{10}, i = 1, \cdots, 10$

- 设计最小距离分类器一

$$\mu_1 = \sum_{X_i \in \omega_1} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_1} \beta_i = \begin{pmatrix} 40 & 50 \end{pmatrix}$$

$$\mu_2 = \sum_{X_i \in \omega_2} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_2} \beta_i = \begin{pmatrix} 80 & 72 \end{pmatrix}$$

  - 决策 $\quad g_1 = (-1, -1, -1, -1, 1, -1, 1, 1, 1, 1)$
  - 错误率 $\quad e_1 = 2/10 = 0.2$

- 修改权系数 $\alpha_1 = \frac{1}{2} ln\left(\frac{1-e_1}{e_1}\right) = 0.6931$

$$\beta = \begin{pmatrix} 0.0625 & 0.0625 & 0.0625 & 0.0625 & 0.2500 \\ 0.2500 & 0.0625 & 0.0625 & 0.0625 & 0.0625 \end{pmatrix}$$

- 设计最小距离分类器二

$$\mu_1 = \sum_{X_i \in \omega_1} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_1} \beta_i = (51.25 \quad 50.00)$$

$$\mu_2 = \sum_{X_i \in \omega_2} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_2} \beta_i = (72.50 \quad 67.50)$$

- 决策 $g_2 = (-1, -1, -1, -1, 1, -1, 1, 1, 1, 1)$

- 错误率 $e_2 = 2/10 = 0.2$

◆ 修改权系数 $\alpha_2 = \frac{1}{2} ln\left(\frac{1-e_2}{e_2}\right) = 0.6931$

$$\beta = \begin{pmatrix} 0.0391 & 0.0391 & 0.0391 & 0.0391 & 0.6250 \\ 0.6250 & 0.0391 & 0.0391 & 0.0391 & 0.0391 \end{pmatrix}$$

□ 设计最小距离分类器三

$$\mu_1 = \sum_{X_i \in \omega_1} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_1} \beta_i = (62.50 \quad 50.00)$$

$$\mu_2 = \sum_{X_i \in \omega_2} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_2} \beta_i = (65.00 \quad 63.00)$$

◆ 决策 $g_3 = (1, 1, -1, -1, -1, 1, 1, 1, 1, 1)$

◆ 错误率 $e_3 = 2/10 = 0.2$

- 修改权系数 $\alpha_3 = \frac{1}{2} ln\left(\frac{1-e_3}{e_3}\right) = 0.6931$

$$\beta = \begin{pmatrix} 0.0977 & 0.0977 & 0.0244 & 0.0244 & 0.3906 \\ 0.3906 & 0.0244 & 0.0244 & 0.0244 & 0.0244 \end{pmatrix}$$

- 设计最小距离分类器四

$$\mu_1 = \sum_{X_i \in \omega_1} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_1} \beta_i = (55.00 \quad 54.61)$$

$$\mu_2 = \sum_{X_i \in \omega_2} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_2} \beta_i = (65.00 \quad 63.00)$$

- 决策 $g_4 = (-1, -1, -1, -1, 1, 1, 1, 1, 1, 1)$

- 错误率 $e_4 = 1/10 = 0.1$

- 修改权系数 $\alpha_4 = \frac{1}{2} ln\left(\frac{1-e_4}{e_4}\right) = 1.0986$

$$\beta = \begin{pmatrix} 0.0543 & 0.0543 & 0.0136 & 0.0136 & 1.9531 \\ 0.2170 & 0.0136 & 0.0136 & 0.0136 & 0.0136 \end{pmatrix}$$

- 设计最小距离分类器五

$$\mu_1 = \sum_{X_i \in \omega_1} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_1} \beta_i = (67.46 \quad 50.78)$$

$$\mu_2 = \sum_{X_i \in \omega_2} \beta_i \times X_i \Big/ \sum_{X_i \in \omega_2} \beta_i = (65.00 \quad 63.00)$$

- 决策 $g_5 = (1, 1, -1, -1, -1, 1, 1, 1, 1, -1)$

- 错误率 $e_5 = 3/10 = 0.3$

- 修改权系数 $\alpha_5 = \frac{1}{2} ln\left(\frac{1-e_5}{e_5}\right) = 0.4236$

$$\beta = \begin{pmatrix} 0.0904 & 0.0904 & 0.0097 & 0.0097 & 1.3951 \\ 0.1550 & 0.0097 & 0.0097 & 0.0097 & 0.0226 \end{pmatrix}$$

- 组合分类器$\sum_{m=1}^{M} \alpha_m g_m(X) =$

$$\begin{pmatrix} -1.3681 & -1.3681 & -3.6017 & -3.6017 & 1.3681 \\ 0.8291 & 3.6017 & 3.6017 & 3.6017 & 2.7544 \end{pmatrix}$$

- 决策 $g = (-1, -1, -1, -1, 1, 1, 1, 1, 1, 1)$

正确率：90%

4.4.2半监督学习

（1）基本概念

- 问题的提出
  - 有小部分有标记的样本（已知类别），有大量未标记样本（未知类别），如何有效利用这些样本提高识别性能？
  - 将未标记样本标记：费时费力
  - 主动学习：Active Learning，用已标记样本训练，判断未标记样本，将新标记的样本加入已标记样本集，再训练，再增加样本…。需要引入额外的专家知识

半监督学习概念

Semi-Supervised Learning，SSL，让学习器不依赖外界交互，自动地利用未标记样本来提升学习性能。

前提假设

未标记样本所揭示的数据分布信息与类别标记相联系的假设：相似的样本具有相似的输出。

- 聚类假设：同一簇的样本属于同一类
- 流形假设：邻近的样本有相似的输出

- 半监督学习方法分类
  - 归纳学习：Indnctive Learning，IL，假设训练样本中的未标记样本并非待预测的数据，希望学习模型能适用于训练过程中**未观察到的数据**。更开放。
  - 直推学习：Transudative Learning，TL，假定学习过程中所考虑的未标记样本恰好是待预测数据，学习的目的就是在这些未标记样本上获得最优泛化性能，仅试图对学习过程中**观察到的未标记样本**进行预测。较封闭。

（2）生成式模型

- 前提假设

有标记和无标记样本由同一潜在模型生成的。采用不同模型得到不同的半监督学习模型，使用模型和真实数据分布相符

- 分析

给定样本X，类别标记为$y \in Y = \{1, 2, \cdots, K\}$，$K$为所有可能的类别的个数。假设样本由高斯混合模型生成，且每一个类别对应一个高斯混合成分，即$p(X) = \sum_{i=1}^{K} \alpha_i p(X|\mu_i, \Sigma_i)$。

$p(X|\mu_i, \Sigma_i)$是样本属于第i个高斯混合成分的概率，混合系数$\alpha_i \geq 0$，$\sum_{i=1}^{K} \alpha_i = 1$为选择第$i$个高斯混合成分的概率。

模型对样本的预测表示为：$f(X) \in Y$；

样本X隶属的高斯混合成分表示为：$\Theta \in \{1, 2, \cdots, K\}$；

最大后验概率法：$f(X) = \arg\max_{j \in Y} p(y = j \mid X)$

$$= \arg\max_{j \in Y} \sum_{i=1}^{K} p(y = j, \Theta = i \mid X)$$

$$= \arg\max_{j \in Y} \sum_{i=1}^{K} p(y = j \mid \Theta = i, X) \cdot p(\Theta = i \mid X)$$

需要已知样本的标记y

$$p(\Theta = i \mid X) = \frac{\alpha_i p(X \mid \mu_i, \Sigma_i)}{\sum_{k=1}^{K} \alpha_k p(X \mid \mu_k, \Sigma_k)}$$

不需要有标记样本

同时利用有标记和无标记样本

算法描述

- 给定有标记数据集$D_l=\{(X_1,y_1),(X_2,y_2),\cdots,(X_l,y_l)\}$，未标记数据集$D_u=\{(X_{l+1},y_{l+1}),(X_{l+2},y_{l+2}),\cdots,(X_{l+u},y_{l+u})\}$，且$l \ll u, l+u=N$。假设所有样本独立同分布，由同一高斯混合模型生成，**通过求解高斯混合模型的参数，进而求最大化后验概率**实现分类。
- 使用极大似然估计求解高斯混合模型参数，对数似然函数为

$$
\begin{aligned}
LL(D_l \cup D_u) = & \sum_{(X_j,y_j)\in D_l} \ln\left(\sum_{i=1}^{K} \alpha_i p(X_j \mid \mu_i,\Sigma_i) p(y_j \mid \Theta=i, X_j)\right) \\
& + \sum_{X_j \in D_u} \ln\left(\sum_{i=1}^{K} \alpha_i p(X_j \mid \mu_i,\Sigma_i)\right)
\end{aligned}
$$

EM算法求解参数

- E步：根据目前的模型各参数计算未标记样本X_j的属于各高斯混合成分的概率γ_{ji}
- M步：根据γ_{ji}更新模型参数α_i、μ_i和Σ_i

$$\mu_i = \frac{1}{\sum_{X_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{X_j \in D_u} \gamma_{ji} X_j + \sum_{(X_j, y_j) \in D_l \wedge y_j = i} X_j \right)$$

$$\alpha_i = \frac{1}{N} \left(\sum_{X_j \in D_u} \gamma_{ji} + l_i \right)$$

$$\Sigma_i = \frac{1}{\sum_{X_j \in D_u} \gamma_{ji} + l_i} \left(\begin{array}{c} \sum_{X_j \in D_u} \gamma_{ji} (X_j - \mu_i)(X_j - \mu_i)^T + \\ \sum_{(X_j, y_j) \in D_l \wedge y_j = i} (X_j - \mu_i)(X_j - \mu_i)^T \end{array} \right)$$

l_i表示第i类中有标记样本的个数

（3）半监督支持向量机

- 原理
  - Semi-Supervised Support Vector Machine,S3VM
  - 转导推理(Transductive Inference)和SVM的结合，Transductive Support Vector Machine,TSVM
  - 针对二分问题，利用已标记数据集训练SVM；用SVM对未标记样本预测标记（伪标记），并尝试对每个未标记样本分别做正例或反例；寻找对于所有样本间隔最大化的超平面；超平面对未标记样本的判别即最终预测结果。

算法描述

给定有标记数据集$D_l=\{(X_1,y_1),(X_2,y_2),\cdots,(X_l,y_l)\}$，未标记数据集$D_u=\{(X_{l+1},y_{l+1}),(X_{l+2},y_{l+2}),\cdots,(X_{l+u},y_{l+u})\}$，且$l\ll u, l+u=N$。二分问题中，标记$y\in\{-1,1\}$

1）利用D_l设计SVM，对D_u中的样本进行标记，再用所有样本求优化超平面

$$\min_{W,b,\hat{y},\xi}\frac{1}{2}\|W\|^2+C_l\sum_{i=1}^{l}\xi_i+C_u\sum_{i=l+1}^{N}\xi_i$$

$$C_u<C_l$$

$$\begin{aligned}
s.t.\quad & y_i\left(W^TX_i+b\right)\ge 1-\xi_i, i=1,2,\cdots,l\\
& \hat{y}_i\left(W^TX_i+b\right)\ge 1-\xi_i, i=l+1,l+2,\cdots,N\\
& \xi_i\ge 0, i=1,2,\cdots,N
\end{aligned}$$

2）找出两个标记指派为异类且很可能发生错误的未标记样本，交换标记，重新求优化超平面

3）重复第二步，直至没有满足条件的

4）增大C_u，重复1）2）3）直到$C_u = C_l$

5）获取最终的未标记样本的预测结果。

- 算法分析

- □ 计算量、计算复杂度十分高
- □ 需要高效优化求解策略，发展出很多方法

（4）半监督聚类

- **原理**
  - 利用少量标记样本对聚类进行辅助，称为半监督聚类，Semi-supervised clustering。
  - 基于约束的半监督聚类：利用监督信息对聚类的搜索过程进行约束
    - 约束K均值算法：Constrained k-means
    - 种子K均值算法：Seeded k-means