基于新型不纯度度量的代价敏感随机森林分类器

师彦文 王宏杰
(西南石油大学计算机科学学院 成都 610500)

摘 要 针对不平衡数据集的有效分类问题,提出一种结合代价敏感学习和随机森林算法的分类器。首先提出了一种新型不纯度度量,该度量不仅考虑了决策树的总代价,还考虑了同一节点对于不同样本的代价差异;其次,执行随机森林算法,对数据集作 K 次抽样,构建 K 个基础分类器;然后,基于提出的不纯度度量,通过分类回归树(CART)算法来构建决策树,从而形成决策树森林;最后,随机森林通过投票机制做出数据分类决策。在UCI数据库上进行实验,与传统随机森林和现有的代价敏感随机森林分类器相比,该分类器在分类精度、AUC面积和Kappa系数这3种性能度量上都具有良好的表现。

关键词 代价敏感学习,随机森林,不纯度度量,分类回归树(CART),不平衡数据
中图法分类号 TP181 **文献标识码** A

Cost-sensitive Random Forest Classifier with New Impurity Measurement

SHI Yan-wen WANG Hong-jie
(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract For the problem of effective classification on imbalanced data sets, a classifier combining cost-sensitive learning and random forest algorithm is proposed. Firstly, a new impurity measure is proposed, taking into account not only the total cost of the decision tree, but also the cost difference of the same node for different samples. Then, the random forest algorithm is executed, K times sampling for the data set is performed, and K basic classifiers are built. Then, the decision tree is constructed by the classification regression tree (CART) algorithm based on the proposed impurity measure, so as to form the decision tree forest. Finally, the random forest algorithm makes the data classification decision by voting mechanism. In the UCI database, compared with the traditional random forest and the existing cost-sensitive random forest classifier, this classifier has good performance in the classification accuracy, AUC area and Kappa coefficient.

Keywords Cost-sensitive learning, Random forest, Impurity measurement, Classification regression tree (CART), Imbalanced data

1 引言

代价敏感学习是机器学习领域内的一个重要研究方向。其对不同类型的错误分类设定不同的惩罚代价,以此来最小化分类器的总体错误分类代价[1]。对于不同数据点误分类产生的代价不同这一问题,通常有两种解决方法[2]:基于分类的代价敏感学习(Class dependent Cost-sensitive,CCS)和基于样本的代价敏感学习(Example dependent Cost-sensitive,ECS)。在CCS中,代价以代价矩阵的方式定义,所有分类错误的代价是相等的;在ECS中,代价是根据每个单独的数据点是否被误分类而计算获得的。因此,标准分类和CCS都可以看作ECS的特例。

目前,已经有多种方法用于解决基于样本的代价敏感学习问题,如根据代价对训练样本重加权[3]。基于决策树[4]的方法具有训练和测试时间短、易于部署、能够以概率形式输出、能够简单拓展到多类场景等优良特性,因此其可以适应于多种不同类型的机器学习问题,将代价敏感学习和决策树相结合具有很大的实际意义[5]。在训练阶段,将代价信息整合到决策树中的方法主要有3种:1)改变数据抽样的方式;2)改变每个节点的类分布从而降低总开销;3)特意为基于样本的代价敏感场景创建一个不纯度度量[6]。本文采用为第三种方法。

基于上述分析,提出一种结合代价敏感学习和随机森林算法的分类器,随机森林是一种集成多个决策树的分类方法。为了使算法能够更好地应对不平衡数据,基于标签向量的任务分数的成对差异,提出了一种新型不纯度度量。与传统决策树的不纯度度量相比,提出的度量不仅考虑了总代价,还考虑了同一节点对于不同样本的代价差异,可以解决其他方法中存在的一些样本限制和代价敏感性限制。在UCI数据库上的实验结果表明,提出的代价敏感随机森林分类器在分类精度、AUC面积和Kappa系数上都具有较优的性能。

2 代价敏感学习与随机森林

2.1 代价敏感学习

给定训练集$\{(\boldsymbol{x}^1,\boldsymbol{y}^1),\cdots,(\boldsymbol{x}^N,\boldsymbol{y}^N)\}$,其中$\boldsymbol{x}^n$为样本的$D$维特征向量,$\boldsymbol{y}^n\in\{1,\cdots,C\}$为分类标签,分类的目标是为未知

师彦文(1991-),男,硕士生,主要研究领域为机器学习、代价敏感、数据挖掘等,E-mail:swpusteven0212@163.com;王宏杰(1992-),男,硕士生,主要研究领域为机器学习、粒计算等。

特征向量 $\boldsymbol{x}^*$ 分配一个分类标签。在大多数的分类场景中，特征向量被认为是互斥的，即每个样本仅属于一个类别[7]。分类问题互斥的基础是假设特征向量可以明确地划分有限类。

代价敏感学习能够应对不互斥的分类场景[8]。对于这类场景，人们希望通过多个分类器来获得准确率最高的结果。在测试时，需要对不同分类器设定合适的权重来作为其对给定问题求解的置信度。在训练时，每位专家(在本文中为分类器)会根据其计算结果和实际结果的比较，获得一个任务分数。在一些特定场景中，某一单一专家可能获得较高的任务分数；而在另外一些场景中，可能有多个专家取得较高分数。此时，需要考虑专家间的差异性，而不是只关心哪个专家取得的分数最高。

在代价敏感分类中，本文使用了一组专家 S 和一组训练集 $\{(\boldsymbol{x}^1,\boldsymbol{y}^1),\cdots,(\boldsymbol{x}^N,\boldsymbol{y}^N)\}$，其中 $C=|S|$，$\boldsymbol{y}^n\in\mathbb{R}^C$ 为标签向量。标签向量中的每个元素 y_c^n 都是连续值($0\leqslant y_c^n\leqslant 1$)，用来表示第 c 个专家的任务分数。任务分数 y_c^n 较高意味着对于特定样本 $\boldsymbol{x}^n$，第 c 个专家的准确度较高。传统的二值分类可以作为 $\boldsymbol{y}^n=(0,1)$ 时的一种特例。具体而言，对于给定的任务实例 $\boldsymbol{x}$，希望找到专家 $c\in S$ 使得任务分数最大，因此需要最小化代价函数，计算公式如下：

$$L_{cs}(c,f(c,\boldsymbol{x}))=1-f(c,\boldsymbol{x}) \tag{1}$$

其中，$f(c,\boldsymbol{x})$ 为对于实例 $\boldsymbol{x}$，专家 c 的任务分数，其最好的结果为 1，最坏的结果为 0。当测试中给定一组新的任务实例 $\boldsymbol{x}^*$ 时，可以根据已有的任务分数为专家分配不同的准确性概率权重。

2.2　随机森林分类

随机森林是一种由多棵决策树集成的分类方法，每棵树都使用数据的一个随机子集独立训练[9]。树从根节点依次迭代生长，对于其中任意一个节点 P，通过一组随机分割决策将数据点分配到其左孩子节点(L)和右孩子节点(R)。随机森林的执行过程如下[10]：

(1)选取训练集。若需要构建 K 个基础分类器，则对数据集作 K 次抽样。

(2)构建随机森林模型。假设训练数据集中具有 M 个属性，从 M 个属性中随机抽取 F 个属性作分类属性集。然后，采用分类回归树(Classification And Regression Tree，CART)[11]方法构建单棵决策树，不限制每棵决策树的生成，且不进行任何剪枝。

(3)投票。随机森林分类器采取贪婪方法来进行决策，利用所构建的 K 棵决策树来对某个数据进行分类，最后进行投票，将得票最多的数据作为随机森林的最终输出结果。

传统基于分类的随机森林(Classification-based Random Forest，CBRF)利用 CART 构建单个二叉决策树。CART 决策树通常使用“Gini 指数”来表示不纯度(impurity)，并用于选择划分属性。Gini 指数 I_{gini} 为：

$$I_{gini}=\sum_{c=1}^{C}\sum_{c'\neq c}p_c p_{c'}=1-\sum_{c=1}^{C}p_c^2 \tag{2}$$

其中，p_c 表示节点上的数据属于第 c 类的归一化后验概率，由该节点上属于 c 类的样本数量比例来计算。Gini 指数反映了数据集中样本类别标记的不一致性。那么，数据集 D 中属性 a 的不纯度为：

$$I(a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}I_{Gini}(D^v) \tag{3}$$

其中，D^v 表示数据集 D 中的第 v 个属性。根据 Gini 指数，将候选属性集合中 Gini 值最小的属性作为最优划分属性。

3　提出的代价敏感随机森林分类器

代价敏感学习是一种能够处理不平衡数据的分类方法，随机森林算法在特征子空间构造决策树，能够处理高维数据。为此，将代价敏感学习和随机森林相结合就能够实现对高维不平衡数据的有效分类。

本节描述了一种新型的代价敏感不纯度度量，构建了一种代价敏感随机森林分类器。在传统随机森林分类器的决策机制中引入了代价函数，在每个节点上通过最小化式(1)所示的代价函数来对数据进行分类。

3.1　不纯度度量

在随机森林中，将节点处的信息增益 $E_{\inf}$ 作为每个潜在分割方案的质量度量，其表达式为：

$$E_{\inf}=I(P)-\left(\frac{N_L}{N}I(L)+\frac{N_R}{N}I(R)\right) \tag{4}$$

其中，N，N_L 和 N_R 分别表示样本落在双亲节点、左孩子节点和右孩子节点的样本数量。为了计算信息增益，需要计算每个节点处的不纯度 $I(\cdot)$，其用来表示在特定节点处数据点标签的不一致程度[12]。对于聚类的场景，如果一个节点处的数据都属于同一类，那么其不纯度应该是最小的；如果一个节点处的数据点均匀分布于各类中，那么其纯度应该最大。目前已经提出了几种用于分类的不纯度度量，除了 2.2 节中提到的 Gini 度量 I_{gini} 外，还有熵度量 I_{ent} 以及误分类率度量 I_{mcl} 等，这些度量的表达式分别如下[13]：

$$I_{ent}=-\sum_{c=1}^{C}p_c\log_2(p_c) \tag{5}$$

$$I_{mcl}=1-\max(\boldsymbol{p}) \tag{6}$$

其中，$\boldsymbol{p}$ 是一个 C 维向量，每一项由 p_c 组成。

一元回归 I_{reg} 的目标是最小化落在一个节点处的所有连续响应值的方差，其表达式为：

$$I_{reg}=\sum_{n\in N^*}(y^n-\mu_y)^2 \tag{7}$$

其中，N^* 是数据点 N 落在节点上的子集，μ_y 是 N^* 的平均标签值。

3.2　代价敏感不纯度度量

传统 CBRF 的不纯度度量 I_{gini} 在训练过程中不能利用任务分数信息，而是依靠样本的后验概率使数据落在节点上。其忽略了任务分数，由获得最高任务分数的专家为样本 x^n 设定分类标签，$c=\arg\max_c y_c^n$。但是，这样操作将导致一些能够提高分类准确性的信息被丢弃。为此，文献[14]在 I_{gini} 的基础上提出了一种代价敏感的 Gini 度量，称为 I_{csg}。其直接使用任务分数来计算每个节点上 C 维类别的后验概率 p，而不是使用落在一个节点上的每类样本的数量(N^*)。修正后的节点后验概率计算公式如下：

$$p_c=\sum_{n\in N^*}y_c^n\Big/\sum_{n\in N^*}\sum_{k=1}^{c}y_k^n \tag{8}$$

3.3　提出的成对差异代价敏感不纯度度量

在实际应用中并不在乎每个数据点的绝对任务分数，但会考虑对于特定样本的每个专家得分的相对差异，本文仅需将专家得分差异明显的样本分割出来。为此，基于标签向量的任务分数的成对差异，定义了一种成对差异代价敏感(PDCS)不纯度度量 I_{pdcs}，用于表示一个专家比另一个专家的优越程度。I_{pdcs} 的表达式为：

$$I_{pdcs}=\frac{1}{C^2-C}\sum_{i=1}^{C}\sum_{j=1}^{C}(f_{i\to j}-f_{i\to j}^2),\forall i\neq j \tag{9}$$

成对的专家后验概率 $f_{i\to j}$ 可以根据 N^* 中每个数据点的分类来计算，公式如下：

$$f_{i\to j}=\frac{\sum_{n\in N^*}(d_{i\to j}^n)^2}{\sum_{n\in N^*}(d_{i\to j}^n+d_{j\to i}^n)} \tag{10}$$

差异向量 $\boldsymbol{d}_{i\to j}$ 是一个大小为 $|N^*|$ 的矢量，其每个元素 $d_{i\to j}^n$ 表示每个特征向量元素的正向差异，即：

$$d_{i\to j}^n=\begin{cases}y_i^n-y_j^n, & \text{if } y_i^n>y_j^n\\ 0, & \text{others}\end{cases} \tag{11}$$

3.4 不纯度度量的对比分析

在训练过程中，决策树中一个节点的潜在分割的接受或拒绝是由该节点的不纯度度量决定的。为了比较 3 种不纯度度量的性能，在两个样本数据集上进行了实验，并计算不纯度得分。两个数据集分别为标准二分类任务 N_1 和存在无互斥情况的分类任务 N_2。

两个数据集的标签向量和相应的不纯度度量值如表 1 所列。表 1 中，y 表示标签向量，y 表示具有较高任务分数的专家的序号。N_1 和 N_2 两个数据集都包含 4 个数据点，唯一的区别在于 N_2 中的一个数据点对于两个专家具有相似的任务分数。一个好的分割必须具有较小的不纯度度量值。

表 1 不同数据集(N_1 和 N_2)的不纯度比较

数据集 N_1					
n	y		y	$d_{1\to 2}$	$d_{2\to 1}$
1	1	0	1	1	0
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	2	0	1
I_{gini}			0.3326		
I_{csg}			0.3326		
I_{pdcs}			0.3326		
数据集 N_2					
n	y		y	$d_{1\to 2}$	$d_{2\to 1}$
1	1.0	0.00	1	1.0	0.00
2	1.0	0.00	1	1.0	0.00
3	1.0	0.00	1	1.0	0.00
4	0.5	0.54	2	0.0	0.01
I_{gini}			0.3326		
I_{csg}			0.2158		
I_{pdcs}			0.0241		

由于标准二分类任务 N_1 包含互斥标签，因此本文新提出的代价敏感度量 I_{pdcs} 与传统的 I_{gini} 和 I_{csg} 两种度量方法得到的结果是相同的。在无互斥情况的 N_2 中，I_{pdcs} 认为其中一个专家能够更好地进行分类，因为它具有较高的任务分数。而 I_{gini} 和 I_{csg} 对任务分数的变化十分敏感，即使专家间的差别非常细微，也不能找到成对差异，可能会忽视一些潜在的最优分割，从而产生较高的不纯度度量。但 I_{pdcs} 却不会对这些微小扰动施加额外的惩罚，原因在于 I_{pdcs} 利用了一个事实，即其中一个专家会对整个集合给出较高的任务分数。

4 实验及分析

4.1 实验设置

在数据挖掘开源平台 WEKA 上实现随机森林分类器，运行环境为配备 Intel Core i5 CPU @ 2.67GHz、8.00GB 内存、64 位 Windows 7 系统的 PC 机。

为了验证提出的随机森林分类器对不平衡数据的有效性，采用一个典型的数据挖掘数据库：UCI 数据库[15]。从该数据库中选择 5 个二分类不平衡数据集进行实验，分别为 SNP，microRNA，ionosphere，GeneChip 和 m_neg_all。这些数据集的相关描述如表 2 所列。

表 2 实验样本集的描述

数据集	样本数量	类别属性数量	不平衡度
SNP	3260	25	15.74
microRNA	8687	32	44.01
ionosphere	351	34	1.78
GeneChip	62	178	2.44
m_neg_all	14698	188	2.72

对于每个实验，将本文提出的基于 I_{pdcs} 不纯度度量的代价敏感随机森林分类器(PD-CSRF)与使用 I_{csg} 的代价敏感随机森林(G-CSRF)和使用传统 I_{gini} 的基于分类的随机森林(CBRF)进行比较。在未达到节点的最小样本数量的情况下，树会持续生长至最大指定深度。标签矢量包含连续的任务分数值，但对于 CBRF 的每个观察，设定分类标签为 y^n 的最大值的索引。分类不依靠分类分数而是任务分数，因为分类分数只能衡量最佳专家决策的概率，而任务分数则衡量了专家使用给定模型所带来的真正效益。

4.2 性能度量

(1)准确性(Accuracy)

本文利用 Accuracy 来表示森林模型在不平衡数据上的整体分类性能，其定义为正确分类的样本比率[13]。使用查准率(precision)和召回率(recall)两个度量来有效表示准确性。precision 表示正确识别的正类样本与所有正类样本的比值，recall 表示正确识别的负类样本与所有负类样本的比值，其表达式分别为：

$$pre=\frac{t_{pos}}{pos},rec=\frac{t_{neg}}{neg} \tag{12}$$

其中，t_{pos} 为正确分类的正类样本的数量，pos 为正类样本的总数量，t_{neg} 为正确分类的负类样本的数量，neg 为负类样本的总数量。因此，准确性可表示为：

$$Accuracy=(pre*\frac{pos}{pos+neg})+(rec*\frac{neg}{pos+neg}) \tag{13}$$

(2)AUC 面积

AUC 面积指 ROC 曲线与 X 坐标轴之间的面积，其越接近于 1，分类器的分类性能越好。ROC 曲线是一种用来表示分类结果的负类样本错误分类比例(FPR)与正类样本正确分类比例(TPR)关系的曲线，ROC 曲线越偏向左上方，表示该分类器能够更好地分类不平衡数据。

(3)Kappa 系数

Kappa 系数用于评估分类器分类的一致性程度，其通过度量两种方法的实际一致性比率和随机一致性比率的差别来计算，表达式如下：

$$Kappa=(po-pe)/(1-pe) \tag{14}$$

$$po=(pre+rec)/(pos+neg) \tag{15}$$

$$pe=(\frac{pos*N_{pso}}{pos+neg}+\frac{neg*N_{neg}}{pos+neg})/(pos+neg) \tag{16}$$

其中，po 表示观察一致性比率，pe 表示期望一致性比率；N_{pso} 为分类为正类样本的总数(包括正确和错误分类)，N_{neg} 为分类为负类样本的总数(包括正确和错误分类)。Kappa 系数的取值范围为[−1,1]，其值越大，表明一致性越好。

4.3 实验结果

在 5 个数据集上进行相关实验，并统计各种随机森林分

类器的性能指标。其中，各种分类器的分类 Accuracy、AUC 面积和 Kappa 系数结果分别如表 3－表 5 所列，其对应的柱状图分别如图 1－图 3 所示。

表 3　各种分类器的分类 Accuracy 比较/%

数据集	CBRF	G-CSRF	PD-CSRF
SNP	91.4	93.2	94.7
microRNA	96.2	97.7	97.8
ionosphere	91.3	92.8	93.5
GeneChip	75.8	83.4	86.2
m_neg_all	93.2	95.5	96.7

表 4　各种分类器的分类 AUC 面积比较/%

数据集	CBRF	G-CSRF	PD-CSRF/
SNP	0.7124	0.7553	0.7618
microRNA	0.8461	0.9143	0.9266
ionosphere	0.9632	0.9804	0.9863
GeneChip	0.6855	0.7262	0.7359
m_neg_all	0.9432	0.9663	0.9701

表 5　各种分类器的分类 Kappa 系数比较/%

数据集	CBRF	G-CSRF	PD-CSRF
SNP	0.5024	0.5946	0.6032
microRNA	0.6271	0.6364	0.6402
ionosphere	0.8213	0.8435	0.8493
GeneChip	0.4010	0.4635	0.4863
m_neg_all	0.8566	0.8783	0.8802

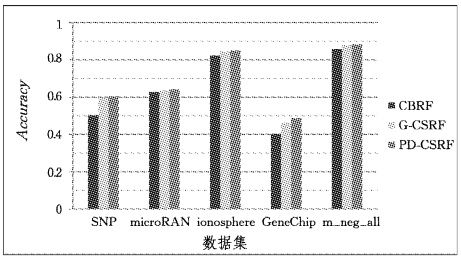

图 1　各种分类器的分类 Accuracy 直方图

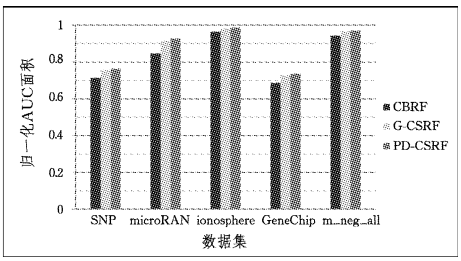

图 2　各种分类器的分类 AUC 面积直方图

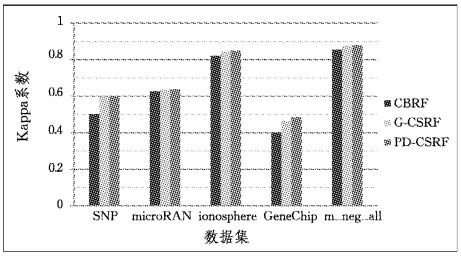

图 3　各种分类器的分类 Kappa 系数直方图

由上述 3 种指标的比较结果可知，提出的 PD-CSRF 分类器都获得了较好的性能。其中，Accuracy 的优势说明了所提方法具有较高的分类精度，AUC 面积和 Kappa 系数的优势说明所提方法能够更好地处理不平衡数据。

传统 CBRF 的各项性能最差，因为其仅依靠 Gini 度量来构建决策树，未考虑误分类代价。G-CSRF 分类器与提出的 PD-CSRF 分类器的性能相近，原因在于其在随机森林中引入了代价敏感，提高了分类准确性。而 PD-CSRF 分类器在结合随机森林和代价敏感学习的基础上改进了不纯度度量，有效地提高了分类性能。

结束语　本文提出了一种新型不纯度度量，并结合代价敏感学习和随机森林算法构建了一种代价敏感随机森林分类器。通过充分利用任务分数信息，构建了更具代表性的分类器来解决不平衡数据集的分类问题。所提出的分类器与传统分类方法相比，在 Accuracy、AUC 面积和 Kappa 系数上都有明显的提升。

参 考 文 献

[1] 刘偲，秦亮曦. 模糊决策粗糙集代价敏感属性约简研究[J]. 计算机科学，2016，43(S2)：67-72.

[2] LÓPEZ V，FERNÁNDEZ A，MORENO-TORRES J G，et al. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics[J]. Expert Systems with Applications，2012，39(7)：6585-6608.

[3] AODHA O M，BROSTOW G J. Revisiting Example Dependent Cost-Sensitive Learning with Decision Trees[J]. 2013，25(6)：193-200.

[4] 邓生雄，雒江涛，刘勇，等. 集成随机森林的分类模型[J]. 计算机应用研究，2015，32(6)：1621-1624.

[5] 赵士伟，卓力，王素玉，等. 一种基于 NNIA 多目标优化的代价敏感决策树构建方法[J]. 电子学报，2011，39(10)：2348-2352.

[6] BAHNSEN A C，AOUADA D，OTTERSTEN B. Example-dependent cost-sensitive decision trees[J]. Expert Systems with Applications，2015，42(19)：6609-6619.

[7] 邓少军，冯少荣，林子雨. 一种新的多分类代价敏感算法[J]. 厦门大学学报(自然科学版)，2017，56(2)：231-236.

[8] THAI -NGHE N，GANTNER Z，SCHMIDT-THIEME L. Cost-sensitive learning methods for imbalanced data[C]// International Joint Conference on Neural Networks. IEEE，2010：1-8.

[9] ZHOU Q，ZHOU H，LI T. Cost-sensitive feature selection using random forest：Selecting low-cost subsets of informative features[J]. Knowledge-Based Systems，2016，95(3)：1-11.

[10] 王爱平，万国伟，程志全，等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报，2011，22(9)：2059-2074.

[11] 张钰，陈珺，王晓峰，等. 随机森林在滚动轴承故障诊断中的应用[J]. 计算机工程与应用，2017，53(6)：312-319.

[12] 胡记兵. 基于决策树的组合分类器的构建和部署[D]. 杭州：浙江工业大学，2008：17-18.

[13] SOFEIKOV K I，TYUKIN I Y，GORBAN A N，et al. Learning optimization for decision tree classification of non-categorical data with information gain impurity criterion[C]// International Joint Conference on Neural Networks. IEEE，2014：3548-3555.

[14] D'AMBROSIO A，TUTORE V A. Conditional Classification Trees by Weighting the Gini Impurity Measure[M]// New Perspectives in Statistical Modeling and Data Analysis. Springer Berlin Heidelberg，2011：273-280.

[15] 黄光鑫. 支持向量数据描述与支持向量机及其应用[D]. 成都：电子科技大学，2011：64-66.