# 一种新的基于SVDD的多类分类算法

缪志敏[1] 潘志松[1] 袁伟伟[1] 赵陆文[2]
(解放军理工大学指挥自动化学院 南京 210007)[1] (解放军理工大学通信工程学院 南京 210007)[2]

**摘 要** 目前的多类学习方法大多将多类问题转化为二类问题,这样处理除了时间开销大,还存在识别盲区。提出了一种直接进行多类学习的算法multi-SVDD。该算法在考虑大样本和多类样本数据中的类内不平衡现象基础上,首先为每类训练样本进行聚类,根据聚类结果由支持向量数据描述(SVDD,Support Vector Date Description)建立多个最小包围球。根据测试样本到SVDD所建立的最小包围球的距离来确定测试样本属于哪个聚类,最终可判断测试样本属于哪个类。multi-SVDD算法在时空开销上相比最小包围球方法没有明显增长,而实验效果则好于最小包围球方法。

**关键词** 多类学习,支持向量数据描述,不平衡学习,聚类
**中图法分类号** TP393.08 **文献标识码** A

## New Multi-class Classification Based on Support Vector Date Description

MIAO Zhi-min[1] PAN Zhi-song[1] YUAN Wei-wei[1] ZHAO Lu-wen[2]
(Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China)[1]
(Institute of Communication Engineering, PLA University of Science and Technology, Nanjing 210007, China)[2]

**Abstract** Most of the multi-class learning methods transfer the multi-class classification problems to two-class classification problems, which not only are time-expensive but also have some region undiscriminating. A direct multi-class learning algorithm named multi-SVDD was proposed. Based on the consideration that there is within-class imbalance in large data sets and multi-class data sets, every class of the training data was firstly clustered. Some minimum bounding hyperspheres were formed by Support Vector Date Description (SVDD) according to the clustering results. A test sample is assigned to the label of hyperspheres if its distance to the sphere center is smaller than or equal to the radius. Compared with minimum enclosing hypersphere algorithm, the multi-SVDD algorithm doesn't become worse in time and space cost, and the experiment result is better.

**Keywords** Multi-class Learning, SVDD, Imbalanced learning, Clustering

在模式分类学习中,通常讨论的是二类学习问题,而现实生活中很多数据都是多类别的。目前普遍采用的方法便是将多类分类问题逐步转化为二类分类问题。从二类学习扩展到多类学习问题不是一个简单的过程。笔者通过对目前流行的多类学习进行分析,增加对多类数据中的不平衡考虑,提出了一种新的解决多类问题算法:multi-SVDD。在UCI数据集上进行的实验表明,本算法在解决多类不平衡问题上是有效的。

## 1 多类学习方法概述

目前对多类分类的研究主要有两个方向:间接解决和直接解决。间接解决多类分类的方法是将多类分类问题转化为二类分类问题,即用多个二类分类器组成一个多类分类器。这类方法主要有以下4种:一对多(OVA,One-Vs-All)方法、一对一(OAO,One-Against-One)方法、DAG(Directed Acyclic Graph)方法和纠错输出码(ECOC,Error Correcting Output Codes)方法。SVM由于其以结构化风险为基础,具有良好的泛化性能。目前将多类学习问题转化成二类问题时大多使用SVM[1]。OVA是一种很简单的多类分类方法,是为每个类构建一个二类分类器,对于$N$个类别的样例,则要构造$N$个二类分类器[2]。对第$i$个类的二类分类器来说,其训练样本集的构成为属于$i$类的样例为正类,而不属于该类的其他所有样本都为负类。OVO方法是对多类别数据进行两两区分,为任意二个类构建分类超平面[3]。对于$N$类数据集,则需要构造$N(N-1)/2$个二类分类器。测试时常用投票法,得票最多的类为测试样本所属类。DAG方法的训练与OVA类似,所以训练时间也与OVO相同。不同的是在进行测试时,DAG方法使用有限Directed Acyclic Graph (DAG)结构来做出判断[4]。ECOC是Bose和Ray-Chaudhuri在1960年提出的一种分布式输出码,1995年Dietterich提出用ECOC解决大类别问题。其主要思想就是对类别进行二进制

到稿日期:2008-06-10 本文受国家自然科学基金"单类分类器和数据不平衡分类问题研究"(No. 60603029),江苏自然科学基金"基于单类分类器的安全审计中的异常检测研究"(No. BK2005009)项目支持。
缪志敏(1978—),女,博士生,研究方向为网络安全、模式识别,E-mail:olivermiao@126.com;潘志松(1973—),男,副教授,研究方向为网络安全、模式识别;袁伟伟(1982—),男,博士生,研究方向为网络管理、网络安全。

编码，将多类分类问题转化为多个二类分类问题[5]。图 1(a)是 OVA 分类方法的示意图，在采用 OVA 方法对 A,B,C 这 3 类数据进行分类时，我们发现如图中被标上“?”的 4 个区域是不能被识别的盲区。如果数据点落在这些区域内，利用 OVA 方法无法辨别这些数据点的类别。图 1(b)是 OVO 分类方法的示意图，在采用 OVO 方法对 A,B,C 这 3 类数据进行分类时，发现图中中心区域是不能被识别的盲区，如果数据点落在这些区域内，利用 OVO 方法无法辨别这些数据点的类别。对于大于 3 类的多类分类问题，依据 OVA 和 OVO 建立的分界面也存在更多的识别盲区[6]。

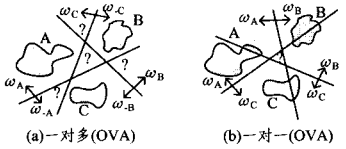


图 1　常用多类分类方法

在间接方法中，除了存在图 1 中的识别盲区，对多类问题的分析也是不全面的，如 OVA 中将导致训练样本中类别的严重不平衡；OVO 中建立两两类别间的分类器，忽视了其他类别的信息。最近几年来，不少研究者试图通过设计直接解决多类问题的 SVM 来解决多类分类问题，同时处理各类数据，没有忽略各类之间的关联信息。这类方法中，最著名的是采用 $K$ 个超球对 $K$ 类数据同时进行描述，每个超球包含一类数据[7]，如图 2 所示。

支持向量数据描述(SVDD,Support Vector Date Description)是一种典型的单类分类器，通过建立包围目标类的超球来拒绝非目标类数据。由于 SVDD 中仅采用一类数据进行学习，有不少研究者将其扩展为多类分类器来解决多类分类问题。用 SVDD 来解决多类分类问题，通常做法如下[7]：

首先给定训练数据 $(x_1,y_1),\cdots,(x_N,y_N)$，$x_i\in R^n$，$y_i\in\{1,\cdots,K\}$，为训练数据中的每一类数据建立一个包围该类所有训练样本的最小超球，每个超球用球心 $a_m$ 和半径 $R_m$ 定义，$K$ 个超球可以通过解决下面 $K$ 个二次规划问题求得：

$$\text{minimize } R_m^2+C_m\sum_{i;y_i=m}\xi_i,m=1,\cdots,K \tag{1}$$

约束条件为

$$\|\phi(x_i)-a_m\|^2\leqslant R_m^2+\xi_i,\forall i,y_i=m,m=1,\cdots,K$$

$$\xi_i\geqslant 0,\forall i \tag{2}$$

其中，$\xi_i$ 是松弛因子，$C_m$ 调节超球面，控制误差。与 SVDD 类似，利用 Lagrange 算子求解二次规划问题，可以写成如下的对偶形式：

$$\text{maximize }\sum_{i;y_i=m}\alpha_iK(x_i,x_i)-\sum_{i,j;y_i,y_j=m}\alpha_i\alpha_jK(x_i,x_j),$$

$$m=1,\cdots,K \tag{3}$$

约束条件为

$$\sum_{i;y_i=m}\alpha_i=1 \tag{4}$$

且 $y_i=m$，有 $0\leqslant\alpha_i\leqslant C_m\quad m=1,\cdots,K$

在 $K$ 个超球确定后，定义测试点与 $K$ 个超球面 $S_m$ 的判定函数，则

$$\text{class of } x\equiv\arg\max_{m=1,\cdots,K}\text{sim}(x,S_m),m=1,\cdots,K \tag{5}$$

不少研究者在通常做法的基础上进行了改进。Pei-Yi Hao 对 $C$ 值进行改进，令 $C_m=\frac{C}{N_m}$，$C$ 为一常数，$N_m$ 为第 $m$ 类数据个数[8]。

还有研究者对判定函数进行了一些变形和改进。最简单的决策函数是通过测试点到超球的球点距离进行判定，判定函数中的相似性函数如下：

$$\text{sim}(x,S_m)=-\|\phi(x)-a_m\|,m=1,\cdots,K \tag{6}$$

Zhu 等引入超球的半径，提出了新的相似性函数[9]。Wu 等通过研究测试点与所建立超球之间的不同关系采用不同的判定函数[10]。

Wang Defeng 等提出了 Structured One-Class Classification 算法[11]，是在考虑数据分布的基础上，将一类目标数据用多个超椭球来描述，以获得对目标数据更有效的描述。该算法在进行多类分类时，依旧采用一对多方法，同样具有间接解决多类分类问题的缺陷。而研究者在采用 SVDD 算法进行多类学习时，则很少考虑多类学习中的类别不平衡问题。在本文中，笔者在现有基于最小包围球的多类学习方法和二类不平衡问题解决方法基础上提出了一种针对多类学习的方法：multi-SVDD，并在 UCI 数据集上对该算法进行实验，与基于最小包围球的多类解决方法进行了比较，证明其有效性。

## 2　multi-SVDD 算法

Weiss 对类别不平衡问题对分类器性能造成影响进行了详细的分析[12]，总结有以下一些原因：不恰当的性能评价准则、不恰当的归纳偏置、样本数目的绝对稀少和相对稀少，以及数据碎片和噪声等。在文献[13]中提到了类内不平衡现象对分类器性能的影响，但目前很少有研究者从类内不平衡着手对不平衡问题进行解决，而类内不平衡在大样本和多类样本数据集中表现更为突出。笔者在基于实例学习中的下采样方法的启发下，综合聚类和单类学习器，提出了一种针对类内不平衡现象的多类不平衡学习算法：multi-SVDD 算法。

在文献[13]中提出通过采样来实现类内间平衡，因为采样过程本身存在信息丢失、过拟合等缺陷，不恰当的采样策略对学习性能的影响更大。样本的类别信息不能揭示类内不平衡现象，无监督学习能在未知样本类别的基础上揭示观测数据的一些内部结构和规律，隐蔽的不平衡信息可以借助无监督学习来发现。最具有代表性的无监督算法是聚类，在 multi-SVDD 算法中采用聚类将每类样本分割成多个子类，通过判定测试样本与子类间的关系来确定测试样本最终的类别。而该算法中采用 SVDD 来建立每个子类的分界面，这是基于以下原因：(1)“物以类聚”，在相同聚类中的样本有很大可能拥有相同的类标(这里指子类标)；(2)单类分类器能依靠一类样本建立分类面；(3)如果依据聚类中心来建立球状分类面难以形成对子类边界细致的刻画，将导致各子类重复区域存在，不利于判定。

multi-SVDD 算法的基本思想是：首先为每类训练样本进行聚类，根据聚类结果建立多个 SVDD 分类器。定义测试样本到单类分类器的判定函数来确定测试样本属于哪个聚类，进而判断测试样本属于哪个类。在 multi-SVDD 算法中，采用的是经典聚类算法：K-means。由于聚类算法的缺陷，每个类内聚类个数的确定目前还没有很有效的方法。以 SVDD 单类分类器为例，给定训练数据 $(x_1,y_1),\cdots,(x_N,y_N)$，$x_i\in R^n$，$y_i\in\{1,\cdots,K\}$，$n_k$ 为第 $K$ 类数据的数量，令 $n'$ 为数量最少类别个数的整数倍，根据 $l_k=n_k/n'$ 确定每类样本的聚类个

数，然后根据聚类算法确定第 $k$ 类样本 $l_k$ 个聚类的中心及每个聚类的样本集。根据每个聚类的样本集建立 $J=l_1+\cdots+l_k$ 个单类分类器，而每个聚类的样本集为 $D_j(j=1,\cdots,J)$。判定函数采用了式(6)，根据测试样本 $x$ 到每个单类分类器构成的超球的球心距离来确定类别，算法如图 3 所示。

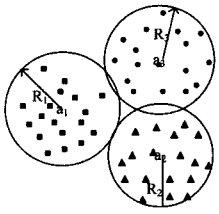


图 2 基于超球的多类分类方法示意图

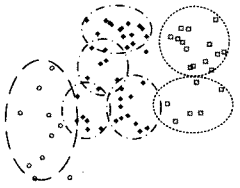

图 3 multi-One Class 算法示意图

与 Structured One-Class Classification 类似，multi-SVDD 可用下面的公式表示：

$$\min\ r_j+C\sum_{m=1}^{|D_j|}\xi_m \tag{7}$$

约束为

$$\|(x_m-a_j)\|\leqslant r_j+\xi_m,$$

$$\xi_m\geqslant 0,\ x_m\in D_j, m=1,\cdots,|D_i|, r_j>0 \tag{8}$$

测试样本 Z，最终的判定函数为

$$\text{class of } Z\in k \text{ class if } \exists j(\|Z-a_j\|\leqslant r_j, j\in\{1,\cdots,J\})$$

$$\text{and } (l_1+\cdots+l_k)<j\leqslant(l_1+\cdots+l_{k+1}) \tag{9}$$

假设通过 K-means 获得 $K$ 个聚类，SVDD 算法的时间复杂度为 $O(n^3)$，则 multi- SVDD 算法的时间复杂度为 $O(Kn^3)$。

multi-SVDD 算法有以下几个特点：

①利用了欠采样策略，实现不平衡类别样本间数量上的平衡。

②数量多的样本数据分布往往更为复杂，类内本身就存在着数据不平衡现象。采用聚类的方法发现数据间的结构性信息，实现对大样本数据的分割。

③对分割好的每一部分数据建立一个单类分类器，从而实现对数量大、分布复杂的数据更加细致的刻画。

④集成学习被运用在最后的测试数据判定中，这样进一步保证了学习器的泛化性能。

多类不平衡问题比二类不平衡问题要复杂。二类不平衡的评价标准不能用于多类不平衡。目前用于多类不平衡的评价标准还很少，除了识别精度、误识率外，常见的还有 VUS 值[15]，MAUC[16]，都是在二类评价准则 AUC 的基础上进行扩展。求解 VUS，MAUC 是一个较复杂的过程。在本文中，笔者在二类不平衡评价标准 GMA[17] 的基础上，提出了一种用于评价多类学习器的简单实用的标准 MGMA(multi Geometric Mean of Accuracy)，与识别率一起评价多类学习算法的学习性能。

定义 $MGMA$ 为所有类别查全率的几何平均：

$$MGMA=\sqrt{r_1\cdots r_i\cdots r_K},\quad i=1,\cdots,K \tag{10}$$

$r_i$ 为第 $i$ 类的查全率，一共有 $K$ 类样本。只有当所有类别的查全率都较高时，$MGMA$ 值才会较大。

## 3 实验结果与分析

本文选用两个数据集对提出的 multi-SVDD 算法进行了验证，其中一个人工数据集、一个 UCI 数据集[18]。实验中取识别精度作为评价标准。所有的测试结果都是通过 10 重交叉验证进行参数选择并取平均记录，算法在 Matlab 6.5 平台上进行运算。

### 3.1 人工数据集

试验中采用的人工数据集由两个 banana 形数据集和一个呈高斯分布的数据集组成。其中两个 banana 数据集数量分别设置标志为“apple”和“banana”，呈参数为[3,7]的高斯分布的数据集设置标志为“pear”。所有数据都为 2 维。

图 4 是采用 multi- SVDD 算法和 Tax 提出的多类分类算法(简称 original- SVDD)对人工数据集中 3 类数据不同比值下进行分类后形成的分界面图。两个算法中 SVDD 算法的参数设置相同，拒识率为 0.1，都采用 rbf 核函数，核参数为 2。从 a 与 b 比较、c 与 d 比较、e 与 f 比较可以清楚地看出，在 3 类数据不同比值下，与 original-SVDD 算法相比，multi-SVDD 算法用更小的面积包围了目标类样本，更好地吻合了数据分布曲线，取得了较好的分类性能。而 Tax 提出的多类分类算法则将大量的数据识别为 3 类之外的野值点，识别效果较差。

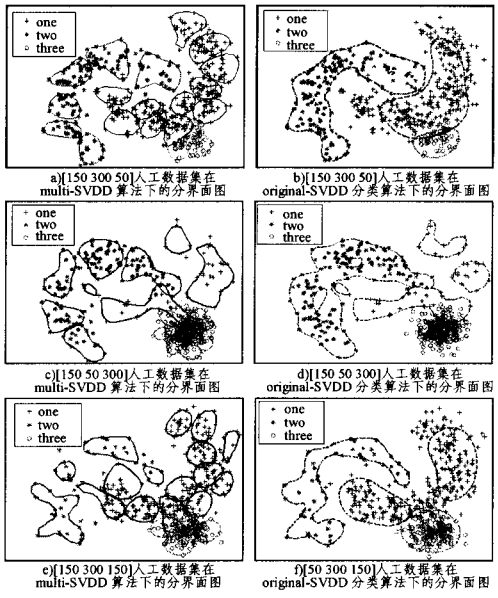


图 4 mult- SVDD 与 original- SVDD 实验结果比较

为了更好地了解 multi-SVDD 算法性能，表 1 给出了在 3 类人工数据不同比值下的各类数据的识别率(r1,r2,r3)以及总识别率(r)、MGMA 值，$n1$，$n2$，$n3$ 为三类数据的数量。分析表 1 的实验数据，我们发现 multi-SVDD 算法比 original-SVDD 算法的识别精度高近 7%，同时 MGMA 值大于 0.24。在进一步证明 multi-SVDD 算法处理多类不平衡问题有效的同时，也说明 MGMA 值虽然计算简单，但作为多类不平衡学习的评价标准是有效的。

表 1 两种多类分类器在人工数据集上的实验结果

| n1/n2/n3 | multi-SVDD | | | | | original-SVDD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | r1 | r2 | r3 | r | MGMA | r1 | r2 | r3 | r | MGMA |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 50/150/300 | 0.820 | 0.993 | 1.000 | **0.976** | **0.914** | 0.820 | 1.000 | 0.907 | 0.926 | 0.744 |
| 150/300/50 | 0.940 | 0.987 | 1.000 | **0.964** | **0.928** | 0.907 | 0.897 | 0.660 | 0.876 | 0.537 |
| 300/50/150 | 0.957 | 0.920 | 1.000 | **0.948** | **0.880** | 0.900 | 0.800 | 0.893 | 0.888 | 0.643 |
| 50/300/150 | 0.900 | 0.9867 | 1.000 | **0.968** | **0.888** | 0.800 | 0.987 | 0.947 | 0.896 | 0.748 |
| 150/50/300 | 0.927 | 0.980 | 1.000 | **0.956** | **0.908** | 0.860 | 0.800 | 0.910 | 0.884 | 0.626 |
| 300/150/50 | 0.967 | 0.980 | 0.980 | **0.952** | **0.929** | 0.947 | 0.873 | 0.860 | 0.916 | 0.711 |
| avg | 0.919 | 0.974 | 0.997 | **0.961** | **0.908** | 0.872 | 0.893 | 0.863 | 0.898 | 0.668 |

### 3.2 iris 数据集

iris 数据集是包含 3 类不同类别的数据，其中第二类和第 3 类是线性不可分的，实验时在三类样本中随机选择不同数量的样本作为训练样本，剩下的数据作为测试样本，采用 multi-SVDD 和 original-SVDD 两种算法进行多类学习，实验结果如表 2 所示。$n1$，$n2$，$n3$ 为训练数据中三类样本的数量。original-SVDD 算法除在“25/10/50”上分类结果优于 multi-SVDD 外，在其他数据集上的结果都比 multi-SVDD 算法差，multi-SVDD 识别精度的平均值比 original-SVDD 高 3.6%，multi-SVDD 的 MGMA 平均值也比 original-SVDD 高 0.034。进一步说明增加了数据类别不平衡考虑的 Multi- SVDD 算法比直接的多类分类方法更有效。

表 2 iris 数据上的实验结果

| n1/n2/n3 | multi-SVDD | | | | | original-SVDD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | r1 | r2 | r3 | r | MGMA | r1 | r2 | r3 | r | MGMA |
| 50/25/10 | 1.000 | 1.000 | 0.900 | **0.988** | **0.900** | 1.000 | 0.946 | 0.900 | 0.902 | 0.851 |
| 50/10/25 | 1.000 | 0.960 | 1.000 | 0.965 | 0.960 | 1.000 | 0.914 | 0.974 | 0.937 | 0.890 |
| 25/50/10 | 1.000 | 1.000 | 0.900 | 0.988 | 0.900 | 1.000 | 0.960 | 0.914 | 0.923 | 0.898 |
| 25/10/50 | 1.000 | 1.000 | 0.960 | 0.976 | 0.960 | 1.000 | 0.980 | 0.992 | **0.982** | **0.972** |
| 10/25/50 | 0.900 | 1.000 | 0.980 | **0.976** | **0.882** | 0.950 | 0.996 | 0.906 | 0.944 | 0.857 |
| 10/50/25 | 1.000 | 1.000 | 0.960 | **0.961** | **0.960** | 0.950 | 0.996 | 0.914 | 0.952 | 0.889 |
| avg | 0.983 | 0.993 | 0.950 | **0.976** | **0.927** | 0.983 | 0.965 | 0.933 | 0.940 | 0.893 |

**结束语** 目前两种流行的多类分类问题的解决方式是间接解决和直接解决多类分类问题。由于间接进行多类学习存在时空开销大和存在分类盲区，直接进行多类学习成为了更优的选择。笔者在前人基于最小包围球解决多类学习和结构单类分类器的基础上提出了一种采用 SVDD 单类分类器解决多类分类问题的算法：multi-SVDD。该算法在时空开销上相比最小包围球方法没有明显增长，而实验效果则好于最小包围球方法。目前笔者在确定每一类超球内的样本时是通过 K-means 聚类来确定，聚类的个数则与各类数据大小比值相关。如何获得更有效的聚类结果还需进一步研究。

## 参 考 文 献


[1] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. IEEE Trans Neural Network, 13: 415-425

[2] Bottou L, Cortes C, Denker J. Comparison of Classifier Methods: A Case Study in Handwriting Digit Recognition[C]// Int. Conf. Pattern Recognition. 1994: 77-87

[3] Kreβel U. Pairwise classification and support vector machines [M]// Schölkopf B, Burges C J C, Smola A J, eds. Advances in kernel methods—support vector learning. Cambridge: MIT Press, 1999: 255-268

[4] Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAG's for multiclass classification[M]. Advances in neural information processing systems. Cambridge: MIT Press, 2000: 547-553

[5] Takenouchi T, Ishii S. Multiclass classification as a decoding problem[C]// Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence. 2007: 470-475

[6] Duda R O, Hart P E, Stork D G. Pattern classification[M]. Second Edition. John wiley & Sons, 2001: 179-180

[7] Ban T, Abe S. Implementing Multi-class Classifiers by One-class Classification Methods[C]// 2006 International Joint Conference on Neural Networks Vancouver. BC, Canada: Shera-ton Vancouver Wall Centre Hotel, July 2006: 327-332

[8] Hao P-Y, Chiang J-H, Lin Y-H. A new maximal-margin spherical-structured multi-class support vector machine[J]. Applied Intelligence, Springer Netherlands, 2007, 10

[9] Zhu M L, Chen S F, Liu X D. Sphere-structured support vector machines for multi-class pattern recognition[C]// Lecture Notes in Computer Science. 2003, 2639: 589-593

[10] Wu Q, Shen X, Li Y, et al. Classifying the multiplicity of the EEG source models using sphereshaped support vector machines [J]. IEEE Trans Magazine, 2005, 41: 1912-1915

[11] Wang Defeng, Yeung D S, Tsang E C C. Structured One-Class Classification[J]. IEEE Transactions on Systems, Man, and Cybernetics-part B: Cybernetics, 2006, 36(6): 1283-1295

[12] Weiss G M. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations, 2004, 6(1): 7-19

[13] Japkowicz N. Concept-learning in the presence of between-class and within-class imbalances[C]// Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence. Ottawa, Canada, June 2001: 67-77

[14] Tax D, Duin R. Support vector data description[J]. Machine Learning, 2004, 54: 45-66

[15] Landgrebe T, Duin R P W. A simplied extension of the Area under the ROC to the multi-class domain

[16] Hand D J, Till R. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning, 2001, 45: 171-186

[17] Kubat M, Holte R, Matwin S. Learning when negative examples abound[C]// Proc. of 9th European Conference on Machine Learning(ECML 1997). Prague, Czech Republic, 1997: 146-153

[18] Blake C L, Merz C J. UCI repository of machine learning databases. Irvine. CA: Dept Inform Computer Science, Univ. California, (online). http://kdd.ics.uci.edu/