2012 年 5 月
第 33 卷 第 5 期
计算机工程与设计
COMPUTER ENGINEERING AND DESIGN
May 2012
Vol. 33 No. 5

改进的一对一支持向量机多分类算法

单玉刚[1,2,3,4]，王 宏[1,2]，董 爽[5]

（1. 中国科学院 沈阳自动化研究所，辽宁 沈阳 110016；2. 沈阳中科博微自动化有限公司，辽宁 沈阳 110179；
3. 空军 93303 部队，辽宁 沈阳 110015；4. 中国科学院研究生院，北京 100049；
5. 北京联合大学 管理学院，北京 100101）

摘 要：支持向量机的一对一多分类算法具有良好的性能，但该算法在分类时存在不可分区域，影响了该方法的应用。因此，提出一种一对一与基于紧密度判决相结合的多分类方法，使用一对一算法分类，采用基于紧密度决策解决不可分区，依据样本到类中心之间的距离和基于 kNN（k nearest neighbor）的样本分布情况结合的方式构建判别函数来确定类别归属。使用 UCI（university of California Irvine）数据集做测试，测试结果表明，该算法能有效地解决不可分区域问题，而且表现出比其它算法更好的性能。

关键词：k 近邻；一对一支持向量机；多分类；不可分区；紧密度
中图法分类号：TP181 **文献标识号**：A **文章编号**：1000-7024（2012）05-1837-05

Improved multi-classification algorithm of one-against-one SVM

SHAN Yu-gang[1,2,3,4], WANG Hong[1,2], DONG Shuang[5]

(1. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; 2. Microcyber Inc, Shenyang 110179, China; 3. Airforce 93303, Shenyang 110015, China; 4. Graduate University, Chinese Academy of Science, Beijing 100049, China; 5. School of Management, Beijing Union University, Beijing 100101, China)

Abstract: Multi-class classification algorithm of one-against-one SVM show good performance, but the algorithm exists an unclassifiable region, which affects the application effect of the algorithm. Hence, a multi-classification algorithm of integration of one-against-one and affinity decision is presented. Firstly, the one-against-one multi-class classification algorithm is used to classify samples, and then the affinity decision is used to solve samples in the unclassifiable region and to determine categories of samples, which using the approach of distance between the sample and centers of classes and sample distribution based on kNN (k nearest neighbor) to create decision function. By adopting UCI data sets for testing, the results show that the algorithm can solve unclassifiable region issues, and show better performance than other algorithms.

Key words: kNN; one-against-one algorithm SVM; multi-class classification; unclassifiable region; affinity

0 引 言

支持向量机（SVM）是由 Vapnik[1] 提出的统计学习理论，在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。已经应用在模式识别、回归预测等领域[2-3]。支持向量机是针对二分类问题提出的。对于多分类问题，以二元分类为基础，通过一定的组合原则，构造多类分类器，实现多类可分。常见的构造方法包括一对一方法（one-against-one，OVO）和一对多方法（one-against-rest，OVR），以及在这 2 种方法基础上的改进算法：有向无环图法（directed acyclic graph，DAG）和二叉树法（binary tree，BT）、纠错输出编码法（error correcting output code，ECOC）等[4]。一对一多分类法具有优秀性能，但这种方法的不足是存在不可分区域。

为了解决一对一多分类存在不可分区域问题，文献[5]在模糊支持向量机（fuzzy support vector machine，

收稿日期：2011-06-16；修订日期：2011-08-20
基金项目：国家 863 高技术研究发展计划基金项目（2007AA041407）
作者简介：单玉刚（1971-），男，辽宁沈阳人，博士研究生，研究方向为现场总线、自动化软件和软件工程；王宏（1963-），男，河北抚宁人，研究员，博士生导师，研究方向为工业控制系统、现场总线和智能仪器仪表；董爽（1969-），女，辽宁沈阳人，博士，讲师，研究方向为企业管理实践和企业信息化。E-mail：shanyg@sia.cn

FSVM）中采用模糊决策法构造隶属函数对不可分区域分类，达到解决问题的目的，但该方法隶属函数构造不完备，分类精度提高有限。Platt 等人[6]在一对一算法的基础上提出了使用有向无环图拓扑结构设计来解决不可分区域。但该方法存在误差积累，错分率大。文献［7］使用迭代方法对不可分区样本迭代构建超平面，直到不可分区无样本，但该方法存在计算量大，并在极端情况如样本都是支持向量时易出现死锁问题。文献［8］使用动态投票法给分类器赋予权重来解决不可分区，权重根据统计设置，但权重不易获取。文献［9］阐明了样本的聚类特点，为本文提供了理论依据。文献［10］在模糊支持向量机中引入 kNN 作为隶属度，实验效果良好，具有启发性。

本文在分析比较多分类的基础上，提出了一种一对一多分类支持向量机与基于紧密度的结合的决策方法。该方法对 SVM 分类器的分类盲区构造样本到类中心之间的距离和基于 kNN 样本分布情况相结合的紧密度判别函数，选择紧密度最大的类为样本所属类。通过对 UCI 数据集测试实验表明了该方法有效提高了样本总体分类精度。

1 支持向量机

1.1 二元分类算法

对于一组带有类别标记的训练样本 $\{x_1, y_1\}$，$\{x_2, y_2\}$，…，$\{x_n, y_n\}$。其中每一个输入点 $x_i \in \mathbf{R}$，$y_i \in \{-1, 1\}$。支持向量机算法目的是构造一个分类超平面 $w \cdot x + b = 0$ 以分割两类不同的样本，使得两类间隔最大。即求解优化问题

$$\begin{cases} Min\ \dfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t.\quad y_i(w^T \cdot x_i + b) - 1 \geqslant 0, i = 1,2,\cdots,n \end{cases} \tag{1}$$

引入 Lagrange 函数求解，得到支持向量机（即判决函数）为

$$f(x) = sgn\left(\sum_{i=1}^{s}\alpha_i^* y_i (x \cdot x_i) + b^*\right) \tag{2}$$

式中：$\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*)$ ——拉格朗日乘子，$w^* = \sum_i \alpha_i^* y_i x_i$ ——权值，b^* ——分类阈值，可以用任一个支持向量求得

$$b^* = \frac{1}{2}[(w^* \cdot x(1)) + (w^* \cdot x(-1))]$$

式中：$x(1)$ ——两类中属于第一类的任意支持向量，$x(-1)$ ——属于第二类的任意支持向量。

对于线形不可分情况下，判决函数

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{s}\alpha_i^* y_i K(x, x_i) + b^*\right) \tag{3}$$

式中：$K(x, x_i)$ ——核函数 $x = (x_1, x_2, \cdots, x_s)$。决策过程如图 1 所示。

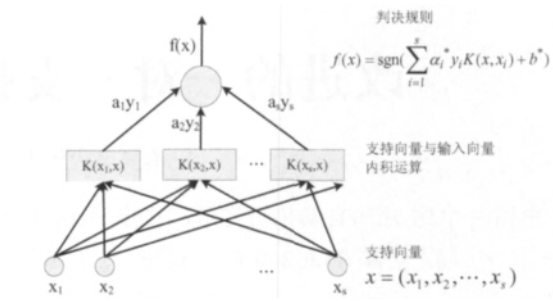

图 1 支持向量机决策过程

1.2 一对一多分类算法

一对一算法的基本思想是：对 k 个类的训练样本进行两两区分，共构造 k（k－1）/2 个 SVM 分类器。使用两类 SVM 分别求得 k（k－1）/2 个判别函数 f_{ij}（x）。组合这些两类分类器并使用投票法，得票最多的类为样本点所属的类。一对一与其它分类方法性能相比[11-12]，在算法训练时间上，一对一法与有向无环图法、二叉树法和纠错输出编码法相当，优于一对多法；在分类精度上，一对一法不需要进行拓扑结构设计，识别结果具有确定性，识别率在上述几个方法中最高；在样本类别不多的情况下，分类速度与有向无环图法、二叉树法和纠错输出编码法相差不多，好于一对多法。缺点在于一对一算法通过计算得票数，来得到最终判别，可能会出现不只一个类别得到最高票数的情况，导致不可分区域，使在此区域的样本不可分。

2 一对一算法不可分区域问题[5]

设第 i 类与 j 类分类函数是 $f^{ij}(x) = \sum_{n=1}^{l} y_n \alpha_n K(x_{ij}, x) + b^{ij}$，l 是第 i 类和 j 类的样本数。在分类过程中，如果输入样本满足条件 f^{ij}（x）＞0，则判定 x 属于第 j 类。但是当有多个 j 满足 f^{ij}（x）＞0，或者没有满足 f^{ij}（x）＞0 时，输入样本 x 不能被确切的定义为属于某一类时，出现了不可分。如图 2（a）所示。阴影部分的样本在 i、j、k 类的最终投票得分均为 1，根据投票结果无法准确决策此区域点的类别所属。

为了解决一对一算法的不可分区域问题，文献［5］在一对一算法的基础上引入模糊判别函数，在特征空间中构造 C（C－1）/2（C 为类别数量）个超平面后，引入决策函数 f^{ij}（x）作为隶属度函数，重新划分分类面。如果有多个类满足 f^{ij}（x）＞0，则 x 属于 f^{ij}（x）最大的类，如果有多个类满足 f^{ij}（x）＜0，则 x 属于 f^{ij}（x）绝对值最小的类。新的分类面如图 2（b）所示。该方法的实质是用支持向量作为类的代表点，判定样本到超平面的距离，样本属于距离最近的类，这种方法的缺点是只用少数样本（支持向量）作为类的代表点，没有用到多数训练样本分布情况，待分类样本 x 与支持向量距离远近并不总能够代表类别隶

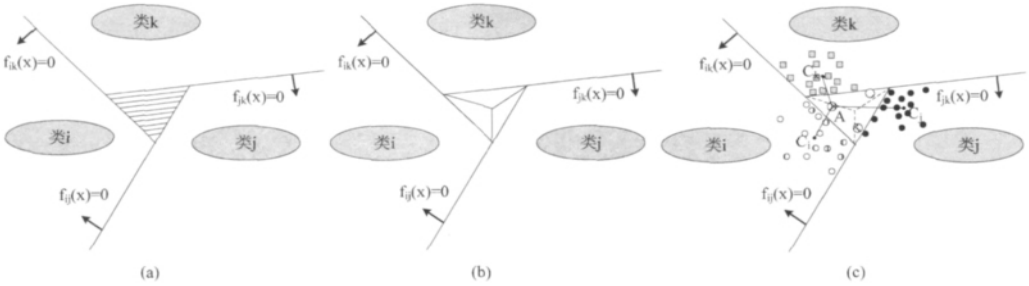

图 2 OVO 不可分区域

属度。如图 2（c）所示根据样本分布，样本 A 更可能属于类 k，按照模糊决策却属于 i，因此在新的分类面交界处附近的样本造成了错分。实际上，同类样本具有易聚类特性[9,13]。由于 kNN 技术成熟，已广泛应用。本文尝试用紧密度方法根据样本多数分布特征来判定样本所属类别来解决样本在不可分区的错分和漏分。

3 基于紧密度的分类

3.1 确定紧密度

对落入拒识区域中的样本点，可以用样本点到类中心的距离判定所属类别，该方法根据相似性原理，距离越小，两个样本可能越相似。但这种方法在样本集存在噪声和野值时，会影响分类精度。所以考虑依据样本到类中心之间的距离和样本分布情况相结合的方法构建判别函数来判别类别归属。这种方法既考虑样本与类中心之间的关系，还考虑了类中各个样本之间的关系。通过样本之间的紧密度来描述类中各个样本之间的关系，能有效提高分类精度。紧密度使用 kNN 方法描述。基于 kNN 紧密度方法是考虑待分样本的某一周围邻域内的训练样本密度来判定类别归属。

定义 1（样本欧式距离） 在训练样本集 $\{x_1, y_1\}$，$\{x_2, y_2\}$，…，$\{x_n, y_n\}$ 中，共有 c 个类别 v_1，v_2，…，v_c，两样本之间的欧式距离

$$d(x_i, x_j) = \| x_i - x_j \| = = \sqrt{\sum_{m=1}^{l} (x_{im} - x_{jm})^2} \quad (4)$$

式中：x_{im}、x_{jm}——测试样本、类中心的第 m 个特征属性。

定义 2（样本中心距离） 第 i 类样本的平均特征为该类样本的样本中心，$o_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j$，$n_i$ 为第 i 类的样本数量。任意样本点 x_j 到第 i 类样本中心的欧氏距离为

$$d(x_j, o_i) = \| x_j - o_i \| \quad (5)$$

定义 3（基于欧式距离类别判别） 样本点离各自类中心的最大距离为 $d_{i\_\max} = \max(d(x_j, o_i) \mid x_j \in \mathbf{R}^i)$。则样本 x 基于欧式距离类别判别函数为

$$s_i(x) = 1 - \frac{d(x, o_i)}{d_{i\_\max}} = 1 - \frac{\| x - o_i \|}{d_{i\_\max}}, i = 1, 2, \cdots, c \quad (6)$$

定义 4（样本紧密度） 基于加权 kNN 的对于不同类别的样本紧密度为

$$\mu_i(x) = \frac{\sum_{j=1}^{k} u_i(x^{(j)}) w_j}{\sum_{j=1}^{k} w_j} \quad (7)$$

式中：w_j——权系数，这里取距离的倒数

$$w_j = \frac{1}{d(x, x^{(j)})} = \| x - x^{(j)} \|^{-1}$$

则

$$\mu_i(x) = \frac{\sum_{j=1}^{k} u_i(x^{(j)})(\| x - x^{(j)} \|^{-1})}{\sum_{j=1}^{k} (\| x - x^{(j)} \|^{-1})} \quad (8)$$

式中：k——k 近邻的个数，$u_i(x)$——测试样本属于第 i 类训练数据的紧密度隶属度，$u_i(x^{(j)})$——第 j 个近邻属于第 i 个类别的隶属度，有 $u_i(x^{(j)}) = \begin{cases} 1, & x^{(j)} \in v_i \\ 0, & x^{(j)} \notin v_i \end{cases}$。

通过赋给 k 个近邻不同的加权系数，可确定样本属于各个类别的紧密程度。

定义 5（基于紧密度的判别函数） 样本中心距离与样本分布结合的分类判别函数为

$$g_i(x) = s_i(x) \times \mu_i(x) \quad (9)$$

通过计算测试样本属于各个种类的紧密度，找出紧密度值最高的类别，作为测试样本的预测结果。Max（$g_i(x)$），i=1，2，…c。

若样本集线性不可分，则可利用非线性映射 $\Phi: R^n \to H$ 把输入空间映射到特征空间，样本集在此特征空间线性可分

$$d^H = \sqrt{K(x,x) - 2K(x,y) + K(y,y)} \quad (10)$$

其中，K（x，y）= Φ（x）·Φ（y），K（x，y）是支持向量机中用到的核函数。

3.2 算法实施步骤

基于紧密度与一对一支持向量机结合的决策算法如下：

（1）设 T 为待识别集，k 为 kNN 的个数，C 为类别数。求解 m=C（C−1）/2 个支持向量机，得到 Langrage 乘子，常数 b，和支持向量，并构造 C（C−1）/2 个分

类器。

（2）如果 T≠Φ，取 x∈T，如果 T=Φ 停止。

（3）将测试样本分别送到 C（C－1）/2 判别函数 $f_{ij}(x)=\sum_{ij} w_{ij}K(x,x_{ij})-b$，若 $f^{ij}(x)=+1$，判样本 x 为 i 类，i 类得一票。统计各个类别最终得票数，计算 $D_i(x)=\sum_{j\neq i,j=1}^{N}\mathrm{sgn}(f_{ij}(x))$，其中，$\mathrm{sgn}(x)=\begin{cases}1, & x>0\\ 0, & x\leqslant 0\end{cases}$，得票最多的类为样本点所属的类，$\arg\max_{i=1,\cdots,N}(D_i(x))$。

（4）若 x（$\arg\max_{i=1,\cdots,N}(D_i(x))$ 类别号不唯一）不可分，按照式（9）计算，取紧密度最大的类为所属类。

（5）取下一个待检样本，T=T－｛x｝跳到步骤（2）。

4 实 验

测试实验选用了由美国 UCI[14] 提供的 4 个公共数据集 Auto mpg、Wine、Car 和 Satimage，涵盖了小样本和大样本情况，类别数多和少的情况，特征维数多和少情况，样本分布均匀和不均匀情况。对于 Satimage 直接对其训练数据进行训练，对测试数据进行测试。对于 Auto mpg、Wine 和 Car 由于所选用的数据集没有提供测试数据，将数据集随机地分成两份，一份作为训练数据，一份作为测试数据。数据集属性如表 1 所示。

表 1 UCI 数据集属性

数据	类别数	特征维数	训练集	测试集
Auto mpg	3	7	242	156
Wine	3	13	90	88
Car	4	6	958	770
Satimage	6	36	4435	2000

所有实验均在酷睿双核 2.0GHz，2GB 内存的微机上进行。实验中，支持向量机的核函数采用 RBF 核函数 $K(x,x_i)=\exp(-\gamma\|x-x_i\|^2)$，采用 SMO 方法进行训练。为了更好地解决不可分区域问题，采用交叉验证的方法获取合适的参数，使得不可分区域包含尽量多的样本数。支持向量机惩罚因子 C 和核函数参数 γ 两个参数分别拥有以下 10 种选择：$\gamma=[2^0, 2^1, \cdots, 2^{-10}]$，$C=[2^0, 2^1, \cdots, 2^{10}]$。在其中选择一组参数使落入不可分区域的样本足够多。k 近邻核函数选用径向基函数，σ 值与支持向量机核参数一致。实验步骤如图 3 所示。一对一算法、模糊支持向量机算法和有向无环图算法对不可分区的实验结果见表 2。

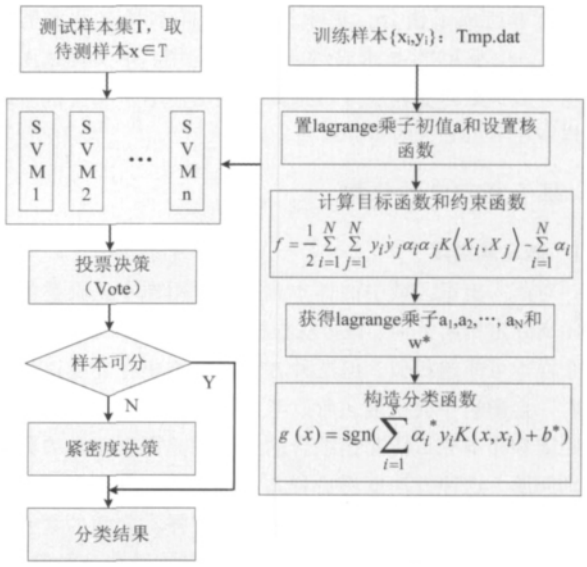

图 3 紧密度与一对一结合训练和分类流程

表 2 紧密度结合一对一法、模糊支持向量机法和有向无环图法的分类性能比较

数据集	C，γ	模糊支持向量机			紧密度结合一对一				有向无环图	
		不可分区精度	总体精度	分类速度/（一个样本/s）	k	不可分区精度	总体精度	分类速度/（一个样本/s）	总体精度	分类速度/（s）
Auto mpg	2^3，2^{-10}	0.45	0.71	0.30	60	0.85	0.82	0.31	0.65	0.24
Wine	1，2^{-5}	0.27	0.53	0.27	90	0.83	0.84	0.30	0.68	0.22
Car	2，2^{-1}	0.60	0.73	0.52	40	0.93	0.91	0.53	0.79	0.30
Satimage	2^4，2^{-10}	0.61	0.60	1.56	400	0.81	0.89	1.40	0.63	0.31

由实验结果可知紧密度分类算法不管在不可分区域的识别精度，总体识别精度都要比其它算法好。单个样本的分类速度与其它算法相差不多。并且，一对一结合紧密度方法对不同类型样本，如：Auto mpg 数据集属于小样本，wine 数据集样本分布不均匀，Satimage 数据集属于大样本，在各种参数设置下都有较高的识别率，表明算法具有很强的鲁棒性和健壮性。

支持向量机惩罚因子 C 和核参数 γ 值的影响不可分区大小。C 值取小值和 γ 取小值时，支持向量机具有较大边缘间隔的分类面，不可分区越大。一对一和 kNN 结合算法对各个测试集分类精度的提高程度不同，在训练样本数量多和属性少情况下，提高较明显，在训练样本数量少和属性多情况下提高幅度较小，这与样本分布状况和数据高维性问题[15]有关。采用基于核的最佳距离 K 近邻算法，对分布不均的样本集经过非线性映射后，变为线性可分，而且距离计算上减少了误差，从而提高了分类准确性。

kNN 算法的 k 值对性能有较大影响。一般 k 值的选取与最少类别训练样本数量关联，$k \leqslant 2\min(n_1, n_2, \cdots, n_c)$，c 为类别数，$n_{cw}$ 为 c 类样本数，k 取整数。以 car 数据集为例，最少类训练样本数为 40，测试 k 取不同值对不可分区分类结果如图 4 所示。分类时间随 k 值的增加依次增加。当 k＝20 分类精度较低，当 k 值在 40 到 60 之间分类精度较高，也比较稳定。当 k 大于 60 时近似误差较大，精度下降。

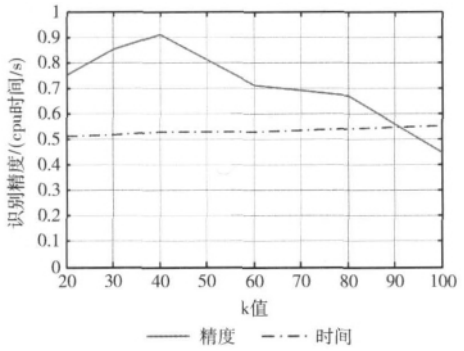

图 4　k 值与精度和时间的关系

5　结束语

在基于支持向量机的传统多分类算法中，一对一算法具有优秀的性能，但此算法在分类时却存在一个不可分区域。本文提出的一对一多分类支持向量机与紧密度决策结合的算法充分利用一对一多分类的性能，并且有效解决了不可分问题，提高了多分类的整体准确率。使用 UCI 数据集进行了测试，实例验证了该方法的实效性，识别率高，分类总体精度也优于模糊隶属度判决分类算法和有向无环图算法，分类时间并无明显增加。k 值的选择可对分类精度进行控制。下一阶段研究参数 k 的合理选择以及算法在传感器故障诊断中的应用。

参考文献：

[1] Vapnik V N. Statistical learning theory [M]. Beijing: PHEI, 2009.

[2] CUI Jianguo, GAO Jian. The application of support vector machine in pattern recognition [C]. 2007 IEEE International Conference on Control and Automation, 2008: 3135-3138.

[3] Maenhout S, De Baets B, Haesaert G, et al. Support vector machine regression for the prediction of maize hybrid performance [J]. Theoretical and Applied Genetics, 2007, 115 (7): 1003-1013.

[4] GOU Bo, HUANG Xianwu. SVM multi-class classification [J]. Journal of Data Acquisition & Processing, 2006, 21 (3): 334-339 (in Chinese). [苟博，黄贤武. 支持向量机多类分类方法 [J]. 数据采集与处理，2006，21 (3)：334-339.]

[5] Takuya I, Shigeo A. Fuzzy support vector machines for pattern classification [C]. Proceedings of International Joint Conference on Neural Networks, 2001: 1449-1454.

[6] Platt J C, Cristianini N, Shawe-Jaylor J. Large margin DAGs for multi-class classification [J]. Advances in Neural Information Processing Systems, 2000, 12 (3): 547-553.

[7] LIU Bo, HAO Zhifeng, XIAO Yanshan. Alternating iterative one-against-one algorithm [J]. Pattern Recognition and Artificial Intelligence, 2008, 21 (4): 425-431 (in Chinese). [刘波，郝志峰，肖燕珊. 交互迭代一对一分类算法 [J]. 模式识别与人工智能，2008，21 (4)：425-431.]

[8] WANG Xiaohong. A novel decision-making method for multi-class SVMs and its application [J]. Information and Control, 2008, 37 (6): 647-652 (in Chinese). [王晓红. 一种新的多类支持向量机决策方法及其应用 [J]. 信息与控制，2008，37 (6)：647-652.]

[9] SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research [J]. Journal of Software. 2008, 19 (1): 48-61 (in Chinese). [孙吉贵，刘杰，赵连宇. 聚类算法研究 [J]. 软件学报，2008，19 (1)：48-61.]

[10] TAO Xinmin, XU Jing, DU Baoxiang, et al. A FSVM based on affinity and its application in bearing fault detection [J]. Journal of Vibration Engineering, 2009, 22 (4): 418-424 (in Chinese). [陶新民，徐晶，杜宝祥，等. 基于紧密度 FSVM 新算法及在故障检测中的应用 [J]. 振动工程学报，2009，22 (4)：418-424.]

[11] HAO Zhifeng, LIU Bo, YANG Xiaowei. A comparision of multiclass support vector machine algorithms [C]. Proceedings of the International Conference on Machine Learning and Cybernetics, 2006: 4221-4226.

[12] YAN Zhigang, DU Peijun. Generalization performance analysis of M-SVMs [J]. Journal of Data Acquisition & Processing, 2009, 24 (4): 469-475 (in Chinese). [闫志刚，杜培军. 多类支持向量机推广性能分析 [J]. 数据采集与处理，2009，24 (4)：469-475.]

[13] Sambasivam S, Theodosopoulos N. Advanced data clustering methods of mining Web documents [J]. Issues in Informing Science and Information Technology, 2006 (3): 563-579.

[14] Asuncion A, Newman D J. UCI machine learning repository [DB/OL]. [2009-08-01]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[15] Hans-Peter K, Kruger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3 (1): 1-58.