研究论文

基于模糊核聚类的多类支持向量机

曹 巍[1,2]，赵英凯[2]，高世伟[1]
([1] 中国石油兰州石化自动化研究院，甘肃 兰州 730060；[2] 南京工业大学自动化学院，江苏 南京 210009)

摘要：传统的支持向量机是基于两类问题提出的，如何将其有效地推广至多类问题仍是一个值得研究的问题。本文在比较常用的几种多类支持向量机分类算法基础上，提出了一种基于模糊核聚类的多类支持向量机分类方法。支持向量机的分类精度和分类速度取决于树结构，新方法利用模糊核聚类生成模糊类，并结合基于二叉树的多类支持向量机分类算法实现多类分类。实验结果表明，该方法是一种效率更高、分类更准确的多类支持向量机分类算法。

关键词：支持向量机；多类分类；模糊核；二叉树
中图分类号：TP 181 **文献标识码**：A **文章编号**：0438－1157 (2010) 02－0420－05

Multi-class support vector machines based on fuzzy kernel cluster

CAO Wei[1,2], ZHAO Yingkai[2], GAO Shiwei[1]
([1] *Automation Institute of Lanzhou Petrochemical Company, PetroChina, Lanzhou* 730060, *Gansu, China*;
[2] *School of Automation, Nanjing University of Technology, Nanjing* 210009, *Jiangsu, China*)

Abstract: Traditional support vector machines (SVM) is originally designed for binary classification. How to effectively extend it to multi-class classification is worthy to research. This paper compared some common support vector machines for multi-class classification problems, and proposed a multi-class support vector machine based on fuzzy kernel clustering algorithm. The classification accuracy and classification speed of support vector machine depended on the tree structure. This multi-class support vector machine used fuzzy kernel clustering algorithm to generated fuzzy class, and combined with the multi-class SVM based on binary tree classification algorithm for multi-class classification. Experimental results showed that the proposed method was more effective and accurate.

Key words: support vector machines; multi-class classification; fuzzy kernel; binary tree

引 言

机器学习是现代智能技术中十分重要的研究领域，它通过对已知数据的学习，找到数据内在的相互依赖关系，从而获得对未知数据预测和对其性质的判断能力[1]。支持向量机（support vector machine，SVM)[2]是数据挖掘中的一项新技术，是借助于最优化方法解决机器学习问题的新工具，它是结构风险最小化方法的近似实现。通过学习，SVM 可以自动寻找那些对分类有较好区分能力的支持向量，由此构造出的分类器可以最大化类之间的间隔，具有较好的推广能力和较高的分类准确率。SVM 在化工领域也有非常广泛的应用[3-4]。SVM 本身是一个两类分类算法，如何将其推广到

2009－10－20 收到初稿，2009－10－28 收到修改稿。
联系人及第一作者：曹巍（1966—），男，硕士，教授级高工。
基金项目：国家高技术研究发展计划项目（2006AA040309）。
Received date: 2009－10－20.
Corresponding author: CAO Wei, caow-ls@petrochina.com.cn
Foundation item: supported by the High-tech Research and Development Program of China (2006AA040309).

多类分类问题，以适应实际应用的需要具有十分重要的意义。本文首先对比分析了几种传统多类 SVM 方法各自的优缺点，然后提出了一种基于模糊核聚类的多类 SVM 分类算法。模糊核聚类通过非线性映射能较好地分辨、提取并放大有用的特征，挖掘更多的细节信息，从而能实现更为准确的聚类。新方法利用模糊核聚类生成模糊类，并结合基于二叉树的多类 SVM 分类算法实现多类分类。

1　多类 SVM

SVM 具有完备的统计学习理论基础，它采用结构风险最小化原则代替传统统计学中的基于大样本的经验风险最小化原则，克服了神经网络受到网络结构复杂性和样本容量的影响大，容易出现过学习或低泛化能力的不足，对于小样本数据分析具有出色的学习能力和推广能力，在模式识别和函数估计中得到了有效的应用。SVM 是一种两类分类器，如何将 SVM 有效地应用于多类分类成为研究的热点。

当前已有的多类 SVM 分类方法大致可分为两种：一次性求解算法和分解重构算法。一次性求解法[5]是在所有训练样本上求解一个大型二次规划问题，同时将所有类别分开。该方法变量个数多，计算复杂度很高，尤其当类别数目较多时，它的训练速度很低，分类精度也不高；分解重构法是一种将多类分类问题转化为多个两类分类问题，并采用某种策略将多个两类分类器组合起来实现多类分类的方法。分解重构法比一次性求解法更适合于实际应用，用它实现多类分类需要解决两个关键问题：模糊类的生成和多个两类分类器的组合策略。当前应用较广泛的 SVM 分解重构算法有 1-a-r（one-against-rest）方法[6]、1-a-1（one-against-one）方法[7]、DAGSVM（directed acyclic graph support vector machines）方法[8]、DT-SVM（decision-tree-based multiclass support vector machines）方法[9-11]以及 H-SVM 方法（hierarchical support vector machines)[12-13]等。

1-a-r 采用最大输出法将多个分类器的输出组合起来实现多类分类，在 1-a-r 分类方法中对 n 个类别仅需构造 n 个 SVM，每个 SVM 分别将某一类的数据从其他类别中分离出来。在测试时，取决策函数输出值最大的类别为测试样本的类别。1-a-r 分类方法简单、有效，可用于大规模数据。但当类别数较大时，某一类的训练样本将大大少于其他类训练样本的总和，这种训练样本间的不均衡将对精度产生影响。由于每次构造分类器都要将整个工作集作为训练样本，当工作集过大时，训练速度将会很慢。同时它存在误分、拒分区域，泛化能力较差。

1-a-1 采用投票法决定未知样本的类别，在 1-a-1 分类方法中，各个类别之间构造分类器，对 n 个类别共需构造 $n(n-1)/2$ 个分类器，每个分类器函数的训练样本是相关的两个类，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。由于每个 SVM 只考虑两类样本，故单个 SVM 容易训练，且其决策边界较 1-a-r 简单；另外，虽然它的复杂度以类数按平方增长，但就分类速度来说，并不比传统的 1-a-r 方法慢；而且其分类精度也较 1-a-r 高。它的缺点是：如果单个两类分类器不规范，则整个分类器将趋向于过学习；分类器的数目随类数急剧增加，导致在决策时速度很慢；存在推广误差无界及误分、拒分区域。1-a-r 和 1-a-1 这两种方法分类时都需要遍历所有的 SVM 分类器，因而识别效率低。

DAGSVM 是针对 1-a-1 存在的拒分现象提出的，算法在训练阶段与 1-a-1 相同，也要构造每两类间的分类面，即有 $n(n-1)/2$ 个分类器。但在分类阶段，该方法将所有分类器构成一个两向有向无环图，包括 $n(n-1)/2$ 个节点和 n 个叶。其中每个节点为一个分类器，并与下一层的两个节点相连。当对一个未知样本进行分类时，首先从顶部的根节点开始，根据根节点的分类结果用下一层的左节点或右节点继续分类，直到达到底层某个叶为止，该叶所表示类别即为未知样本的类别。DAGSVM 简单易行，对 n 类问题，只需使用 $n-1$ 个决策函数即可得出结果，较 1-a-1 方法提高了测试速度，而且不存在拒分区域；另外，由于其特殊的结构，故有一定的容错性，分类精度较一般的二叉树方法高，但该方法的泛化能力与各子分类器在有向无环图中的位置有关。DT-SVM 和 H-SVM 也是采用树结构的组合策略，具有较高的训练和分类速度，但是它们生成的模糊类交叠严重，而分类树又存在错分积累，因而分类精度较低。

基于二叉树的 SVM 多类分类算法是将所有类别分为两个子类，每个子类又划分为两个子子类，如此循环，直到划分出最终类别，每次划分后两类

分类问题的规模逐级下降。这样得到一个倒立的二叉树，每个决策点用 SVM 实现分类。二叉树分类的优点是不存在不可分区域，分类时不一定需要遍历所有的分类器，测试时间较短，有较高的分类效率。但是，二叉树的结构对其推广能力影响很大，对于同一多类分类问题不同的二叉树结构分类性能也不尽相同。二叉树分类算法解决了传统分类算法中的不可分区域问题，然而二叉树的结构对分类效果影响很大，不同的二叉树结构其推广能力也不相同。如何确定二叉树的生成规则，使分类器的性能有所提高，是问题的关键。在数据集非线性可分情况下，SVM 在进行分类的过程中，需要将样本数据映射到特征空间。因此，在特征空间中计算类间距离，能够更真实地反映类间距离情况。

模糊类生成方法决定了两类分类器的个数、各两类分类器的训练样本数目和模糊类间的交叠程度，进而对多类分类器的训练速度、分类速度和精度有较大影响。本文提出一种基于二叉树的 SVM 多类分类方法，该方法基于模糊核聚类算法反复将训练样本集划分为两个子集，直到每个子集都只包含一个类成员。基于隶属度，每次从两个子集中选择可分性强的子类为树的当前结点定义分类子任务，将可分性弱的子类移至下层结点，使得树的上层结点具有好的分类性能，从而降低树结构的错分积累，提高整个二叉树的泛化性能。

2 模糊核聚类

聚类是模式识别和数据挖掘中广为使用的数据分析手段，是将物理或抽象的对象按照对象间的相似性进行区分和分类的过程。模糊 C-均值（fuzzy C-means，FCM）聚类算法[14]是模糊聚类算法中应用最广泛的典型形式。相对传统的聚类算法，模糊聚类算法由于引入了模糊集理论，因此有更好的数据表达能力与聚类性能。FCM 算法的目的在于，将向量空间的样本点按照某种距离度量划分成子空间。可以不需要训练样本，直接通过机器学习达到自动分类的目的。FCM 算法抗噪性能较差，而且忽略了数据集对应几何空间的位置不同拥有的特殊性，对于数据集中含有不均等类的情况下极易形成误判。

将经典的聚类算法推广到核 Hilbert 空间，是无监督学习领域近年来的研究热点之一。模糊核聚类依照核方法思想，首先用非线性映射把输入空间的数据映射到高维特征空间，扩大模式类之间的差异，然后在特征空间中对数据进行模糊聚类的方法[15]。核聚类方法在性能上比经典的聚类算法有较大的改进。它通过非线性映射能够较好地分辨、提取并放大有用的特征，从而实现更为准确的聚类，算法收敛速度也较快。

依照核方法的思想，用非线性映射 $\phi(\cdot)$ 把输入模式矢量空间变换到一个高维特征空间，在该特征空间扩展 FCM 算法，对变换后的特征矢量 $\phi(x_i)$（$i=1, 2, \cdots, N$）进行模糊聚类分析。

设原空间样本集为 $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$，$\boldsymbol{x}_j \in R^d$，$j=1, 2, \cdots, N$，$C$ 是事先确定的簇数，$m \in (1, \infty)$ 是模糊加权指数，对聚类的模糊程度有重要的调节作用；核非线性映射为 $\phi: x \rightarrow \phi(x)$，若在高维特征空间采用 Euclid 距离，则模糊核 C-均值聚类的目标函数为

$$J_{\mathrm{m}}(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} \left\| \phi(\boldsymbol{x}_j) - \phi(\boldsymbol{v}_i) \right\|^2 = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} [K(\boldsymbol{x}_j,\boldsymbol{x}_j) - 2K(\boldsymbol{x}_j,\boldsymbol{v}_i) + K(\boldsymbol{v}_i,\boldsymbol{v}_i)] = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} d_{Kij}^{2}(\boldsymbol{x}_j,\boldsymbol{v}_i) \quad (2 \leqslant C < N) \tag{1}$$

按模糊 C-均值算法优化方法，隶属度应满足

$$u_{ij} = [1/d_{Kij}^{2}(\boldsymbol{x}_j,\boldsymbol{v}_i)]^{1/(m-1)} \Big/ \sum_{j=1}^{C} [1/d_{Kij}^{2}(\boldsymbol{x}_j,\boldsymbol{v}_i)]^{1/(m-1)} \tag{2}$$

式中 $\boldsymbol{v}_i$ 为第 i 类的类中心，$\phi(\boldsymbol{v}_i)$ 为该中心在相应核空间中的像，且有

$$\phi(\boldsymbol{v}_i) = \sum_{k=1}^{N} u_{ik}^{m}\phi(\boldsymbol{x}_k) \Big/ \sum_{k=1}^{N} u_{ik}^{m} \quad (i = 1,2,\cdots,C) \tag{3}$$

为最小化目标函数，需要计算 $K(\boldsymbol{x}_j, \boldsymbol{v}_i)$ 和 $K(\boldsymbol{v}_i, \boldsymbol{v}_i)$，由 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)=\langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$ 可得到

$$K(\boldsymbol{x}_j,\boldsymbol{v}_i) = \langle\phi(\boldsymbol{x}_j),\phi(\boldsymbol{v}_i)\rangle = \sum_{k=1}^{N} u_{ik}^{m} K(\boldsymbol{x}_k,\boldsymbol{x}_j) \Big/ \sum_{k=1}^{N} u_{ik}^{m} \tag{4}$$

$$K(\boldsymbol{v}_i,\boldsymbol{v}_i) = \langle\phi(\boldsymbol{v}_i),\phi(\boldsymbol{v}_i)\rangle = \sum_{k=1}^{N}\sum_{s=1}^{N} u_{ik}^{m} u_{is}^{m} K(\boldsymbol{x}_k,\boldsymbol{x}_s) \Big/ \Big(\sum_{k=1}^{N} u_{ik}^{m}\Big)^2 \tag{5}$$

模糊核聚类算法的步骤为：

（1）设置迭代停止条件 ε、模糊指数 m、迭代次数 T、聚类数 C；

（2）对样本集作归一化处理，并选择核函数及合适的参数；

（3）初始化类中心 $\boldsymbol{v}_i$（$i=1, 2, \cdots, C$）；

(4) 计算每个样本在特征空间的隶属度 u_{ij} ($i=1,2,\cdots,C$；$j=1,2,\cdots,N$)；

(5) 计算新的 $K(\boldsymbol{x}_j,\boldsymbol{v}_i)$ 和 $K(\boldsymbol{v}_i,\boldsymbol{v}_i)$，更新隶属度 u_{ij} 为 $\hat{u}_{ij}$；

(6) 若 $\max\limits_{j,i}|u_{ij}-\hat{u}_{ij}|<\varepsilon$ 或迭代次数等于预定迭代次数 T 则算法停止，否则转到步骤 (4)。

3　SVM 二叉树的层次构造

二叉树泛化性能的好坏与树层次结构紧密相关，分类错误发生的地方距二叉树根结点越近，二叉树的错分积累就会越大。如果直接采用模糊核 C-均值生成的子集作为 SVM 的正负类定义树结点的分类子任务，若子集间的可分性差且当前结点处于二叉树的顶层，则会导致整个二叉树具有较高的错分积累。

根据隶属度 u_{ij} 的定义，可以得出如下结论：当类 j 与第 i 个模糊类中心的距离越近，与其他模糊类中心的距离越远时，则 u_{ij} 越大，即类 j 属于第 i 个模糊类的程度越大。假定对一个多类样本集按照上述模糊核 C-均值算法进行粗分得到两个子集 C_1 和 C_2，若设定阈值 T_i，从 C_i 中仅仅选择满足 $u_{ij}>T_i$ 的成员类组成相应子类 S_1 和 S_2，则能有效地减小分类子任务的规模，使得 S_1 和 S_2 间的可分性强于 C_1 和 C_2 间的可分性。各子类中成员类的隶属度越大，各子类间的可分性越强。

令 u_i 和 σ_i 分别表示 C_i 中所有成员类关于第 i 个模糊类的隶属度的均值和方差

$$u_i=\frac{1}{n_i}\sum_{\substack{j=1\\ w\in C}}^{n}u_{ij},\ \sigma_i^2=\frac{1}{n_i}\sum_{\substack{j=1\\ w\in C}}^{n}(u_{ij}-u_i)^2 \tag{6}$$

其中，n_i 为 C_i 中的成员类个数。

定义阈值

$$T_i=u_i-\alpha\sigma_i \tag{7}$$

其中，α 为常数因子，基于式 (6)、式 (7) 可以自适应地从各个子集 C_i 中选择隶属度大于 T_i 的成员类组成各子类 S_i。α 越小，则 T_i 越大，各子类间的可分性越强；α 越大，则 T_i 越小，各子类间的可分性越弱。

SVM 二叉树计算法如下。

指定初始训练集 R 为 R_0，R_0 表示原始的多类样本集。

(1) 粗划分　基于模糊核 C-均值算法对训练集进行粗划分得到子集 C_1 和 C_2。

(2) 细划分　根据式 (6)、式 (7) 定义阈值 T_i。对 $\forall w_j\in C_i$，若 $u_{ij}\geqslant T_i$，则将类 w_j 划分到子类 S_i，同时将 w_j 从 C_i 中去除。

(3) 构造二叉树结点　将 S_1 和 S_2 分别作为正负类，训练得到 1 个两类 SVM 分类器。若 R 为 R_0，则将该 SVM 作为二叉树的根结点；若 R 为 R_1，则将该 SVM 作为当前结点的左孩子；否则，作为当前结点的右孩子。

(4) 对训练集进行更新

$$R_i=S_i\cup C_i\cup C_j$$

(5) 对 R_1 和 R_2 重复执行上述步骤，直到每个训练集都只包含 1 个类。

值得注意的是，在细划分中定义阈值 T_i 时应根据不同子集的大小选择相应的 α，α 的大小影响着子类之间的可分性和二叉树的高度。对于含有较多成员类数量的子集，应选择较小的 α 以增大阈值，减少子类含有的成员类数量，从而减小分类子任务的规模，增加子类间的可分性；反之，应选择较大的 α 以减小阈值，避免分类子任务规模过小而导致过度增加树的高度。

4　实验结果及分析

实验采用机器学习数据集 (UCI Machine Learning Reposity) 中的 Letter 数据集对算法进行测试，该数据集含有 26 个类别，每个样本的特征维数为 16，共 20000 个样本[16]。实验中选取 15000 个样本组成训练样本集，剩下的 5000 个样本组成测试样本集。模糊核 C-均值采用 $\gamma=2$ 的 RBF 核函数，基于子集中含有的成员类数量 $\Omega(\Omega<26)$ 的大小自适应地选取 α：$\Omega>15$，则 $\alpha=0.5$；$\Omega<10$，则 $\alpha=2$；其他情况 $\alpha=1$。将本文方法与 1-a-1、1-a-r 和 DAGSVM 算法 3 种多类分类方法进行了比较，所有 SVM 均采用 RBF 核函数，采用 10 折交叉验证法选择参数，核参数 γ 和惩罚系数 λ 的选择范围分别为 $\gamma=[2^5,2^4,\cdots,2^{-3}]$，$\lambda=[2^{10},2^9,\cdots,2^{-2}]$。采用了 Lin 等[17]开发的支持向量机工具箱 LIBSVM。算法在 VC++6.0 上实现，实验平台为 AMDAthlon 64 X2 5000+，2G。

表 1 列出了各种算法的支持向量数、分类阶段需要遍历的 SVM 个数、分类精度以及测试时间，支持向量数为所有 SVM 含有的唯一支持向量总数。实验结果表明，本文基于模糊核聚类的多层次 SVM 分类树大大降低了树结构的错分积累，提高

了分类精度。由于实现了更为准确的分类子任务层次的定义，本文算法的支持向量数和分类阶段需要遍历的 SVM 个数都很少，因此有效地加快了分类速度。

表 1 不同算法的支持向量数、分类需要遍历的 SV 个数、SVM 个数、分类精度、误分样本数以及测试时间

Table 1 SVs, SVMs, classification accuracy, error number of samples and test time

Algorithm	SV number	SVM number	Classification accuracy /%	Error number of samples	Test time/s
1-a-1	7644	325	97.86	107	21.216
1-a-r	7013	26	97.5	125	19.804
DAG	7646	25	97.8	110	11.657
new algorithm	6996	<26	98.06	97	7.856

5 结 论

本文基于模糊核 *C*-均值算法中隶属度衡量类间可分性，提出了一种具有良好泛化性能的 SVM 二叉树多类分类方法。实验中无论是在识别精度上还是在速度上都获得了令人满意的效果，与其他传统方法的比较体现了本文方法的高效性。

References

[1] Erin L Allwein, Robert E Schapire, *et al*. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 2000 (1): 113-141

[2] Vapnik V. Nature of Statistical Learning Theory. New York: Springer-Verlage, 1995

[3] Dai Bo (戴波), Zhao Jing (赵晶), Zhou Yan (周炎). Ultrasonic in-line inspection of pipeline corrosion based on support vector machine. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2008, **59** (7): 1812-1817

[4] Song Xiaofeng (宋晓峰), Yu Huanjun (俞欢军), Chen Dezhao (陈德钊), Hu Shangxu (胡上序). Modeling delayed coking process by adaptive support vector machine. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2004, **55** (1): 147-150

[5] Angulo C, Parra X, Català A K-SVCR. A support vector machine for multi-class classification. *Neurocomputing*, 2003, **55** (1/2): 55-77

[6] Bottou L, Cortes C, Denker J, *et al*. Comparison of classifier methods: a case study in handwriting digit recognition//Shmuel Peleg, Shimon Ullman. Proceedings of the 12th International Conference on Pattern Recognition. Los Alamitos, California: IEEE Computer Society Press, 1994: 77-87

[7] Krebel U. Pairwise classification and support vector machines//Bernhard Schoölkopf, Christopher J C Burges, Alexander J Smola. Advances in Kernel Methods. Cambridge, MA: MIT Press, 1999: 255-268

[8] Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification//Sara A Solla, Todd K Leen, Klaus-Robert Müller. Advances in Neural Information Processing Systems 12. Cambridge, MA: MIT Press, 2000: 547-553

[9] Takahashi F, Abe S. Decision-tree-based multiclass support vector machines//Wang Lipo. Proceedings of the 9th International Conference on Neural Information Processing. Singapore: IEEE Press, 2002: 1418-1422

[10] Sungmoon C, Sang H O, Soo-Young L. Support vector machines with binary tree achitecture for multi-class classification. *Neural Information Processing-Letters and Reviews*, 2004, **2** (3): 47-51

[11] Bennett K, Blue J. A support vector machine approach to decision trees [R]. Rensselaer Polytechnic Institute, Troy, NY: R. P. I Math Report, 1997: 97-100

[12] Schwenker F. Hierarchical support vector machines for multi-class pattern recognition//Howlett R J, Jain L C. Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies. Brighton, UK: Institute of Electrical & Electronics Engineers, 2000: 561-565

[13] Weston J, Watkins C. Support vector machines for multi-class pattern recognition//Michel Verleysen. Proceeding of the 7th European Symposium on Artificial Neural Networks. Bruges, Belgium: D-Facto Publications, 1999: 219-224

[14] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981

[15] Wu Zhongdong (伍忠东), Gao Xinbo (高新波), Xie Weixin (谢维信). A study of a new fuzzy clustering algorithm based on the kernel method. *Journal of Xidian University* (西安电子科技大学学报), 2004, **31** (4): 533-537

[16] Asuncion A, Newman D J. UCI machine learning repository [EB/OL]. [2009-05-06]. http://www.ics.uci.edu/~mlearn/MLRepository.html

[17] Chang C, Lin C J. LIBSVM: a library for support vector machines [EB/OL]. [2009-04-08]. http://www.csie.ntu.edu.w/cjlin/libsvm