第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)18—0193—03 文献标识码：A 中图分类号：TP391

基于核聚类的手写金融汉字识别方法

陈增照[1,2]，杨 扬[1]，何秀玲[1,2]，喻 莹[1,2]，董才林[2]
(1. 北京科技大学信息工程学院，北京 100083；2. 华中师范大学最优控制与离散数学重点实验室，武汉 430079)

摘 要：根据手写体金融汉字的特点，利用核聚类方法将原始样本特征映射到高维特征进行聚类分组，对每一组使用一个支持向量机二值分类器进行分类，并用这些二值分类器组成决策树的结点，构成一个决策分类树。给出了金融汉字的分组方法和决策树的生成算法，提出利用交叠系数来控制交叠，可以克服错分积累，提高分类准确率。实验结果表明，采用该方法，手写体金融汉字识别的速度和正确率都达到了实用的要求。

关键词：手写体汉字识别；支持向量机；决策分类树；核聚类

Handwritten Financial Chinese Characters Recognition Approach Based on Kernel Clustering Algorithm

CHEN Zeng-zhao[1,2], YANG Yang[1], HE Xiu-ling[1,2], YU Ying[1,2], DONG Cai-lin[2]
(1. School of Information Engineering, University of Science & Technology Beijing, Beijing 100083;
2. Center for Optimal Control & Discrete Mathematics, Central China Normal University, Wuhan 430079)

【Abstract】 According to the characteristics of handwritten financial Chinese characters, original features are mapped to higher dimension by applying kernel clustering. Each group is classified by a support vector machines (SVM) classifier. These binary classifiers are seen as the nodes of decision tree, and construct a decision classifying tree. The algorithm of grouping financial Chinese characters and creating decision tree is given. The method of controlling overlap by overlap coefficient is proposed and it can overcome misclassification accumulation. Experimental results show that, with this approach, the speed rate and accuracy rate of recognition meet the requirements.

【Key words】 handwritten Chinese character recognition; support vector machines(SVM); decision classifying tree; kernel clustering algorithm

手写体金融汉字的识别是金融票据处理中最困难的问题之一，目前很多票据处理系统都因手写体金融汉字的识别率太低而放弃对其的处理。手写体金融汉字在金融票据中(如支票的大写金额与日期)是非常关键的要素，金融机构迫切希望对其进行识别处理。

在银行票据处理系统中，对金融汉字的识别速度和准确率要求比较高。近年来对手写体金融汉字识别虽然有很多研究，但往往存在如下缺点：(1)实验所用的样本集少，或者样本是请专人所写，没有从实际的银行票据中提取，这样虽然在实验中识别率达到了要求，但实际应用中识别效果并不理想；(2)识别率太低；(3)识别速度慢，一些系统的识别速度常常只有几个字/s，这些都达不到银行票据处理系统的要求。

本文采用核聚类的方法[1]，把输入空间的样本映射到高维特征空间，在特征空间中进行聚类，将手写体金融汉字分为不同的组(模糊类)，并采用决策分类树结构将多个 2 类分类器组合起来实现多类分类。核聚类方法能够较好地分辨、提取并放大有用的特征，实现更为准确的聚类，利用决策分类树结构大大提高分类的速度，同时在生成决策分类树时，引入交叠系数δ控制子类的交叠(即同一个原始类别同时划分到 2 个子类中)，从而在一定程度上克服了使用决策树时存在的错分积累，使得手写体金融汉字识别的速度与准确率都有明显的提高。

1 支持向量机分类器

支持向量机(support vector machine, SVM)是基于Vapnik等人建立的以解决有限样本机器学习问题为目标的统计学习理论(SLT)发展起来的一种新的学习机器[2]，它是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以获得最好的推广能力。由于其在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，使其在许多领域都获得了良好的应用。但经典的SVM是针对 2 类问题的分类而提出的，实现对多类问题的分类识别方法大致有 2 种：(1)扩展SVM的基本算法[2,3]，即同时考虑多个类别的样本，用一个二次优化问题描述整个多类别问题；(2)构造多个二值分类器(binary classifier)，对二值分类器的分类结果进行比较分析并得出最后的结果[2,4]。

扩展 SVM 方法因计算复杂度过高、精度较低而不实用。利用二值分类器组合的方法主要有 1-a-1 (one-against-one)方法、1-a-r(one-against-rest)方法以及构造多层决策分类树的方法等。

对于N类问题(N>2)，1-a-1 方法需要构造$N(N-1)/2$ 个 2 类SVM分类器[2,4]，将其他的每个类别单独作为反例进行训练，从而每个类别对应N-1 个分类器，在分类过程中，采用投票法统计所有分类器的分类结果，选择得票数最多的类别作为待分类对象应属的类别。这种算法的缺点是分类器数目大、

基金项目：湖北省科技攻关计划基金资助项目(2003BDST004)
作者简介：陈增照(1971－)，男，博士研究生、副研究员，主研方向：模式识别与图像处理；杨 扬，教授、博士生导师；何秀玲、喻 莹，博士研究生；董才林，博士、副研究员
收稿日期：2006-09-25 **E-mail：**llczz@tom.com

分类速度慢。例如，常用的手写体金融汉字有 21 个，共需要构造 210 个分类器，每一次识别都需要进行 210 次分类，并对分类结果进行投票，使得分类速度非常慢(往往只有几个字/s)，达不到实用要求。

1-a-r方法是将N类问题转化为N个 2 类问题[2,4]，即在训练某一类别时，将剩余N-1 个类别的所有样本作为它的反例。这种方法在训练每个分类器时需要优化所有N个训练样本，并且每个二值分类器的反例太多导致识别精度不高。

多层决策分类树的方法近年来得到了广泛的应用。该方法将原有的多类问题分解成了一系列的 2 类分类问题，分类树中的每个节点都是完成一个预定义的分类子任务的 2 类分类器[5]。这类方法的分类器结构简单清晰，可根据样本特征向量的特性定义子任务，推广误差只取决于类和节点上的类间间隙。同时，算法所需的基本 2 类分类器数目远小于 1-a-1 算法所需的 2 类分类器数目，而且也不需要计算庞大的拓展优化问题，因此，算法的速度也较快。可见，合理定义分类子任务的层次，多层决策分类树算法与前面几种算法相比具有明显的优势。本文提出的就是一种基于核聚类方法预定义子任务的多层次SVM分类树。

2 基于核聚类的手写体汉字分类方法

聚类属于非监督模式识别问题，其特点是输入空间的样本没有期望输出。对聚类问题，按照某种相似程度的度量，使相似的样本归为一类，而不相似的样本归于不同的类，即聚类的过程完全依赖于样本之间的特征差别。比较经典的聚类方法有传统的 C-均值方法和模糊 C-均值聚类方法，这些方法都没有对样本的特征进行优化，而是直接利用样本的特征进行聚类，因此，这些方法的有效性很大程度上取决于样本的分布情况。对于手写体汉字来说，样本分布混乱，聚类的结果就会面目全非。

支持向量机的独特之处在于它将输入空间中非线性可分的样本映射到高维空间，使其变得线性可分。向高维空间的非线性映射，实际上起到了分辨、提取并放大有用特征的作用。从预分类的阶段便引入非线性映射，在预定义子分类任务时采用基于Mercer核的聚类方法[6] (简称核聚类方法)，应用统计学习理论的有关思想，整个算法实际上在高维空间完成，使其性能得到提高。

假设输入样本为 $x_k \in R^n, k=1,2,...,l$，被某一非线性映射 Φ 映射到一个高维的核空间 H 中，得到 $\Phi(x_1),\Phi(x_1),...,\Phi(x_l)$，则输入空间的点积在高维核空间可以用 Mercer 核表示为

$$K(x_i,x_j)=\Phi(x_i)\bullet\Phi(x_j) \tag{1}$$

由所有的样本组成的核函数矩阵 $K_{i,j}=K(x_i,x_j)$，高维空间的 Euclidean 距离可表示为

$$\begin{aligned} d_H(x,y)&=\sqrt{\|\Phi(x)-\Phi(y)\|^2}\\ &=\sqrt{\|\Phi(x)\bullet\Phi(x)-2\Phi(x)\bullet\Phi(y)+\Phi(y)\bullet\Phi(y)\|} \end{aligned} \tag{2}$$

一般情况下，非线性函数的表达式是未知的，因此，由式(1)、式(2)可知

$$d_H(x,y)=\sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)} \tag{3}$$

将式(3)作为聚类相似性的度量函数。聚类的准则是最小化下列的目标函数：

$$J=\sum_{i=1}^{C}\sum_{\substack{j=1\\x\in C_i}}^{N_i}\left[K(x_j,x_j)-\frac{2}{N_i}\sum_{k=1}^{N_i}K(x_j,x_k)+\frac{1}{N_i^2}\sum_{k,p=1}^{N_i}K(x_k,x_p)\right] \tag{4}$$

其中，C为聚类的总类别数；N_i是第C_i类样本的个数，该类中心的模为

$$\|W_i\|^2=\frac{1}{N_i^2}\sum_{k,p=1}^{N_i}K(x_k,x_p) \tag{5}$$

根据式(3)、式(4)，可以建立聚类算法[1]。

对于手写体金融汉字的识别问题可以看作是一个多类分类问题，这里的类别数 N=21，构造手写体金融汉字识别分类树的算法如下：

(1)设聚类类别数为 C=2，交叠系数 $\delta\in[0,1]$。

(2)将全部的l个原始样本 $<x_i,y_i>, i=1,2,...,l, y_i\in[1..N]$，按照文献[1]提供核聚类算法($C$=2)，将原始样本分为 2 个模糊子类$S_1$和$S_2$。对于原始类别$I$，可以定义其样本集为

$$N_i=<x,i>,\ i=1,2,...,N$$

原始类别N_i被划分到子类C_j的概率为

$$P(i,j)=\frac{|N_i\cap S_j|}{|N_i|},\ i=1,2,...,N,\ j=1,2$$

则对于子类C_1和C_2，原始类别i的分配方案如下：

1)如果 $|P(i,1)-P(i,2)|\ge\delta$，则若 $P(i,1)>P(i,2)$，类别i划分到C_1类；否则，类别i划分到C_2类；

2)如果 $|P(i,1)-P(i,2)|<\delta$，则将类别i同时划分到C_1和C_2类。

(3)对于每一个子类，重复应用步骤(2)进行划分，直到所有的子类中都仅包含一个原始类别。

(4)根据步骤(2)、步骤(3)的预分类结果，定义分类子任务，每一个分类子任务对应决策树中的一个节点，每个节点都是一个 SVM 二值分类器。

从算法中可以看出，该方法允许部分类别存在交叠(即同一个原始类别同时划分到 2 个子类中)，这样可以克服错分积累。通常交叠会降低分类速度，但可以提高分类的准确率，这里引入了交叠系数 δ 来调整允许交叠的程度：若 δ=1，则交叠可能非常严重，甚至会造成 2 个子类中都包括了全部的原始类别，这种情况会导致算法失败；若 δ=0，则不允许交叠存在，这时由于错分积累的原因会导致分类准确率降低。在应用中，需要根据实际情况选择一个合适的 δ。

在本文的算法中，所有数据样本的预分类、分类子任务的预定义及分类树的建立都是映射到高维空间中完成的，统计学习理论保证了这一非线性映射过程，将使样本之间的特征区别更加明显，使分类效果更优。

3 实验结果及分析

为了保证实验结果的真实有效性，笔者从银行实际使用的票据中采集样本，票据从不同地区(包括武汉、北京、青岛、上海、天津、兰州、广州等)的各个银行(包括中国银行、交通银行、建设银行、工商银行、城市商业银行、农村信用社等)采集，并选择在不同类型的票据上提取，手写体金融汉字包括零、壹、贰、叁、肆、伍、陆、柒、捌、玖、拾、佰、仟、万、亿、元、角、分、正、整、弌，共 21 个汉字，每个汉字 1 200 个样本，建立了一个实际的手写体金融汉字库。对每个汉字任取 800 个作为训练样本，另外 400 个作为测试样本。每个字符按 64×64 点阵进行归一化处理，按照文献[7]的方法进行 8×8 弹性网络变换得到 256 维向量，按照本文提出的方法建立分类决策树，然后对测试样本进行识别测试，实验平台为P4 2.6GHz CPU, 1 024MB RAM, Windows 2000 操作系统，所有算法均采用并用VC++ 6.0 实现，二值SVM分类算法是在的 Chih-Chung Chang 和 Chih-Jen Lin 提供的

LIBSVM工具包的基础上修改实现的，其中，SVM的核函数选择径向基函数，采用交叉验证的方法来选择训练参数(C,γ)，结果为$(2^3, 2^{-7})$，生成的手写体金融汉字决策分类树如图 1 所示(仅列出了左子树部分)。

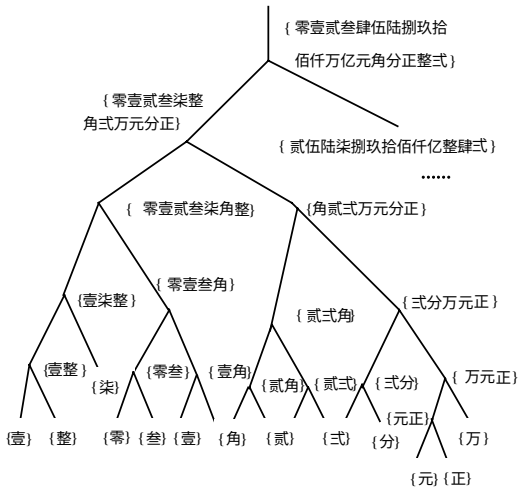

图 1　手写体金融汉字决策分类树(部分)

实验同时采用了 1-a-1 和 1-a-r 方法对这些样本进行训练识别，并与本文的方法进行比较，结果如表 1 所示。

表 1　实验结果比较

分类方法	分类时间/s	准确率/%
1-a-1	2 238	89.83
1-a-r	967	84.24
核聚类方法	235	93.13

实验结果表明，本文提出的基于核聚类的手写体金融汉字识别方法，从速度和准确率方面比 1-a-1 和 1-a-r 方法都有较大提高。从图 1 也可以看出，利用核聚类的方法，仅需要执行 6 次分类就可以得到结果，而 1-a-1 方法需要 210 次分类，1-a-r 方法需要 20 次分类，因此，利用核聚类方法的分类速度大大提高，超过了 35 字/s，达到了实际需求。

在训练阶段首先根据按照设定交叠系数 δ 进行 2 类核聚类，然后再进行 2 类 SVM 分类器的训练。为了得到较好的识别结果，需要调整 δ 的值，甚至每次核聚类时采用不同的 δ，这大大增加了训练时间，但实际使用中，训练与识别是分离的，并且训练完成后可以长期使用，因此，可以容忍较长的训练时间。本实验中取 δ=0.05。

4　结语

本文针对手写体金融汉字别问题进行了研究，提出一种基于核聚类和决策树结合的多类分类方法，该方法充分利用 SVM 的特点，在训练阶段利用核函数将原始样本特征非线性映射到高维特征空间，样本在经过很好的分辨、提取和放大后，可以实现更准确的聚类，根据聚类的结果来生成决策分类树，本文给出了决策分类树的生成算法。实现表明，该方法比 1-a-1、1-a-r 方法在速度和准确率方面都有较大的提高，基本可以满足银行票据处理中对手写金融汉字识别的要求。进一步的研究工作包括对决策树底层的分类器进行改进，采用更能体现字符细节的特征，提高对相似字符的识别能力。

参考文献

1 张　莉, 周伟达, 焦李成. 核聚类算法[J]. 计算机学报, 2002, 25(6): 587-590.

2 Vapnik V N. 统计学习理论[M]. 许建华, 张学工, 译. 北京: 电子工业出版社, 2004.

3 Crammer, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines[J]. Journal of Machine Learning Research, 2001, 2(3): 265-292.

4 Chih-Wei H, Chih-Jen L. A Comparison of Methods for Multi-class Support Vector Machines[J]. IEEE Transactions on Neural Networks. 2002, 13(2): 415-425.

5 Schwenker F. Hierarchical Support Vector Machines for Multi-class Pattern Recognition[C]//Proc. of the 4th Int'l Conf. on Knowledge Based Intelligent Engineering Systems & Allied Technologies, Brighton. 2000.

6 Girolami M. Mercer Kernel Based Clustering in Feature Space[J]. IEEE Transactions on Neural Networks, 2002 ,13(3): 780-784.

7 金连文, 彭秀兰, 尹俊勋. 一种手写体汉字特征提取新方法: 小波变换及弹性网格技术的应用[J]. 中国图象图形学报, 1998, 3(7): 549-552.

(上接第 177 页)

此外，图 6 给出了水印图像在受到压缩攻击后，本文的基于预测攻击的检测算法和普通的未加预测攻击的检测算法提取水印随压缩品质因子 Q 的变化对比情况。从图 6 可以看出，本文的基于预测攻击的检测算法能明显地改善水印系统的检测性能，提高水印的鲁棒性。并且，从它的理论推导过程来看，可以看出该检测算法不失一般性，不仅可以用在小波域中，也可以用在其他的变化域和各种各样的算法中(事实上，作者在 dct 域的另一算法中也做了实验，该检测方案的应用同样极大地改善了检测性能)，因此具有较好的通用性。

3　总结

本文阐述了为使水印系统达到更好的性能，水印的检测方案也很重要这个观点。并首次提出了一种新的基于预测攻击的相关检测方案，通过充分运用原始载体作品信息，按照一定的方法判断嵌入水印后的图像所受到的攻击类型，再对原始载体作品进行相同的攻击处理，以此来提高原始作品和水印作品的相关性，最大限度地排除了噪声的干扰，从而更有效地恢复水印信息。仿真结果表明，该方案可以大大提高水印检出率，并不失一般性。因此，该新型的检测方案有较强的可实现性和应用前景。

参考文献

1 Swanson M D, Kobayashi M, Tewfik A H. Multimedia Data-embedding and Watermarking Technologies[J]. Proceedings of the IEEE, 1998, 86(6): 1064-1085.

2 Delaigle J F, Vleeschouwer C, Macq B. Psychovisual Approach to Digital Picture Watermarking[J]. Journal of Electronic Imaging, 1998, 7(3): 628-640.

3 Cox I J, Miller M L. Digital Watermarking[M]. 北京: 电子工业出版社, 2003: 32-57.

4 周亚训, 徐铁峰. 基于二值运算的隐形签名数字水印算法[J]. 通信学报, 2002, 23(2): 107-112.

5 丁　玮, 齐东旭. 基于 Arnold 变换的数字图像置乱技术[J]. 计算机辅助设计与图形学学报, 2001, 13(4): 338-341.