学校编码：10384　　分类号_____ 密级_____

学号：19020141152630　　UDC_____

厦门大学

硕 士 学 位 论 文

改进的AP-SVM算法研究及其在字母识别的应用

Research on Improved AP - SVM Algorithm and Its Application in Letter Recognition

刘 晓 红

指导教师姓名：谭　忠　教授

专 业 名 称：应　用　数　学

论文提交日期：2017　年　4　月

论文答辩日期：2017　年　5　月

学位授予日期：2017　年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2017 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。本学位论文属于：

（　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

中文摘要

字母识别是字符识别的一个特例，正确识别字符信息并实现自动录入在信息化建设快速发展的今天拥有越来越重要的意义。在过去几十年中，很多学者将多种分类算法应用于字母识别领域，包括贝叶斯分类器、BP 神经网络、支持向量机等。其中，支持向量机（SVM）分类算法因其趋于完善的理论研究和算法实现研究，及克服维灾难和过拟合的优点，在字符识别领域取得了理想的效果。但支持向量机的算法复杂度受到样本规模的直接影响，且对噪声和孤立点较为敏感。因此，如何有效挑选经典样本成为提高模型分类性能的关键。

考虑到AP聚类算法具有如下两个优点：其一，算法无须事先指定聚类个数；其二，算法具有很低的均方误差，因此可借助AP聚类算法解决上述问题。本文主要研究改进的AP-SVM 模型，首先分别阐述支持向量机（SVM）和AP聚类算法的基本思想和主要实现算法；随后介绍改进的AP-SVM模型的主要思路及算法实现步骤，即将AP聚类算法作为一种数据预处理手段，利用其求得聚类中心，选取这些聚类中心及每个中心的边缘分布点作为典型样本重构样本空间，然后在新样本空间中用SVM训练样本并预测。在实证研究部分中，通过比较支持向量机模型与改进的AP-SVM模型的英文字母分类效果，证实后者的分类正确率高于前者，并且通过参数调优寻得改进模型的最优参数，使字母识别正确率有效提升。

关键词：字母识别；支持向量机；AP聚类；改进的AP-SVM

Abstract

Letter recognition is a special case of character recognition. The correct recognition of character information and achieve automatic entry have more and more important significance in the rapid development of information technology today. In the past few decades, many scholars have applied a variety of classification algorithms to letter recognition, including Bayesian classifier, BP neural network, support vector machine and so on. Among them, SVM classification algorithm is more and more perfect. It can overcome the difficulties of disaster and over-fitting in the field of character recognition. However, the complexity of the support vector machine is directly affected by the size of the sample, and it is more sensitive to the noise and the isolated points. Therefore, how to effectively select the classical sample becomes the key to improve the classification performance of the model.

Considering that the AP clustering algorithm has the following two advantages: First, it does not need to specify the number of clusters in advance. Second, the algorithm has a very low mean square error. Therefore, AP clustering algorithm can solve the above problem. In this paper, we mainly study the improved AP-SVM model. Firstly, we introduce the basic idea and main implementation algorithm of support vector machine and AP clustering algorithm respectively. Then, we introduce the main idea and algorithm of the improved AP-SVM model. In other words, the AP clustering algorithm is used as a data preprocessing method. The clustering centers are obtained by using the AP algorithm. Clustering centers and the edge distribution points of each cluster are seen as typical samples to reconstruct the sample space, and then use the support vector machine in the new sample space. In the empirical

study, by comparing the support vector machine model with the improved AP-SVM model, it is proved that the classification accuracy of the latter is higher than that of the former. At last, the optimal parameters of the improved model are obtained by parameter tuning to effectively improve letter recognition accuracy.

Key words: Letter recognition; SVM; AP clustering; Improved AP-SVM

目 录

Contents

第一章 引言

1.1 研究背景及意义

字母识别是字符识别的一个特例，准确而快速地识别字母在信息化建设快速发展的今天具有重要的意义。通常，字符识别和信息处理涵盖两大方面：一方面是文字信息，主要包含许多国家的文字信息；另一方面则是数字信息，无论是金融领域中的公司财务报表还是社会生活领域中的大规模统计调查，如人口普查等，都会产生大量的数字信息。当今处于一个信息爆炸时代，频繁地依靠计算机获取及处理信息是人们必然的需求。为了解决人工输入信息导致效率极低的弊端，如何正确识别这些信息并实现自动录入成为亟待解决的问题。

字母识别长期以来都是一个备受关注的研究课题。字母识别需要解决的问题主要包括数据采集、特征提取、识别分类器的选择等。在分类器的选择方面，很多学者进行了多种探索，在字符识别领域应用较为广泛的模型有朴素贝叶斯、支持向量机、BP神经网络等。其中，支持向量机（SVM）有坚实的理论基础和趋于完善的算法研究，并且可以有效克服维灾难和过拟合等缺陷，具有较好的鲁棒性，因而受到广泛应用。而支持向量机被Vapnik及其实验室团队第一次提出后，首个应用就是手写阿拉伯数字的识别。但考虑到支持向量机训练效率受到样本数量的直接影响，且对噪声和孤立点较为敏感。因此，如何有效挑选经典样本成为提高模型分类性能的关键。若是随机选取样本，会导致分类准确率明显降低，若是人工选取样本，将耗费较多的人力，降低模型的整体分类效率。而聚类算法只需要给出数据，通过寻找数据间的规律，自动聚类，虽然效率较高，但往往准确率较低，是一种粗糙的分类方法。若能将两者结合起来，将聚类视为一种数据预处理的方法，通过聚类快速找寻典型样本，利用这些典型样本重构样本空间，用支持向量机在新的样本空间训练

和测试，提高分类正确率，对字母的识别有较大的意义。

1.2 国内外研究现状

在机器学习研究不断发展的若干年中，很多学者尝试了不同的字母识别算法，从最初的逻辑推理法到现在流行的BP神经网络、支持向量机等经历了一个不断发展的过程。下文对主要方法做简要介绍。

(1) 逻辑推理法

逻辑推理法的主要思想为，站在研究对象的角度，从数据库找寻一系列规则进行推理，得到结果，而每种结果都对应某个类别。

(2) 模板匹配法

模板匹配法是一种思路简单且容易实现的识别方法，其主要思想为：每个字母都对应着一个标准模板，计算每个未知样本的点阵图像距各个字母模板的距离，距离小则表明匹配程度高，分类策略为距离最小的那一类。常用的距离测度公式有：

$$D(i,j)=\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(S_{ij}(m,n)-T(m,n))^2}$$

或

$$D(i,j)=\sum_{m=1}^{M}\sum_{n=1}^{N}|(S_{ij}(m,n)-T(m,n))|$$

其中，$S_{ij}(m,n)$表示子图S_{ij}的$m\times n$个像素，$T(m,n)$表示模板的$m\times n$个像素。

(3) 模糊判别法

模糊判别法的理论基础为模糊数学，在标准模式库中，通过评估每个未知实例与标准模式库样本的相似程度，根据最大隶属原则将未知实例分类。此方法可以反映样本的主要特征，且在模糊模式下有较强的抗干扰能力，但是隶属函数通常难以确定。

(4)贝叶斯判别

贝叶斯判别是根据贝叶斯准则对未知样本进行推断的一种多元统计学分类方法，以二分类为例说明其主要思路：设q_1，q_2分别为两个总体G_1，G_2的先验概率，密度函数分别为$f_1(x)$，$f_2(x)$，对于待分样本x，根据贝叶斯公式计算得到它属于第$k(k=1,2)$个总体的后验概率如下：

$$P(G_k|x)=\frac{q_k f_k(x)}{\sum_{k=1}^{2} q_k f_k(x)}$$

分类策略是将未知样本x归属为后验概率$P(G_k|x)$最大的总体，使期望损失最小。由贝叶斯公式可以看出，各个类别的概率分布估计对分类的影响很大，而在实际问题中概率密度函数的分布是很难估计的。

(5)BP神经网络

BP神经网络模型是由Rumelhart和McCelland带领的的科学研究小组于1986年首次提出，是一种将误差逆向传播的具有层次的前馈网络。它的拓扑结构由输入层（input layer）、隐藏层（hidden layer）和输出层（output layer）三部分组成。模型训练时，信息首先逐层前向传播，当经输出层向外界发布的预测并非期望结果时，将待测样本的预测结果与已知的目标值之间的误差向隐藏层传递，即转入后向传播阶段，并利用误差梯度下降法不断修正各层之间的连接权重和偏倚，使预测结果逐渐逼近目标值，反复迭代直到模型的全局误差低于阈值或者学习次数达到最大时停止[1]。模型学习过程其实为权重和偏倚不停更新的过程。BP 神经网络模型的建立不需要先验知识和判别函数，具有较强的抗干扰能力。

BP神经网络具有良好的非线性映射能力和高容错能力，在字符识别领域被广泛应用[2]-[3]。吴聪等曾基于此分类方法研究车牌号码的识别，通过数据仿真实验验证了该算法的时效性和鲁棒性[4]；杨杰等曾在研究图像分类时，通过比较不同的分类方法，指出了BP神经网络模型的缺陷，如对初始权重敏感，很容易收敛至局部极小值，且训练时间较长，隐藏层结点个数对分类准确率有直接影响等。

(6)支持向量机

支持向量机（Support Vector Machine，简称SVM）模型是由Corinna Cortes和Vapnik及其实验室成员于1995年提出来的[5]，是一个典型的二分类器，通过寻找分类间隔最大的超平面$\omega^T x + b = 0$，使不同类的数据点恰好位于该超平面两侧，从而实现最优分类。最佳分离超平面的搜寻实质为求解如下凸二次规划问题：

$$min \quad \frac{1}{2}\|\omega\|^2$$

$$s.t. \quad y_i(\omega^T x_i + b) \geq 1, \quad i = 1, 2, \cdots n$$

等价于求解对偶问题：

$$max_\alpha \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j)$$

$$s.t. \quad \alpha_i \geq 0, \quad i = 1, 2, \cdots n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

对于待测样本x，决策函数为:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i (x_i, x) + b)$$

支持向量机能够高效地解决非线性边界问题，并且模型的复杂度$J(f)$受训练样本规模的直接影响，与数据的维度无关，故在高维数据分类建模方面体现了突出的优越性。十几年来，很多学者在理论研究与算法改进方面取得了很多成果，在很多领域已经被成功应用，如字符识别、数据挖掘[6]、回归分析[7]等，逐渐成为重要的研究热点。

1.3 本文主要内容及结构安排

本文主要研究改进的AP-SVM算法及其在英文字母识别的应用。支持向量机（SVM）本质是基于线性二值分类机理，寻找间隔最大的分类超平面。为了提升模型处理非线性数据的能力和拥有良好的泛化能力，支持向量机模型巧妙地引入核技巧和惩罚项，并且突破传统求解小型规划问题的标准算法，采用启发式方法求解二次规划问题，以较快搜索最优解。然而，SVM的算法复杂度受到样本规模的直接影响，且对噪声和孤立点较为敏感，这些实例点以及对分类贡献较小的点影响分类性能。此时，合理挑选有效样本非常必要。因此文本在利用SVM训练之前首先通过AP 聚类算法，产生高质量的簇中心，利用这些典型的簇中心及每个簇中心的边缘分布实例点重构样本空间，然后再进行SVM 建模，以此提高分类性能，并在UCI 字母识别数据集进行实证研究，实验结果表明改进的AP-SVM 算法使得字母识别正确率有效提升。

本文分为六个部分展开，具体安排如下：

第一章为引言部分。首先介绍了字母识别的研究背景及意义，随后讲述字母识别研究的发展现状，侧重介绍应用于该领域的分类方法。

第二章和第三章为本文的基础理论部分。其中，第二章详细介绍了SVM的基础知识，包括SVM的基本思想、理论推导，算法求解方法，尤其对SMO快速训练算法进行了详细的阐述，最后将分类策略从标准的二分类推广至多分类。第三章为AP聚类算法的理论介绍。包括常用聚类方法的主要思路、模型特点及缺陷，从而引出AP聚类模型，详尽表述了AP 聚类的主要思想、算法实现和参数调优等内容。

第四章为本文核心内容，也是本文创新部分。将AP聚类方法与SVM分类器有效结合，称为改进的AP-SVM算法，即借助AP聚类搜索簇中心，利用寻得的簇中心及每个簇中心的边缘分布点重构样本空间，在新样本空间用SVM训练及预测。

第五章为实证研究部分。首先通过SVM，改进的AP-SVM算法对四个数据集训