学校编码：10384　　分类号_____密级_____
学号：B200431005　　UDC_____

博　士　学　位　论　文

支持向量机的核选择

Research on Kernel Selection of Support Vector Machine

罗林开

指导教师姓名：林成德　　教授

专　业　名　称：控制理论与控制工程

论文提交日期：2007 年　　月

论文答辩时间：2007 年　　月

学位授予日期：2007 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2007 年 10 月

厦门大学学位论文原创性声明

兹呈交的学位论文，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的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“ √ ”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘　要

由 Vapnik 等人提出的支持向量机（Support Vector Machine，SVM）技术，由于具有极强的模型泛化能力，不会陷入局部极小点，以及很强的非线性处理能力等特点，近十年来取得了全面飞速的发展，获得了大量成功的应用，已成为模式识别中最为活跃的研究领域之一。

当前，选择合适的核函数及其参数（核选择）已成为 SVM 进一步发展的关键点和难点。核函数决定了 SVM 的非线性处理能力，也决定着分类函数的构造，而对具体问题而言，选择合适的核函数及其参数，还存在着许多的实际困难。

针对 SVM 中的核选择问题，本文对 SVM 的模型问题、特征空间线性可分的结构问题、核学习中基核的选择问题、以及核函数及其参数的评判准则问题开展了深入的探讨，主要的工作有：

1．在 SVM 的模型方面，给出了 L2-范数下平分最近点原理问题；然后得到了它的解与最大间隔原理问题的解之间的关系，建立了它与最大间隔原理的等价性；指出它还具有模型性质更好、几何意义更直观、能利用求解凸包之间距离的内点算法等优点；最后给出了它的 SMO（Sequential minimal optimization）求解算法。

2．在特征空间线性可分的结构方面，利用平分最近点原理模型，通过对核矩阵零空间的深入分析，得出特征空间中样本线性可分与核矩阵零空间关系的一个充要条件。

3．在基核矩阵的选取方面，首先提出矩阵的秩空间差异性（Rank Space Diversity，RSD）概念，其次将其作为基核矩阵的差异性度量，由此导出选择基核矩阵的一个定量规则“基核矩阵的秩空间差异性越大越好”。我们还给出了基于 L2-范数下平分最近点原理的核学习模型和模型求解算法；最后通过实验验证了基核矩阵选择规则的有效性。

4．在核函数及其参数的评判准则方面，首先从分类函数抗样本扰动的“泛化性能”出发，分析了传统最大间隔原理的不足，提出了分类函数的鲁棒度概念；探讨了鲁棒度的性质；并提出用最大鲁棒度作为核选择的评判准则；通过与经典

的交叉验证方法和最小支持向量数方法的实验对比，表明最大鲁棒度准则克服了交叉验证方法时间代价高，最小支持向量数方法测试准确率不稳定的缺点，获得了很好的结果。

5．在核学习方面，提出了按单属性设计基核，以最大鲁棒度为优化目标的核学习方法，给出了鲁棒度的梯度计算公式和模型的求解算法，并用实验表明了该方法的有效性和优越性。

关键词：SVM；核选择；秩空间差异性；鲁棒度；平分最近点

Abstract

In the last ten years there have been very significant developments in the theoretical understanding of Support Vector Machines (SVMs) , proposed by Vapnik and others, as well as algorithmic strategies for implementing them, and applications of the approach to practical problems.

Nowadays, the selection of the SVM-kernel with suitable form and parameters (Kernel Selection) has become a key-point both in theoretical research and application consideration. In fact, the nonlinear processing ability of SVM and the structure of the separating function are both largely decided by the choice of individual kernel function, and actually there are still a lot of difficulties on practice.

As a research work focused on kernel selection of SVM, this paper has mainly discussed the following problems:

1. On the modeling of SVM, the principle of bisecting closest points under L2-norm is firstly introduced. The relation between the solutions based respectively on the bisecting closest points principle and the maximum margin principle is then deduced, and the equivalence is established on these two solutions. The advantage of bisecting closest points method is showed, including of the better model character, the more intuitive geometric significance, and the optional nearest point algorithm. A SMO typed algorithm for the model based on bisecting closest points principle under L2-norm is also presented.

2. On the aspect of linear separable structure of sample set in feature space, a necessary and sufficient condition is obtained based on null space of kernel matrix.

3. On the aspect of selecting the base-kernels in kernel learning, a new concept of rank space diversity of matrices is firstly proposed; it is considered as a diversity measure for the base-kernel matrices. “Rank space diversity of base-kernel matrices should be as big as possible” is then deduced as a rule for the selection of base-kernel matrices. The kernel learning model based on bisecting closest points principle under L2-norm, as well as it’s solving algorithm, are given, and the validity of this rule is showed by some experiments.

4. On the aspect of the criterion of kernel evaluation, a robustness concept on separating function is firstly proposed based on the anti-disturbance ability of samples. By its properties, the maximum robustness of separating function is proposed to be a criterion for kernel evaluation. Experiments on the comparison among classic k-fold cross validation, minimum support vectors and maximum robustness methods show that our proposition is efficiency, which overcomes the shortages of high time cost for k-fold cross validation and the unstable testing accuracy for minimum support vectors.

5. On the aspect of kernel learning, a new method is proposed, in which the base-kernels are designed on each attribute and the robustness of separating function is maximized. The corresponding solving algorithm of this kernel learning model is presented, and the validation and advantages of our method is shown by some numerical experiments.

Keywords: SVM, Kernel Selection, Rank Space Diversity, Robustness, Bisecting Closest Points.

目　录

Contents

第一章 绪 论

1.1 本文的研究背景和研究意义

分类是人类认识世界的一个基本手段，人类的生活和生产实践都离不开分类活动，例如人们能识别以前认识的人和事物，靠的是分类，控制科学中对控制对象的识别、跟踪等也都要依赖分类识别技术。随着上个世纪计算机的诞生和发展，信息科学与技术得到了飞速的发展，现代社会快速进入了信息时代，分类更成为人们分析、处理和理解现代信息世界的重要工具。信息时代产生了大量、复杂的数据，需要人们去分析和处理。如何有效地利用、理解数据，如何从海量数据中发现规律，获取知识，已成为一个富有挑战性、极具价值的研究领域，而分类作为人们获取知识的一个基本手段，在其中起着重要的、不可或缺的作用[1]。

近几十年来，分类学得到了飞速的发展，出现了许多有效的智能分类方法，如决策树方法[2,3,4]、神经网络方法[5-8]、贝叶斯分类[9,10]、K-近邻算法[11,12]、模糊集方法[13,14]和支持向量机方法（Support Vector Machine, SVM）[15-18]等，它们被成功地应用到许多实际工作中，极大地促进了分类学的发展。

值得特别指出的是，由 Vapnik 等人[15]提出的 SVM 技术，由于具有极强的模型泛化能力，不会陷入局部极小点，以及很强的非线性处理能力等特点，近十年来取得了全面飞速的发展，现已成为机器学习和数据挖掘领域的标准工具，它在字符识别、文本自动分类、人脸检测、头的姿态识别、生物识别、疾病诊断、医学图像处理、智能控制、经济模式分类、预测等方面获得了大量成功的应用，已成为模式识别中最为活跃的研究领域之一[19-42]。

作为使 SVM 具有强大非线性处理能力的核方法，在 SVM 中占据着举足轻重的地位，它不仅是 SVM 的热点[43-49]，而且成为 SVM 进一步发展的关键点和难点。核方法之所以成为 SVM 的关键点，是因为核决定了 SVM 的非线性处理能力，也决定了分类函数的构造；核方法之所以成为 SVM 的难点，是因为在具体问题中，选择合适的核函数及其参数（核选择），还存在着许多的实际困难。

选择合适的核函数及其参数是 SVM 的重点和热点研究方向。近年来，国际上一些著名学者在这方面已经开展了一些有益的工作，如 Vapnik[15]的交叉验证，

Chapelle 和 Vapnik 等人[50]的半径-间隔界，Cristianini 等人[51]的核排列，Lanckriet 等人[52]基于半定规划的核学习等，这方面的相关文献请见[15,50-60]。2006 年 11 月 27-28 日,在比利时的布鲁塞尔召开了“International Workshop on Current Challenges in Kernel Methods (CCKM06)”，国际许多著名学者，如 Schölkopf, Cristianini 等都在此次会上发表了主题演讲，提出了核方法面临的挑战和未来的发展方向,此次会议必将为核方法的研究带来一个新的高潮。

虽然在选择合适的核函数及其参数方面，国内外已经开展了一些有益的工作，为其发展奠定了一定的基础，但还存在许多有待进一步研究的基础问题。例如，核函数及其参数的评判准则问题，特征空间的结构问题，核学习中基核的选取问题，核的构造问题等，这些问题的解决对 SVM 的进一步发展起着重要的关键作用。

考虑到以上因素，本文以 SVM 中的核选择作为研究方向，对 SVM 的模型问题，特征空间中线性可分的结构问题，核学习中基核的定量选择规则问题，以及核函数及其参数的评判准则问题开展了深入的探讨，期待得出一些有益的研究成果，为 SVM 的进一步发展作出有益的贡献。

综上所述，本文的研究符合 SVM 的发展方向，研究的内容属于 SVM 中关键的、基础性的，且目前尚未开展的工作，本研究目标的达成，对于促进 SVM 的进一步发展，具有重要的科学价值；此外，由于本研究属于 SVM 的基础性工作，其成果可直接应用在 SVM 的所有应用领域，因此它也具有极大的应用前景。

1.2 研究现状与发展方向

在核选择方面，大致有核学习、核设计和流形学习三个主要方向，下面分别对它们作一个简单的综述。

- 核学习

当样本在输入空间线性不可分时，SVM 采用核函数将输入空间映射到一个特征空间（高维的 Hilbert 空间）中，然后在特征空间中求最优分类超平面。在这

种处理方法中，至少就存在两个模型参数的选择问题，一个是核参数，另一个是对误分的惩罚系数。

最早的、也是最经典的参数选择方法是 Vapnik 等[15]提出的 k-折交叉验证方法。k-折交叉验证方法通过在参数空间逐一比较各个参数的 k-折误分率，来选择最优参数；为了计算 k-折误分率，首先必须将训练样本随机地分成 k 个互不相交的子集$S_1,\cdots,S_k$，每个子集的大小大致相等，然后进行 k 次训练与测试，其中第 i 次的训练集是$S_1,\cdots,S_k$中扣除S_i的并集，测试集为S_i，最后用 k 次测试的平均误分率作为 k-折误分率。当 k 等于训练样本的个数时，k-折交叉验证又称为留一法（Leave-One-Out，LOO），相应的 k-折误分率称为 LOO 误差，已经证明 LOO 误差是模型泛化误差的一个很好的估计。k-折交叉验证的缺点是计算代价大，特别是 LOO 估计，若样本数较大，计算代价更是惊人，使得它能调节的参数个数较少（一般不超过 2 个参数）。

2002 年 Chapelle 和 Vapnik[50]针对 k-折交叉验证的缺点，提出了根据泛化误差的各种上界，选择多个模型参数的方法，在[50]中，提到的界包括支持向量数，Jaakkola-Haussler 界，Opper-Winther 界，半径-间隔（Radius-Margin，RM）界和 Span 界等，其中半径-间隔界由于较易计算，被作为其重点推荐的优化目标。此后，Keerthi[61]也对优化 RM 界的算法进行了研究，指出了 RM 界的可行性和有效性，2003 年 Chung 等人[62]在 L1-SVM 下对高斯径向基核的 RM 界进行了研究，提出了一些 L1-SVM 下的启发性 RM 界。2002 年 Chapelle 和 Vapnik 的工作，不仅可以应用到选择多个核参数，而且还可以应用到特征选择中，可以说 Chapelle 和 Vapnik 当年的这个工作有着重大的理论意义。但是 RM 界在 SVM 的实际工作中，却一直未能取得特别理想的效果[62]。

2004 年 Lanckriet 等人[52]提出了基于半定规划的核学习方法。该方法首先设定一个可选核的集合，这个集合通常是某些基核的线性组合，然后在该集合中选择一个达到最好分类性能的核。Lanckriet 成功地将该问题转化为一个半定规划（Semidefinite Program,SDP）问题[63]，然后利用 SDP 技术求解，Lanckriet 的实验显示了组合核分类器取得了能与单个最好分类器竞争的效果。在 Lanckriet 的工作中有两个重要的特点，一个是面向转导推理，也就是待标号样本的输入属性值是已知的，另一个是面向凸优化问题。转导推理的引入简化了学

习问题的难度，使得核函数的学习转化为训练集和测试集上的 Gram 矩阵学习，凸优化问题的好处在于我们总能获得全局的最优解。Bach 和 Lanckriet[64]随后将其发展到多核学习（Multiple Kernel Learning，MKL），并引入最小序贯方法（Sequential minimal optimization，SMO）求解模型，提高了求解效率，并被应用到半监督学习问题中[65]。

Lanckriet 等人的工作开创了一个解决核选择问题的新框架，有重大的理论意义和实用价值，但同时也还存在许多有待进一步研究的问题，如基核的选取问题，以及核的评判准则问题等。

2006年Sonnenburg[54]将Lanckriet等人的多核学习归结为半无穷线性规划（Semi-Infinite Linear Program，SILP)，大大提高了模型求解的效率（能在合理的时间里完成百万级和上百个核的训练），并将其拓广到回归和一类问题，但Sonnenburg依然未对基核的选取问题，以及核的评判准则问题进行探讨。

- 核设计

许多实际应用中的模型空间并不具有欧氏空间的特点，例如字符串、图、集合等对象，这时为这些特定对象设计核，就很有必要。1999 年出现了根据递归关系定义的串核[66,67]，这些思想在2002年被Lodhi等人应用到文本归类领域，此后应用到生物信息学领域[68-71]。此外，很多情况下，我们知道一些关于生成数据过程的信息，例如，DNA 序列是通过对祖先序列进行一系列修改、进化而成的，时间序列可以由某一类型的动态系统生成，二维图像由三维图像的投影生成等，此时可以建立含先验信息的核函数，P-核[66]和 Fisher 核[72,73,74]是这些工作的代表。核设计的难点是如何构造含先验信息的核，这方面比较成功的例子是文本核[75]。

- 流形学习

核通过非线性映射将输入空间嵌入到特征空间，特征空间一般来说维数很高，但是对原问题来说，牵涉到的常常仅是一个低维的流形，近年来，许多学者开始研究分类和聚类问题的拓扑特征，形成了流形学习（Manifold Learning）方向，提出了核的黎曼度量等概念，开始研究核所蕴涵的几何度量[76-81]。把