Research and Application of Kernel Methods Based on Manifold Regularization

基于流形正则化的核方法研究与应用

刘兵 著

中国矿业大学出版社
China University of Mining and Technology Press

基于流形正则化的核方法研究与应用

刘　兵　著

中国矿业大学出版社

内 容 提 要

本书针对分类器设计中类不平衡、核优化以及快速半监督分类三个方面的问题，结合科研工作中的研究案例，介绍了模糊聚类、可能性模糊支持向量机模型、非参数核优化方法以及极速学习机等理论和算法。

本书内容全面，理论与实例分析相结合，可作为信息科学、计算机科学与技术、自动化等领域的从业人员和科研工作者的参考书。

图书在版编目(CIP)数据

基于流形正则化的核方法研究与应用/刘兵著. —徐州：中国矿业大学出版社，2016.6

ISBN 978-7-5646-3150-5

Ⅰ. ①流… Ⅱ. ①刘… Ⅲ. ①正则化—研究 Ⅳ. ①O177

中国版本图书馆 CIP 数据核字(2016)第 143548 号

书　　名　基于流形正则化的核方法研究与应用
著　　者　刘　兵
责任编辑　张　岩　郭　玉
出版发行　中国矿业大学出版社有限责任公司
（江苏省徐州市解放南路　邮编 221008）
营销热线　(0516)83885307　83884995
出版服务　(0516)83885767　83884920
网　　址　http://www.cumtp.com　**E-mail**:cumtpvip@cumtp.com
印　　刷　江苏徐州新华印刷厂
开　　本　880×1230　1/32　**印张** 7.125　**字数** 185 千字
版次印次　2016 年 6 月第 1 版　2016 年 6 月第 1 次印刷
定　　价　26.00 元

前　言

分类器设计一直是模式识别领域研究的重要课题之一。近十年来在以支持向量机为代表的核分类方法的基础上，又涌现出了一些新的研究热点，例如，海量高维数据的分类、类重叠和噪声干扰下的数据分类、多标记数据分类、类不平衡数据的分类、非线性分类中的核函数（矩阵）优化以及非线性快速分类等。本书研究的内容主要涉及不平衡数据分类方法、基于非参数核优化的分类方法以及快速半监督分类方法三个方面的相关研究内容。在前期工作的基础上，建立了多种分类和学习模型，提出了新的学习算法，并使用标准数据集和多个人脸数据集对算法进行了测试。通过与相关算法进行对比，进一步验证了本书提出的算法的有效性。

全书共分六章。第1章阐明了本书研究课题的研究背景和意义，针对模式分类中的不平衡数据处理、核矩阵优化和半监督学习等问题指出其中不完善的方面和需要解决的问题。第2章针对线性可分问题和线性不可分问题，对经典的线性和非线性支持向量机模型进行了描述。这些理论和方法是后面章节具体研究课题的理论基础。第3章针对实际应用中样本重叠以及噪声干扰问题，提出了一种基于样本加权的可能性模糊聚类算法和一种鲁棒可能性模糊核聚类算法。所提出的聚类算法不仅可以同时处理线性不可分和部分重叠数据集，而且具有更强的鲁棒性，

在噪声干扰下能够取得较好的聚类准确率。针对实际应用中正负样本数量分布不平衡分类问题，利用本章提出的鲁棒聚类算法，建立了可能性模糊支持向量机模型，提出了基于可能性模糊聚类的不平衡数据分类方法。所设计的分类器较好地解决了分类中的类不平衡、孤立点和噪声干扰问题。第4章针对多核学习效率较低以及需要预先定义一组核函数等缺陷，建立了无监督非参数核学习模型，所建立的模型把谱核学习和间隔最大化标准进行有机结合，充分利用了数据的低维流形结构，增强了决策函数的光滑性，同时可以有效利用未标记数据进行最大间隔分类。实验验证在有监督和无监督情况下，提出的非参核学习方法的性能均优于多核学习方法。第5章为解决半监督快速学习问题，建立了扩展的流形正则化框架，所提出的算法是流形正则化极速学习机(MRELM)分类算法，它是传统核分类的近似算法，实验结果验证了MRELM算法的有效性。第6章对本书研究内容进行总结，同时提出待于进一步研究和完善的问题。

本书的写作过程得到学院领导和同事的大力支持，在此向他们表示感谢。此外，还要感谢博士研究生刘明明，硕士研究生杨烁、缪健、马丁等为本书的出版提供的帮助。最后，感谢“中央高校基本科研业务费专项资金”(项目号：2014QNA45)的资助和国家自然科学基金青年科学基金项目(项目号：61403394)的支持。

由于本书作者知识水平有限，书中不妥之处在所难免，恳请读者批评指正。

著　者

2015年8月

目　录

1 绪 论

1.1 研究背景及意义

机器学习是智能信息处理的核心研究内容之一，重点研究如何通过计算机模拟或实现人类的学习行为，从而获取新的知识，重新组织已有的知识结构，并不断改善自身性能。机器学习以客观世界的数据作为观测对象，利用观测数据寻找潜在的规律，进而对未知数据进行决策。在机器学习领域，模式分类是最重要的研究课题之一，近年来取得了长足的进展。一些经典的学习算法，例如支持向量机等，最初都是为解决分类问题而提出的，然后又拓展到了机器学习的其他研究领域。

模式识别(pattern recognition)以模式作为基本处理对象，对表示模式的各种数值的、文字的和逻辑关系的信息进行处理和分析，最终实现对模式的描述、辨认、分类和说明[1,2]。它通常包括训练和测试两个阶段，具体是指在训练阶段用已知训练数据对设计的分类模型进行训练，然后在测试阶段利用训练出的分类器确定未知数据的类别[3]。完整的模式识别系统如图 1-1 所示，包含四个处理阶段：数据采集、特征选择和提取、分类器设计以及评价。

数据采集，是指通过传感器等采集设备捕获观测对象的信

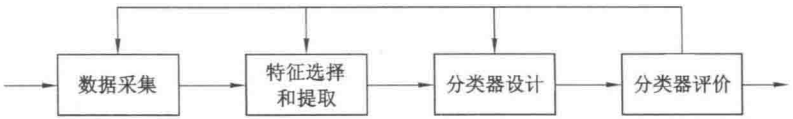

图 1-1　模式识别系统的构成

息，通过量化等方法将采样信息进一步转换为数据向量等形式，并最终输入计算机[2]；特征选择和提取，指利用变换方法把原始空间的数据映射到特征空间，然后在特征空间选择或提取低维的数据典型特征[2]；分类器设计，指在特征选择和提取的基础上，利用选择或提取出的特征作为训练样本对分类器进行训练[4]；分类器评价，是指对训练出的分类器的推广能力进行评价，若分类器对未知样本的测试结果不理想，则需要返回前面的阶段，重新设计分类器，直到最终获得较好的模式识别性能[5]。

模式分类起源于 20 世纪 50 年代，根据采用的理论和方法的不同大致可以分为三个阶段[3-6]：线性分析方法阶段、基于多层神经网络和非线性映射的非线性分析方法阶段和基于核函数的非线性分析方法。线性分析方法阶段主要利用单层计算单元的前向神经网络模拟人脑的学习能力。感知器(perceptron)的提出标志着科研人员对机器学习的研究上升到理论阶段。基于多层神经网络和非线性映射的非线性分析方法阶段的代表方法是 1986 年 Rumelhart 等人提出的多层感知器反向传播学习(back propagation，BP)算法[7]，此后，机器学习开始进入实际应用阶段。然而，这种方法缺乏完备的数学理论基础，容易出现过拟合和欠学习等问题。20 世纪 90 年代中期出现了以支持向量机为代表的核学习方法，简称核方法(kernel methods)，模式分类进入基于核函数的非线性分析方法阶段[8-13]。其理论基础是统计学习理论

(statistical learning theory)。传统的以支持向量机 SVM(support vector machine)为代表的核方法目前在模式识别等领域获得了广泛的应用[2,3]。核函数的引入使得 SVM 可以方便地处理非线性可分问题,复杂的特征映射通过核函数隐式定义,能够在特征空间中高效地计算内积,因此核函数的引入可以大大降低非线性变换的计算量。这种思想迅速渗透到一些传统的学习算法,通过把线性学习方法核化,产生了基于核的非线性学习算法。例如核 Fisher 判别分析、核主成分分析、核独立成分分析、核聚类分析等。这些方法是传统线性算法引入核函数后的非线性推广,统称核方法。各种核方法的共同点在于:原始数据在映射空间的点积通过核函数间接求解,这种特征映射是隐含的,无需求解,实现了在一个高维映射空间的数据的线性可分,因此通过使用线性算法解决了线性不可分数据的处理问题,避免了神经网络不能解决的"维数灾难"问题。

近年来,在以支持向量机为代表的核分类方法的基础上,围绕核方法的鲁棒性、计算复杂度、核函数(矩阵)优化、半监督核方法等几个方面,涌现出了新的研究热点。这些新的热点往往针对的是传统模式分类方法存在的弊端,如:海量高维数据的分类,多标记数据分类,类不平衡和代价敏感数据的分类,核函数参数选择、有序分类等[14]。本书重点研究基于聚类和流形正则化的分类方法,提出解决类不平衡、核优化以及快速分类的方法,该研究将丰富分类问题的解决途径,具有一定的理论研究意义和较好的应用前景。

1.2 国内外研究现状

本书主要研究基于聚类和流形正则化的分类方法及其应用,

涉及不平衡数据分类方法、基于无参核矩阵优化的分类方法以及半监督分类方法三个方面的相关研究内容。下面分别针对这三个方面的国内外研究现状进行阐述。

1.2.1 不平衡数据分类

传统分类算法的性能在高度不平衡且类重叠（highly imbalanced and overlapping，HIO）的数据集上具有局限性。HIO 数据集是指其中一类（多数类）样本远多于（不平衡比例接近或超过 1∶10）另一类（少数类），并且两类样本在特征空间存在不可分的情况。不平衡数据分类包括数据预处理和改进 SVM 算法两种主要方法。数据预处理方法的主要思想是：首先预处理训练样本，使样本基本满足平衡要求，从而可以进一步通过分类器进行正常训练，其具体包括过采样（oversampling）和欠采样技术（undersampling）。过采样对数据特征的依赖较大，而欠采样则可能会导致有用数据的丢失。改进 SVM 模型的方法主要对模型引入衡量样本重要程度的先验信息，同时尽可能不增加模型复杂度。结合这两种方法又可以设计不同的不平衡数据分类方法。

到目前为止，许多学者专注于抽样技术的研究。过采样方法通过增加少类样本的数量来弥补不平衡性。其中，复制少数样本最为简单，但其仅增加了少量样本的个数，而没有增加样本信息，这会导致分类器的决策区域变小，引起过学习。为克服这个问题，Chawla 等人提出了 SMOTE 算法[15]，算法利用启发式策略在少数样本和同类近邻的连线上进行过抽样。H. Han 等人在此基础上，进一步缩小少数类样本过抽样的区域，算法执行的区域范围限制为分类边界。一些学者又提出了一些结合方法，例如 Chawla 等人提出的 SMOTEBoost 算法、Seiffert. C 等人提出的 RUSBoost、Lianshengzhuang 等人提出的 Boost-BFKO 等，这些

方法的思路均是把样本处理与 Boosting 算法进行整合。文献[16]中提出了一种 SMOTE 的改进方法——边界 SMOTE，只对边界样本进行过采样。另外文献[17-20]也提出了几类不同的 SMOTE 改进方法，例如引入代价敏感(cost-sensitive method)学习，再训练 SVM，从而解决不平衡问题。S. Tang 等通过整合 SVM 和 SMOTE，最终实现过抽样[21]。

欠采样方法通过减少多类样本的数量来间接提高少类的分类准确率，例如可以通过随机地约减部分多类样本的方法进行欠采样，这种方法的缺点是会丢失多类样本的一些重要分类信息[22]。常用的欠采样方法，例如一致子集缩减、单边采样[23]等，均通过启发式策略对事先定义的距离测度进行欠采样。Dehmeshki 等人利用定义好的规则实现数据欠采样[24]，T. Yu 等人引入数据压缩中的矢量量化，提出了一种压缩后进行欠采样的方法[25]，这种方法同随机欠采样相比，对于样本集中的分类信息损失较少。同时能够消除与分界面距离较远或导致冗余的样本，从而在分类精度和算法效率之间达到较好的平衡。

与以上两种技术相比，基于 SVM 算法改进的方法能够充分利用现有的信息。例如，肖健华等设计了基于 SVM 惩罚因子参数选择改进模型[26]，这些参数保证分类器的推广性能和错分率之间的平衡。若惩罚因子太小可能会导致过多的错分样本，若惩罚因子值太大则可能会影响算法的泛化能力。Strohmann 等设计的 BMPM(biased minimax probability machine)算法通过两类样本均方差矩阵获得衡量样本重要程度的权值[27]，最终保证分类平衡，因此均方差矩阵的估计尤为重要。Wu 等人设计了核边界排列[28,29]，根据样本的几何特征修改核矩阵，最终实现平衡分类。但是该方法需要反复构造 SVM 分界面，然后通过支持向量估计

真实分类面，因此 KBA 方法计算速度较慢。文献[30]提出了基于单类 SVM 和异常检测学习矢量量化的方法，用来解决不平衡数据集对模型产生的影响。Fu 等人提出了基于样本加权的 SVM[31]。Ding 等人提同模糊粒度支持向量机算法。Batuwita 等人提出 FSVM-CIL 算法，通过利用模糊 SVM 的鲁棒性处理不平衡分类问题[33]。

此外，Su 等人提出 GSVM (granulation support vector machine) 算法[34]，该算法基于信息粒度的思想，不仅可以对平衡数据集进行分类，而且针对欠采样问题将多数和少数类样本的减少对分类性能的影响减少到最低，这对解决不平衡数据集的分类问题提供了一种新的思路。谢纪刚等提出了一种加权 Fisher 线性判别方法 WFLD(weighted fisher linear discriminant)[35]，这种方法本质上是一种特殊的过采样方法，对两类样本进行不同倍数的过采样来消除两类样本的不平衡。该方法虽然从数量上解决了数据不平衡，但是并没有从根本上解决过采样的问题。文献[36]设计了基于样本加权的粗糙集方法，通过加权熵度量不同属性中包含的信息量，提出了一种基于权重熵的属性约减算法。此外，还有引入模糊集和决策树[37,38]的方法。

重抽样尽管对分类性能有一定提升，但其影响了样本的分布特征。过抽样加入了新的数据，而欠抽样可能丢弃了一些关键数据，它们都会对分类面产生较大影响。已有的不平衡分类方法对不平衡程度较低的问题较为有效，而对高度不平衡问题尤其是线性不可分情况的分类性能有待提高。例如，高度不平衡数据集的几何分布特性难以利用，高度不平衡线性不可分数据集的特征映射本身容易受到不平衡数据的影响，从而会增加用线性方法解决不平衡问题的难度。所以，寻找一种有效的方法解决 HIO 数据

集的分类问题显得尤为重要。

1.2.2 核优化方法

目前，关于核学习方法的主要研究大致可以从核函数中参数的优化、多核学习及其大规模算法和快速非参数核学习三个侧面加以分析。

(1) 核函数中参数的优化

从时间上讲，随着支持向量机算法的发展与广泛应用，人们开始关注 SVM 中的模型选择问题。从概念上讲，SVM 的模型选择包括了 SVM 的核函数选择。SVM 参数选择的目的是选取适当的模型参数使得算法的性能最优。因此传统的核函数中的参数优化问题的常用解决方法主要包括交叉验证和最小化学习算法错误率的上界。

交叉验证技术的本质是一种穷举法，首先对参数进行组合，然后训练 SVM，通过不断修正参数，使分类错误率最小，最终得到优化参数。经典的交叉验证方法包括 k-折交叉验证(k-fold cross-validation)和留一法(leave-one-out，LOO)。留一法在理论上已被证明是关于真实错误率的无偏估计，k-折交叉验证则是留一法的推广，计算量相对较小。然而，在大数据集上，使用交叉验证技术选择核参数是没有效率的，当参数较多时更加难以实现。

为了克服交叉验证技术的不足，许多学者提出使用最小化学习算法错误率上界的方法来优化核参数。这些误差界包括 Radius-margin(RM)、GACV、Xi-Alpha 等[39-41]。其中，RM 界是较常用的一种误差界，但其只适用于硬间隔 SVM 和二范数软间隔 SVM。Yan 等人改进了 RM 界[41]，使其适用于一范数软间隔 SVM。常群等人基于改进的 RM 界进行了多个 Gaussian 核参数的优化[42]。Wang 等人将二分类下的 RM 界推广到了多分类的

情况[43]。这些基于 RM 界的方法通常采用基于梯度下降的优化算法优化核参数。和交叉验证技术相比，这种最小化学习算法错误率上界的方法进一步减少了计算量，在一定程度上解决了多参数优化问题。但是这种方法的问题在于通常只能取得核参数的局部最优解，并且算法的每一次迭代都需要训练 SVM 以及求解一个二次规划问题，因此同样不适用于大样本情况下的参数优化。

核参数优化的其他方法包括基于核函数（核矩阵）度量的核参数优化[44-52]、核路径优化算法[53]、基于特征空间簇间距的优化方法[54]、基于核相似性差异最大化的优化算法[55]等。这些优化算法在一定程度上提高了核参数优化的效率，但均为基于单个特征空间的单核方法。在解决实际问题的过程中，许多研究人员和学者发现需要处理异构数据源，在采用单个核函数处理这类问题时往往取得的结果不太理想。例如，假定处理的数据来自两个数据源，一个服从高斯分布，另一个服从多项式分布，这种情况下需要使用多个核函数来处理多源异构数据。

（2）多核学习及其大规模算法

2004 年，Lanckriet 等人[55,56]提出了基于多个基核组合形式的多核学习（multiple kernel learning，MKL）框架，随后，多核学习算法受到了许多学者和研究人员的广泛关注。多核学习理论的基本思想是：预先定义多个核函数，利用多个核函数隐式定义的多个非线性映射$\varnothing_j(x)(j=1,\cdots,m)$，将原始数据空间 X 中的向量 x 映射到多个特征空间$H_j(j=1,\cdots,m)$，然后在多个特征空间中根据不同的准则建立相应的算法。其中，多个核函数可以由不同类型的核函数或者由不同核参数的同一类型的核函数组成。因此，多核学习一方面增强了数据表示和描述的灵活性，另一方

面，通过求解优化模型的全局最优解可以获得各个基核的最优组合系数，从而提供了一种选择最优核函数的方法，在一定程度上解决了核方法中的核参数选择问题。在多核学习中，目前常用的准则函数主要有两类：一种是基于支持向量机准则的优化模型[57-59]；另一种是基于鉴别准则的优化模型[60-62]。在不同的优化模型中，研究的重点主要集中在以下三个方面：如何根据相应的准则设计取得全局最优解的凸优化模型，研究不同的基核组合形式对学习算法准确率的影响以及如何设计适用于大规模问题的多核学习算法。

在优化模型的设计方面，Lanckriet 等人以支持向量机的结构风险为优化目标函数[55]，将多核学习等价为求解一个半正定规划(semi-definite programming，SDP)问题。Yang 等人把稀疏化技术引入多核学习模型[63]，提出了一种稀疏广义多核学习模型，然后基于一组核函数在函数空间优化该模型。在此基础上，Gönen 等人通过定义一个参数门控模型，提出了局部多核学习的思想[57]。为了降低多核学习的计算复杂性，Lanckriet 等人将半正定规划优化模型放松为一个二次约束的二次规划(quadratic constrained quadratic programming，QCQP)问题[55]。Bach 等人进一步提出了 QCQP 的二阶锥规划(second order cone programming，SOCP)对偶形式[56]，并给出了对应的 SMO 算法。除了支持向量机准则，一些学者和研究人员提出了基于鉴别准则的优化模型。例如文献[61，62]基于核鉴别分析建立了一组异质核的稀疏多核学习模型，通过迭代算法解决了两类分类问题的核学习问题。目前，基于上述两种准则设计的多核学习算法计算复杂度较高，且算法主要为解决两类问题设计，直接拓展到多类问题则较为困难。

多核学习的另一个研究方向是研究如何合理地对基核进行组合。l_1-MKL 是最常见的一种基于多个基核凸组合的多核学习模型。使用这种模型求解的权系数具有稀疏性，即只有一少部分基核的权系数不为零。其优点在于可以减少数据特征冗余，提高运算效率。但对于一些正交性特征则会导致信息的丢失，从而降低学习的泛化能力。为此，Cortes 等在多核模型中引入了 l_2 范数[64]，提出了非稀疏多核学习模型 l_2-MKL。Han 等进一步将 l_2-MKL 推广到 $l_p(P>1)$ 范数多核学习[65]，即 l_p-MKL，从而增强了算法的鲁棒性和通用性。此外，一些研究人员还提出了基于权系数的混合范数(Mixed-norm)的多核学习模型[66]，Cortes 等人提出了基于权系数非线性组合的多核学习[67]。上述多核模型主要研究了不同的基核组合形式对学习算法准确率的影响，没有从根本上解决多核学习的高计算复杂度问题。

为了使多核学习适用于大规模问题，近年来，许多研究者提出了基于支持向量机准则的大规模多核学习算法。例如，Bach 等人将多核学习优化问题转换为对偶形式[56]，提出了基于 SMO (sequential minimal optimization)的多核学习方法，其本质是一种交替优化的方法，优化过程在基核的权系数和 SVM 训练之间交替进行，直至算法收敛。类似的方法还包括文献[68]提出的扩展的水平集(extended level set)方法、文献[69]提出的基于分组 Lasso 的方法等。这类方法均利用成熟的 SVM 算法进行快速训练，然后采用不同的方法更新基核的权系数。

大规模多核学习算法在一定程度上降低了多核学习算法的计算复杂度，但仍然存在以下问题：① 多核学习需要预先定义一组核函数，限制了其解决复杂模式分析问题的能力；② 多核学习算法大多基于支持向量机准则，面向两类问题设计，拓展到多类