摘 要

在很多实际应用中，随着数据采集技术和存储技术的发展，获取大量的无标号样本已变得非常容易，而获取有标号样本通常需要付出很大的代价。因而，相对于大量的无标号样本，有标号样本通常会很少。传统的无监督学习只能利用无标号样本进行学习，监督学习只利用少量的有标号样本学习，而半监督学习则能同时利用大量的无标号样本和少量的有标号样本来进行学习，因此，半监督学习是非常有意义的研究课题。半监督学习包括半监督分类、半监督回归、半监督聚类和半监督维数约减等几个方面。本文以半监督学习为基础，主要做了以下一些工作：

在半监督维数约减方面，提出了一种新的算法 ISSDR。一方面，它能够充分利用正负约束信息，使得在低维空间中不属于同一类的数据离的越远越好，而属于同一类的数据靠的越近越好。另一方面，引入剩余的大量未标记数据，利用隐藏在未标记数据中的潜在信息，能很好的保持数据集的全局以及局部结构。实验结果表明，该算法能从大量的未标记数据以及有限的成对约束中学习出有用的知识，实验证明了该算法的有效性。

在半监督分类方面，提出了一种新的集成算法 E-LNP。它选用一种基于图的半监督学习算法 LNP 作为子学习器，通过选择不同的特征个数以及学习参数，利用子学习器分别多次训练进行预测，然后将其预测结果按投票方式进行集成，从而得到最终的学习结果。实验表明 E-LNP 算法比仅使用单一的半监督分类器有更强的泛化能力，能有效的提高分类精度。

关键词：半监督学习　维数约减　集成学习

Abstract

In many practical applications, along with the development of data mining and data storing it is relatively easier to acquire large number of unlabeled data than labeled data, so there often amount of unlabeled data and few labeled data. The traditional unsupervised learning can only use unlabeled data and the supervised learning only use few labeled data. The superiority of semi-supervised learning is that it can use both unlabeled data and labeled data, so it's a very significance research subject. The semi-supervised learning contains semi-supervised classification, semi-supervised regression, semi-supervised clustering and semi-supervised dimensionality reduction, ect. Based on the semi-supervised learning, the following works are accomplished:

In terms of semi-supervised dimensionality reduction, a new algorithm ISSDR is proposed: On the one hand, it can preserve the positive and negative constraints, which makes distances in the transformed low-dimensional space between instances involved by the negative constraints set as large as possible, while distances between instances involved by the positive constraints as small as possible; On the other hand, it can also preserve the local and global structure by using the potential information in the numerous unlabeled data. Experimental results demonstrate that it can get helpful information from the constraints and the large amounts of unlabeled data, so it is an effectiveness method.

In terms of semi-supervised classification, a new ensemble learning algorithm E-LNP is proposed in this paper, and its base learners LNP (Linear Neighborhood Propagation) are selected from one of SSL learning approaches with graphs. We choose different input attributes and learning parameters to produce a series of component classified learners, and combine the predictions of these component learners via majority voting, at last attain the final learning prediction results. The experimental results show that the E-LNP algorithm performs better than just a single learner do, and it can improve the classification precision effectively.

Keyword: Semi-supervised learning Dimensionality reduction Ensemble learning

西安电子科技大学

学位论文创新性声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切的法律责任。

本人签名：赵玲玲　　日期 2010.1.20

西安电子科技大学

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律署名单位为西安电子科技大学。（保密的论文在解密后遵守此规定）

本学位论文属于保密，在___年解密后适用本授权书。

本人签名：赵玲玲　　日期 2010.1.20

导师签名：周宏生　　日期 2010.1.20

第一章 绪论

1.1 机器学习

学习，为什么会学习，如何更好的进行学习，学习的内在机理是什么，这些都是人类一直在积极探索的话题。学习能力是智能行为的一个非常重要的特征，但至今人们对学习的机理尚不完全清楚。使计算机具有学习能力，这是人工智能研究者几十年来一直梦寐以求的目标，机器学习研究的目的也是为此。

自从计算机问世以来，人们就想知道它们能不能自我学习。如果我们理解了它们学习的内在机制，即怎样使它们能够根据经验来自动提高自身的学习能力，那带来的影响将是空前的。想象一下，在未来，计算机能够从医疗记录中学习，获取治疗新疾病最有效的方法；住宅管理系统能够分析住户的用电模式，以降低能源消耗；个人助理软件能够跟踪用户的兴趣，为其选择最感兴趣的在线早间新闻。对计算机学习的成功实现将开辟出许多全新的应用领域，并使其计算能力和定制性上升到新的层次。

通过一些专项成果，我们可以看到机器学习这门技术的研究现状：计算机已经能够成功地识别人类的讲话、预测肺炎患者的康复率、检测信用卡的欺诈、在高速公路上自动驾驶汽车、以接近人类世界冠军的水平对弈西洋双陆棋等。在过去的几十年中，无论是应用、算法、理论，还是生物系统的研究，都取得了令人瞩目的发展和进步。

那么，什么是机器学习呢？

机器学习一般被定义为一个系统自我改进的过程。从最初的基于神经元模型以及函数逼近论的方法研究，到以符号演算为基础的规则学习和决策树学习的产生，以及之后的认知心理学中归纳、解释、类比等概念的引入，至最新的计算学习理论和统计学习的兴起，当然还包括基于马尔可夫过程的增强学习，机器学习一直都在相关学科的实践应用中起着主导作用，研究人员们借鉴了各个学科的思想来发展机器学习。

近几年来机器学习发展迅速，已经渗透到模式识别、计算机视觉、数据挖掘等多个领域，随着对计算机认识的日益成熟，机器学习将会在计算机科学和技术中扮演越来越重要的角色。

按照学习过程中有没有教师的参与，学习可以分为三种基本形式：监督学习、增强学习和无监督学习。而近年来随着机器学习的发展，出现了第四种学习形式，即本文中将要讨论的半监督学习。

监督学习：在监督学习方式中，存在一个“教师”，它可对一组给定的输入提供应有的输出结果，学习按照“教师”的监督信息进行着，如分类。

增强学习：增强学习的典型做法是，给定一个输入样本，计算它的输出类别，把它与已知的类别标记作比较，根据差异来改善分类器的性能。“教师”对这次分类任务的完成情况只给出“对”或“错”的反馈。

无监督学习：在无监督学习中并没有显式的“教师”，系统对输入样本“自动”形成聚类或自然的组织。

半监督学习：半监督学习是近年来机器学习领域新近提出的一种学习方式，它结合了监督学习和无监督学习两种学习方式，存在有限的“教师”信息，学习算法能借助这些少量的“教师”信息来更好的进行学。

1.2 半监督学习的研究背景及内容

随着互联网的普及，网上信息正在呈指数级增长。如何合理地组织这些信息，以便从茫茫的数据世界中检索到期望的目标，并有效地分析这些信息，以便挖掘出新颖和潜在的有用模式，正在成为网上信息处理的研究热点。网上信息的分类目录组织是提高检索效率和检索精度的有效途径，如在利用搜索引擎对网页数据进行检索时，若能提供查询的类别信息，必然会缩小与限制检索范围，从而提高查准率。同时，分类可以提供信息的良好组织结构，便于用户进行浏览和过滤信息。很多大型网站都采用这种组织方式，如 Yahoo[1]采用人工方式来维护网页的目录结构；Google 网站采用一定的排序机制，使与用户最相关的网页排在前面，便于用户浏览。Deerweser 等人[2]利用线性代数的知识，通过矩阵的奇异值分解来进行信息滤波和潜在语义索引，它将文档在向量空间模型中的高维表示投影到低维的潜在语义空间中，这一方面缩小了问题的规模，另一方面也从一定程度上避免了数据的过分稀疏现象，它在语言建模视频检索及蛋白质数据库等实际应用中取得了较好的效果。聚类分析是文本挖掘的主要手段之一[3]，它的主要作用是：(1)通过对检索结果的聚类，将检索到的大量网页以一定的类别提供给用户，使用户能够快速定位期望的目标；(2)自动生成分类目录；(3)通过相似网页的归并，便于分析这些网页的共性。K-均值聚类是比较典型的聚类算法，另外，自组织映射神经网络聚类和基于概率分布的贝叶斯层次聚类等新的聚类算法也正在不断地被研制与应用。然而大部分的这些聚类算法属于无监督学习，它对解空间的搜索带有一定的盲目性，因而聚类的结果在一定程度上缺乏语义特征。同时，在高维情况下，选择合适的距离度量标准变得相当困难。而网页分类是一种监督学习，它通过对一系列训练样本的分析来预测未知网页的类别归属，目前已有很多有效的算法来实现网页的分类，如 Naive Bayesian，SVM 等。遗憾的是，获得大量的带有

类别标注的样本的代价是相当昂贵的，而这些方法只有通过大规模的训练才能获得较高精度的分类效果。此外，在实际应用中，分类体系常常是不一致的，这为目录的日常维护带来了一定的困难。针对这些问题，Kamal Nigam 等人[4]提出从带有类别标注和不带有类别标注的混合文档中分类 Web 网页，它一方面克服了无监督学习中对求解空间搜索的盲目性；另一方面，它不需要对大量训练样本进行类别标注，只需提供相应的类主题变量，把网站管理人员从繁琐的训练样本的标注中解脱出来，提高了网页分类的自动性，这种方法就属于半监督学习算法。

随着数据采集技术和存储技术的发展，获取无标记样本已变得非常容易。另一方面，由于有标记样本的获取需要相关领域的专家对样本进行标记，因而相对比较困难而且代价昂贵。例如，在医学影像处理中，很容易从医院得到大量的医学影像，但医学专家不可能花费大量的时间和精力来对所有的病灶都进行标记，只能选择其中的很少一部分进行标记。所以在许多实际应用中，通常会有大量的无标记的样本，而有标记样本只占很小的比例。当用传统的监督学习方法（比如分类）来处理此类问题时，由于有标记样本较少，因而训练出来的分类器精度有限，同时仅使用少量“昂贵的”有标记样本而不利用大量“廉价的”未标记样本，则是对数据资源的极大的浪费。另一方面，传统的无监督学习方法（如聚类）则没有利用宝贵的已有标记样本指导聚类，因而限制了聚类性能的提高。因此，如何利用大量的未标记样本来辅助有限的标记样本来提高学习的精确性是面临的一个新问题。为了能更好地处理此类问题，目前在机器学习领域逐渐形成了一种介于传统监督学习和无监督学习之间的新的机器学习方法，称之为半监督学习（Semi-Supervised Learning）。与只利用有标记样本的监督学习以及只利用无标记样本的无监督学习相比，半监督学习同时利用有标记样本和无标记样本来学习。由于在诸如文本分类等实际问题上的出色性能，半监督学习已在国际上引起高度重视。半监督学习开始成为当前国际机器学习界的一大研究热点。近几年来随着机器学习在数据分析和数据挖掘中的广泛应用，半监督学习的理论研究成果已经部分地应用于实际问题的解决。

半监督学习包括半监督分类、半监督回归、半监督聚类和半监督降维。到目前为止关于半监督分类方法的研究相对较多，并且有很广泛的应用范围，包括网页分类、人脸识别、目标识别；信息检索，如形状信息检索和手写数字检索；图像分割等。相比之下，半监督降维方面的研究相对较少。然而，我们在现实生活中遇到的待处理数据大部分是高维空间中的样本集。比如随着数码产品的发展，我们可以得到更清晰的图片，但为此要付出的代价是更高维的数据，即更大的存储空间和更长的处理时间，这会给计算机处理带来更大负担。而且高维的数据中往往还带有大量的冗余信息，在机器学习中难以发现模型的真正变量。因此很多时候降维作为一种预处理过程是非常必要的。

1.3 国内外对半监督学习研究的进展及现状

半监督学习是模式识别和机器学习中的重要研究领域。近几年随着机器学习理论在数据分析和数据挖掘的实际问题，例如网页检索和文本分类、基于生物特征的身份识别、图像检索和视频检索、医学数据处理等问题中的广泛应用，半监督学习在理论和实际应用研究中都获得了长足的发展。半监督学习主要关注当训练数据的部分信息缺失的情况下，如何获得具有良好性能和推广能力的学习机器，这里的信息缺失涵盖数据的类别标签缺失或者存在噪声，数据的部分特征缺失等多种情况。半监督学习的理论研究对于我们深入理解机器学习中的许多重要理论问题，例如数据的流形与数据的类别信息的关系、缺失数据的合理处理、标注数据的有效利用、监督学习和非监督学习之间的联系、主动学习算法的设计等都有非常重要的指导意义。

自20世纪八九十年代以来国际机器学习界研究者在半监督学习研究领域展开了广泛深入的探讨和研究。其涵盖的范围非常广泛，例如半监督回归问题[5]；利用标签和特征维都缺失的数据集进行学习[6]；标签有噪声时的数据处理[7]；对各种监督学习算法进行修改，探讨如何融入非监督数据信息[8]或者对于非监督学习算法进行修改，探讨监督数据信息的引入[9]；利用有限混合模型对数据的概率分布进行建模或者利用其他模型对数据标签关于特征维的条件概率进行建模，利用 EM 算法学习模型参数的半监督学习的研究[10]；引入合适的数学方法进行半监督学习，例如基于核矩阵的谱分析[11]，高斯随机场的利用[12]，利用图论中的方法来对于样本集进行聚类分析[13]；半监督数据的流形分析[14]等。研究者同时开展了将半监督学习和传统模式识别和机器学习中的一些问题相结合的研究，例如基于半监督学习的特征提取[16]，半监督学习和集分类器的设计[17]等。国际研究者同时开展了与半监督学习有着密切关联的一些相关研究，具有代表性的是利用半监督数据和数据的不同特征子集在数据的不同视图上同时训练具有良好性能的学习机器[18]。

目前半监督学习的研究正在继续从广度和深度上不断进行扩展。就广度而言，一方面不断有各种传统的或者新提出的监督非监督算法的半监督情况下的修改算法出现，另一方面，不断有新的数学方法引入半监督学习。同时，半监督学习探讨的对象已经由简单的利用半监督数据训练扩展到半监督数据的流形分析，半监督数据和图模型的关系，半监督数据和基于核学习的关系。换而言之，半监督学习已经与当前机器学习研究的各热点和重点问题的研究紧密联系在一起。就深度而言，已经有许多研究探讨基于各种不同的有限混合模型的半监督学习的统一方法[19]，基于数据的特征视图的半监督学习机器的性能分析[20]，半监督学习和主动学习相结合提高学习机器性能[21]，半监督学习的聚类假设的显式数学表达[22]等等。

半监督学习的理论研究在未来的一段时间将一直是机器学习研究的重点和热点，这些研究对于我们理解学习机器的学习机理以及人机交互都具有重要的理论意义。

1.4 论文的研究内容及安排

半监督维数约减和半监督分类是半监督学习的重要方面，本文主要从以下两个方面做了一些工作：

第一，在对样本集进行维数约减方面，本文提出了一种新的半监督维数约减算法 ISSDR，它不仅能够分别利用正负约束对中的信息，而且还利用了所有剩余的大量未标记数据，通过将其分为邻域内和邻域外两个部分来挖掘隐藏在其中的潜在信息，实验结果表明，该算法能从大量的未标记数据以及有限的成对约束中学习出有用的知识，实验证明了该算法的优越性。

第二，在对样本集进行分类方面，本文提出了一种新的集成算法 E-LNP，它选用一种基于图的半监督学习算法 LNP 作为子分类器。每个子分类器分别选择不同的样本特征以及学习参数进行训练，然后将得到的训练结果用投票方式进行集成，从而得到最终的学习结果。实验表明 E-LNP 算法比仅使用单一的半监督分类器有更强的泛化能力。

本文的章节内容安排如下：

第一章，首先简要介绍了什么是机器学习以及机器学习的分类，引出了半监督学习的概念，接着介绍了半监督学习算法的研究背景及意义，分析了国内外研究进展及现状，并概述了本文的研究内容和章节安排。

第二章，对半监督学习理论进行了整体的介绍。首先介绍了半监督学习的基础知识，然后对已有分类算法的半监督学习框架和比较成熟的半监督学习算法分别进行了详细介绍。

第三章，从无监督，监督，半监督三个方面介绍了已有的维数约减方法，提出了一种半监督维数约减方法 ISSDR，同时进行了实验验证。

第四章，提出了一种基于集成算法的半监督算法。首先介绍了集成算法的相关知识以及 LNP 算法的内容，然后对提出的算法进行了详细的介绍并进行了实验验证。

第二章 半监督学习理论

2.1 半监督学习的基础知识

一般认为，半监督学习的研究始于 B. Shahshahani 和 D. Landgrebe 的工作[23]，但未标记样本的价值实际上早在上世纪 80 年代末就已经被一些研究者意识到了。D. J. Miller 和 H. S. Uyar[24]认为，半监督学习的研究起步相对较晚，可能是因为在当时的主流机器学习技术（例如前馈神经网络）中考虑未标记样本相对比较困难。随着统计学习技术的不断发展，以及利用未标记样本这一需求的日渐强烈，半监督学习才在近年来逐渐成为一个研究热点。

半监督学习的基本设置是给定一个来自某未知分布的有标记样本集 $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{|L|},y_{|L|})\}$，以及一个未标记样本集 $U=\{x_1',x_2',\cdots,x_{|U|}'\}$，期望学得函数 $f:X\to Y$ 可以准确地对样本 x 预测其标记 y。这里 $x_i,x_j'\in X$ 均为 d 维向量，$y_i\in Y$ 为样本 x_i 的标记，$|L|$ 和 $|U|$ 分别为 L 和 U 的大小，即它们所包含的样本数，一般情况下 $|L|<<|U|$。

2.1.1 半监督学习的有效性

在介绍具体的半监督学习技术之前，有必要先探讨一下为什么可以利用未标记样本来改善学习性能。关于这个问题，有不少研究者给出了解释。例如，D. J. Miller 和 H. S. Uyar[24]从数据分布估计的角度给出了一个直观的分析。假设所有数据服从于某个由 L 个高斯分布混合而成的分布，即

$$f(x|\theta)=\sum_{l=1}^{L}\alpha_l f(x|\theta_l) \tag{2-1}$$

其中 $\sum_{l=1}^{L}\alpha_l=1$ 为混合系数，$\theta=\{\theta_l\}$ 为参数。这样，标记就可视为一个由选定的混合成分 m_i 和特征向量 x_i 以概率 $\mathrm{P}(c_i|x_i,m_i)$ 决定的随机变量。于是，根据最大后验概率假设，最优分类由式(2-2)给出：

$$h(x_i)=\arg\max_k \sum_j \mathrm{P}(c_i=k|m_i=j,x_i)\mathrm{P}(m_i=j|x_i) \tag{2-2}$$

其中 $P(m_i = j \mid x_i) = \frac{\alpha_j f(x_i \mid \theta_j)}{\sum_{l=1}^{L} \alpha_l f(x_i \mid \theta_l)}$。

这样，学习目标就变成了利用训练样本来估计 $P(c_i = k \mid m_i = j, x_i)$ 和 $P(m_i = j \mid x_i)$。这两项中的第一项与类别标记有关，而第二项并不依赖于样本的标记。因此，如果有大量的未标记样本可用，则意味着能够用于估计第二项的样本数显著增多，这会使得第二项的估计变得更加准确，从而导致式(2-2)更加准确，也就是说，分类器的泛化能力得以提高。此后，文献[25]中进一步分析了未标记样本在半监督学习中的价值，并指出如果一个参数化模型如果能够分解成 $P(x, y \mid \theta) = P(y \mid x, \theta)P(x \mid \theta)$ 的形式，那么未标记样本的价值就体现在它们能够帮助更好地估计模型参数从而导致模型性能的提高。

2.1.2 半监督学习的两个基本假设

实际上，只要能够合理建立未标记样本分布和学习目标之间的联系，就可以利用未标记样本来辅助提高学习性能。在文献[24]中，这一联系是通过对生成式模型参数的估计来体现的，但在更一般的情况下就需要在某些假设的基础上来建立未标记样本和目标之间的联系。目前，在半监督学习中有两个常用的基本假设，即聚类假设和流形假设。

(1) 聚类假设

聚类假设是指处在相同聚类中的样本有较大的可能拥有相同的标记。根据该假设，决策边界就应该尽量通过数据较为稀疏的地方，从而避免把稠密的聚类中的数据点分到决策边界两侧。在这一假设下，大量未标记样本的作用就是帮助探明样本空间中数据分布的稠密和稀疏区域，从而指导学习算法对利用有标记样本学习到的决策边界进行调整，使其尽量通过数据分布的稀疏区域。聚类假设简单、直观，常以不同的方式直接用于各种半监督学习算法的设计中。例如，文献[26]中提出了 TSVM 算法，在训练过程中，该算法不断修改 SVM 的划分超平面并交换超平面两侧某些未标记样本的可能标记，使得 SVM 在所有训练数据（包括有标记和未标记样本）上最大化间隔，从而得到一个既通过数据相对稀疏的区域又尽可能正确划分有标记样本的超平面；文献[27]中通过修改高斯过程中的噪音模型来进行半监督学习，他们在正、反两类之间引入了“零类”，并强制要求所有的未标记样本都不能被分为零类，从而迫使学习到的分类边界避开数据稠密区域；文献[28]中通过使用最小化熵作为正则化项来进行半监督学习，由于熵仅与模型在未标记样本上的输出有关，因此，最小化熵的直接结果就是降低模型的不确定性，迫使决策边界通过数据稀疏区域。

(2) 流形假设

流形假设是指处于一个很小的局部邻域内的样本具有相似的性质，因此，其标记也应该相似。这一假设反映了决策函数的局部平滑性。和聚类假设着眼整体特性不同，流形假设主要考虑模型的局部特性。在该假设下，大量未标记样本的作用就是让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性，使得决策函数能够更好地进行数据拟合。流形假设也可以容易地直接用于半监督学习算法的设计中。例如，文献[12]中使用高斯随机场以及谐波函数来进行半监督学习，他们首先基于训练例建立一个图，图中每个结点就是一个（有标记或未标记）样本，然后求解根据流形假设定义的能量函数的最优值，从而获得对未标记样本的最优标记；文献[29]中根据样本相似性建立图之后，让样本的标记信息不断向图中的邻近样本传播，直到达到全局稳定状态。

值得注意的是，一般情形下，流形假设和聚类假设是一致的。由于聚类通常比较稠密，满足流形假设的模型能够在数据稠密的聚类中得出相似的输出。然而，由于流形假设强调的是相似样本具有相似的输出而不是完全相同的标记，因此流行假设比聚类假设更为一般，这使其在聚类假设难以成立的半监督回归中仍然有效。

2.2 几种已有分类算法的半监督学习框架

2.2.1 基于 SVM 的半监督学习

首先我们回顾传统的 SVM 优化问题，给定包含 l 个样本的训练集：(x_i, y_i)，$i=1,\ldots,l$，此处 $x_i \in R^n$，考虑二元模式识别问题，即 $y_i \in \{-1,+1\}$，支持向量机构造的分类超平面为 $w\cdot x+b=0$。该问题可以转化为求解下列的原始最优化问题：

$$\begin{aligned} &\min_{w,b,\eta} \quad C\sum_{i=1}^{l}\eta_i+\frac{1}{2}\|w\|^2 \\ &s.t. \quad y_i(w\cdot x_i-b)+\eta_i\geq 1, \eta_i\geq 0, i=1,\ldots,l \end{aligned} \tag{2-3}$$

若在给定的条件中另外加入新的工作集，考虑到 $x_j, j=l+1,\ldots,l+k$ 为未标记数据，S^3VM 优化问题可以描述如下：

$$\begin{aligned} \min_{w,b,\eta,\xi,z} \quad & C\left[\sum_{i=1}^{l}\eta_i+\sum_{j=l+1}^{l+k}\min(\xi_j,z_j)\right]+\|w\| \\ s.t. \quad & y_i(w\cdot x_i-b)+\eta_i\geq 1, \quad \eta_i\geq 0, \quad i=1,\ldots,l; \\ & w\cdot x_j-b+\xi_j\geq 1, \quad \xi_j\geq 0, \quad j=l+1,\ldots,l+k; \\ & -(w\cdot x_j-b)+z_j\geq 1, \quad z_j\geq 0. \end{aligned} \tag{2-4}$$

第二个不等式约束当点属于 1 的那一类，第三个不等式约束当点属于-1 的那类。目标函数为计算可能的错误分类的最小值，可以用整数规划来解这个问题，基本的思想是：对工作集中的每个x_j，加入取值为 0 或 1 的变量d_j。若$d_j=1$，则点属于 1 的那一类；若$d_j=0$，则点属于-1 的那一类，从而得到了以下的混合整数规划：

$$\begin{aligned} \min_{w,b,\eta,\xi,z} \quad & C[\sum_{i=1}^{l}\eta_i+\sum_{j=l+1}^{l+k}(\xi_j+z_j)]+\|w\| \\ s.t. \quad & y_i(w\cdot x_i-b)+\eta_i\geq 1 \quad \eta_i\geq 0 \quad i=1,\ldots,l \\ & w\cdot x_j-b+\xi_j+M(1-d_j)\geq 1 \quad \xi_j\geq 0 \quad j=l+1,\ldots,l+k \\ & -(w\cdot x_j-b)+z_j+Md_j\geq 1 \quad z_j\geq 0 \quad d_j=\{0,1\} \end{aligned} \tag{2-5}$$

选择足够大的常量$M>0$，使得当$d_j=0$时$\xi_j=0$可行，当$d_j=1$则$z_j=0$可行。

随着未标记样本数目的增加，计算量也迅速增加。为了解决计算量的问题，文献[30]中提出了 VS^{3}VM，将原来的非线性规划寻优问题转化为每次迭代为线性规划的迭代寻优问题以减少计算代价，使 S^{3}VM 可以处理未标记样本规模较大的问题。

2.2.2 基于核方法的半监督学习

文献[31]中提出了基于核的半监督学习，它的主要思想是对核矩阵的特征谱进行调整，使得属于同一聚类的样本点间距离更小而不同聚类的样本点间距离更大。调整后的核矩阵，再加上数据的标记信息就可以进行半监督学习。基于核矩阵的半监督学习的一个重要的思路是对于大量的未标记数据，可以获得边缘概率 P(x)或者在边缘概率已知的情况下获得能够应用于判决分类器（例如 SVM）的核矩阵，或者说利用边缘概率信息直接获得标记数据的核矩阵，然后利用该矩阵设计判决分类器。核的设计方法主要有如下几种形式：

a. 来源于混合模型的核：Seeger[32]设计了 Mutual Information 核。在单混合模型下，Fisher Kernel[33]是这种 Mutual Information 的近似；

b. 基于 Markov 随机游走的核；

c. 基于聚类表述的核：

(1) 计算相似阵(RBF 核矩阵)，对角元素为 0 而不是 1；

(2) D 为对角阵,对角元素为 K 的行或者列的和,计算 $L=D^{-1/2}KD^{-1/2}$；

(3) 寻找 L 的前 k 个特征向量；

(4) 对于点的坐标进行单位化处。

将基于核聚类的模型进行扩展，将几种不同的核矩阵的设计方法统一在一个

框架下，其基本算法步骤为：

(1) 从监督和非监督数据直接计算RBF相似阵K（对角设置为1），以及D，D为对角阵，对角元素为K的行或者列的和；

(2) $L=D^{-1/2}KD^{-1/2}$，其特征值分解为$L=U\Lambda U^T$；

(3) 设置变换方程φ，$\tilde{\lambda}_i=\varphi(\lambda_i)$，$\lambda_i$为$L$的特征值，重构$\tilde{L}=U\tilde{\Lambda}U^T$；

(4) $\tilde{D}$为对角阵而且$\tilde{D}_{ii}=1/\tilde{L}_{ii}$，计算$\tilde{K}=\tilde{D}^{1/2}\tilde{L}\tilde{D}^{1/2}$。

于是根据不同的变换方程，有不同的核：

a. 线形$\varphi(\lambda)=\lambda$相当于无变换。

b. 单步核：如果$\lambda\geq\lambda_{cut}$，$\varphi(\lambda)=1$，否则为0。该结果就是前文的基于聚类表述的核方法。

c. 线形单步：$\lambda\geq\lambda_{cut}$，$\varphi(\lambda)=\lambda$，否则为0。

d. 多项式$\varphi(\lambda)=\lambda^t$，那么有$\tilde{K}=\tilde{D}^{1/2}D^{-1/2}(D^{-1}K)^tD^{-1/2}\tilde{D}^{1/2}$，该矩阵可以认为是Markov随机游走表述核矩阵$D^{-1}K$的归一和对称化结果。

对于新来的测试样本，可以利用如下方法进行分类，将其理解为已有标记样本和未标记样本的线形组合，那么：

$$\alpha_0=\arg\min_{\alpha}\left\|\Phi(x)-\sum_{i=1}^{n+n_r}\alpha_i\Phi(x_i)\right\|=K^{-1}\nu \tag{2-6}$$

然后，我们有新的表述，并基于该表述分类：

$$\tilde{K}(x,x_i)=(\tilde{K}\alpha_0)_i=(\tilde{K}K^{-1}\nu)_i \tag{2-7}$$

2.2.3 基于K均值的半监督学习

这一类方法主要是将监督样本纳入K均值流程，在K均值算法中对监督样本的使用进行一些限定，同时利用这些监督样本来确定各个聚类的类别，它属于比较直观的利用监督样本帮助非监督聚类的算法。文献[34]中提出利用问题的背景信息以及数据间的必然联系(Must-link)和不能联系(cannot-link)来指导K-means聚类（COP-KMeans），算法流程为：

(1) 设置初始聚类；

(2) 分配点到不违反限制的最近类中，如果没有这样的类，返回失败；

(3) 更新类别中心；

(4) 重复上面步骤(2)-(3)，直至收敛。

文献[35]中进一步提出两种半监督的 K-均值聚类，主要想法是利用监督样本产生初始聚类种子,同时利用监督样本的约束指导聚类过程。在某些样本集上的实验说明，这些算法比在 COP-KMeans 半监督学习算法要好。首先对于考虑 K 均值的另一种表述－SPKMeans。如果点 X=$\{x_1,\cdots,x_N\}$，$x_i \in R^d$，那么 KMeans 产生一个 K 分割，$\{X_l\}_{l=1}^K$,$\{\mu_1,\cdots,\mu_K\}$为各分割中心。用 L_2 范数对样本点和中心进行归一化，这样，可以将目标函数形式变化，转化为最大化目标函数：

$$J_{spkmeans} = \sum_{l=1}^{K} \sum_{x_i \in X_l} x_i^T \mu_l \tag{2-8}$$

那么两种半监督算法流程描述如下：

a. Seeded-KMeans：输入为样本点和分割的数目 K，初始的种子集合 $S = \cup_{l=1}^{K} S_l$，输出为目标函数最优的 K 分割：

(1) 初始化：$\mu_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{x \in S_h} x, h = 1,\ldots,K; t \leftarrow 0$；

(2) 重复下列步骤直至收敛：

① 分配聚类：不考虑 seed 集合中的点，根据 h^* 重新分配点到各类，

$h^* = \arg\min_h \left\| x - \mu_h^{(t)} \right\|^2$；

② 重新估计均值。

b. Strained-KMeans：输入为样本点和分割的数目 K，初始的种子集合 $S = \cup_{l=1}^{K} S_l$，输出为目标函数最优了的 K 分割：

(1) 初始化：$\mu_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{x \in S_h} x, h = 1,\ldots,K; t \leftarrow 0$；

(2) 重复下列步骤直至收敛

① 分配聚类：重新分配点到各类 h^* 中，如果 $x \in S_h$，直接将其放入 h 聚类，否则根据 $h^* = \arg\min_h \left\| x - \mu_h^{(t)} \right\|^2$；

② 重新估计均值。

2.3 四类半监督学习算法

近年来国内外对半监督学习算法开展了大量的研究，取得了很多研究成果[15]。下面分别从四个方面来进行介绍。

2.3.1 基于 EM 的半监督学习算法

EM(Expectation Maximization)算法是一种迭代算法，是 1977 年提出的求参数极大似然估计的一种方法[36]，它可以从非完整数据集中对参数进行最大似然估计，是一种非常简单实用的学习算法。这种方法可以广泛地应用于处理缺损数据，截尾数据，带有讨厌数据等所谓的不完全数据。

基本原理可以表述如下：我们可以观察到的数据是 Y，完全数据 $X=(Y,Z)$，Z 是缺失数据，θ 是模型参数。θ 关于 Y 的后验分布 $p(\theta|Y)$ 很复杂，难以进行各种不同的统计计算。假如缺失数据 Z 已知，则可能得到一个关于 θ 的简单的添加后验分布 $p(\theta|y,z)$，利用 $p(\theta|y,z)$ 的简单性我们可以进行各种统计计算。然后，回过头来，又可以对 Z 的假定作检查和改进。如此进行，就将一个复杂的极大化抽样问题转化为一系列简单的极大化或抽样问题。

基于 EM 的半监督学习算法是最早的半监督学习方法，在半监督学习的问题中，由于大部分样本是未标记样本，也可以把它们看成不完整的数据。它的主要思想是：首先由有限的标记样本训练一个分类器，再通过计算未标记样本类标的期望对其进行加权概率标记，然后在每一次迭代中使用所有的样本及之前得到的类标训练一个新的分类器。可以利用它对文本进行分类人脸定位判别任务，然而文献[37]中指出这种方法很受模型假设的限制。

下面介绍一种基于双重高斯混合模型的半监督 EM 学习算法。这里的双重高斯混合模型指的是在全体学习样本的概率分布中，上一重为大高斯，拟定这些样本符合高斯混合分布，并且样本的类别数就是高斯数。在每一个样本里，又都分别含有一个高斯混合模型，也就是下一层小高斯，小高斯里样本的类别数也就是高斯数。所以假设定义了全体学习样本的概率分布，在第一重高斯中，高斯数是样本的类别数，设类别数是 M，则第一重中需要学习的参数是 $\{\alpha_1,\cdots,\alpha_M\}$，即每个高斯(每个类别)的先验概率，并且满足 $\sum_{i=1}^{M}\alpha_i=1$；在第二重中，每个高斯的子高斯数相等并且都为 n，需要学习的参数是 $\{\alpha_{1|1},\cdots\alpha_{H_1|1},\cdots\alpha_{1|M},\cdots\alpha_{H_M|M}\}$ 以及 $\{\mu_{11}\cdots\mu_{1H_1},\cdots\mu_{M1},\cdots\mu_{MH_M},\Sigma_{11}\cdots\Sigma_{1H_1},\cdots\Sigma_{M1},\cdots\Sigma_{MH_M}\}$。其中 H_i 表示第 i 个高斯包含的高斯数，即为 n；$\alpha_{j|i}$ 表示第 i 个高斯的条件下，第 j 个子高斯的条件先验概率，并且满足 $\sum_{j=1}^{H_i}\alpha_{j|i}=1$，$i\in\{1,2,\cdots M\}$。$\mu_{ij}$ 和 Σ_{ij} 分别表示第 i 个高斯的第 j 个子高斯

的均值向量和方差矩阵。从而，双重高斯混合模型参数定义为

$$\Theta = \{(\alpha_1,\cdots,\alpha_M),(\alpha_{1|1},\cdots\alpha_{H_1|1},\cdots\alpha_{1|M},\cdots\alpha_{H_M|M}),$$
$$(\mu_{11}\cdots\mu_{1H_1},\cdots\mu_{M1},\cdots\mu_{MH_M},\Sigma_{11}\cdots\Sigma_{1H_1},\cdots\Sigma_{M1},\cdots\Sigma_{MH_M})\}$$

假设学习的样本集是由部分已标记的和大部分未标记的样本组成，即 $S = S^l + S^u$，

$$S = \{(X_1,y_1),\cdots,(X_L,y_L),X_{L+1},\cdots,X_{L+U}\}$$

对应于未标记的样本 $X_{L+1},\cdots,X_{L+U}$ 的标记为 $y_{L+1},\cdots,y_{L+U}$。整个训练样本集 S 共有 M 个类，y_i 是其相应的类标，所以 $y_i \in \{c_1,c_2\cdots,c_M\}$，每一个 X_i 可以表示为一个在 d 维空间内的由 N 个特征向量集组成的有 K 个成分的高斯混合模型

$$X_i =< x_1,x_2,\cdots,x_N >, x_j \in R^d$$

每个 S 中的 (X_i,y_i)，认为它们都是相互独立且同分布的。对于每个未标记的样本 X_i，定义表示类别数目的 M 个隐含变量 z_{ij}，$j=1,2,\cdots M$，

$$z_{ij} = \begin{cases} 1 & if\ y_i = c_j \\ 0 & otherwise \end{cases} \tag{2-9}$$

利用已标记的样本，计算出初始的迭代参数 θ^0。然后再按照如下的步骤迭代计算：

[E-step] Set $\hat{z}^{(t+1)} = E[z \mid S;\hat{\theta}^{(t)}]$

[M-step] Set $\theta^{(t+1)} = \arg\max_\theta P(S,z^{(t+1)} \mid \theta)$

在 E-step 中，根据最大后验概率，给每个未分类的样本一个类别标签。可以用式(2-10)来计算求得 z：

$$\begin{aligned} E[z_{ij} \mid S;\theta] &= P(y_i = c_j \mid S;\theta) = P(c_j \mid X_i;\theta) \\ &= \frac{p(X_i \mid c_j)P(c_j)}{p(X_i)} \\ &= \frac{[\prod_{n=1}^{N}(\sum_{k=1}^{K}\partial_{jk}p(x_n \mid \mu_{jk},\Sigma_{jk}))]^* P(c_j)}{\sum_{j=1}^{M}[[\prod_{n=1}^{N}(\sum_{k=1}^{K}\partial_{jk}p(x_n \mid \mu_{jk},\Sigma_{jk}))]^* P(c_j)]} \end{aligned} \tag{2-10}$$

其中 K 是第 j 类中混合成分的数目，μ_{jk} 是混合权重。$p(x \mid \mu_{jk},\Sigma_{jk})$ 是多元高斯混合分布的概率密度函数。在 M-step 中，根据新分配的类标和原有的类标，按照最大似然，重新计算参数 θ_j。

高斯混合模型的参数 $\theta_j = \{\partial_{jk},\mu_{jk},\Sigma_{jk}\}$ 是由 EM 算法来估计的，对任何一个未标记的 X，通过计算最大后验概率来确定其所属类别。由此，我们可以计算出后

验概率 $P(c_j \mid X)$，使得：

$$j^* = \arg\max_j P(c_j \mid X) \tag{2-11}$$

2.3.2 增量半监督学习算法

在许多时候，训练样本不能一次获得。当新的训练样本到来时，若对所有训练样本进行重新学习，计算开销将会很大，同时，为了保存原训练样本也需要相当的存储空间。神经网络中的增量学习便能解决此类问题，它利用已有训练样本完成学习后能够通过不断学习新的样本来提高分类能力，适应“只知部分不知全局”的情况和动态的环境。通常，增量学习算法需要满足两点：(1)当分类器遇到新的样本时，能够学习其新的信息；(2)分类器学习新的信息时，不会或尽可能不忘记已经学过的知识。

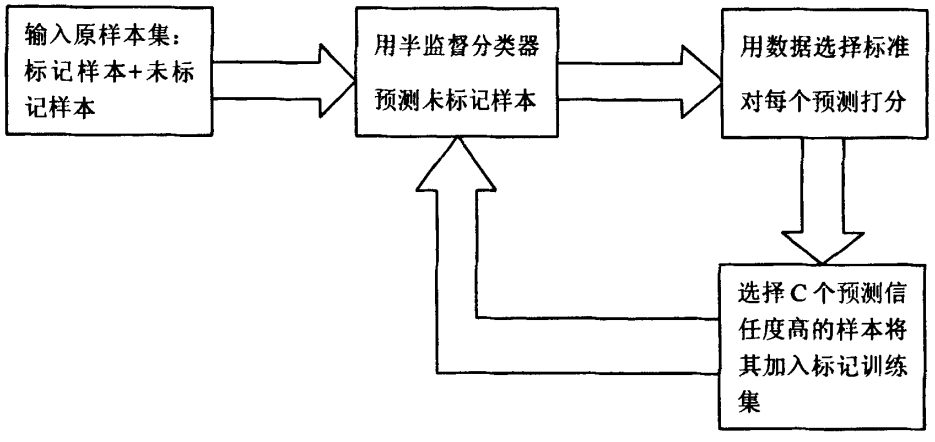

图 2.1 增量学习的实现过程

半监督学习中的增量学习算法是基于以下事实的：一般情况下，提供的监督信息越充分，即标记的样本越多，对未标记样本的预测就越准确，这一点在实际问题的应用中也得到了验证[38]。如图 2.1 所示，它是以迭代方式来实现的：在每次迭代时选择一些预测信任度较高的样本加入到标记样本集，从而为下一次的迭代做准备，如此迭代直到满足某一终止条件。所以，如何选择一个可分性好的分类器及如何构造一个合适的计算预测信任度的准则是非常重要的。

常见的增量学习算法包括自训练算法和协同训练算法。

自训练算法的主要思想是：首先使用标记样本训练一个分类器，其次用它来分类未标记样本，然后选出信任度较高的未标记样本及其对应的预测类标加入训练集进行重新训练，重复此过程直到满足某一终止条件。它可以被应用于一些自

然语言处理任务，词义消歧，主观名词识别，图像中的目标检测等。

协同训练算法的基本思想是：首先分别训练标记样本集的两个子特征集合得到两个不同的分类器，再用每个分类器标记未标记样本，然后分别用由其中一个分类器输出的预测信任度较高的样本来“教导”另一个分类器。每个分类器由另一个分类器提供的部分训练样本重新训练，如此重复这个过程。协同训练算法基于很强的假设：(1)样本特征可以分成两个集合；(2)每个特征子集合都能足够用来训练一个好的分类器；(3)当给定类标的条件下两个子集合是相互条件独立的。

2.3.3 基于图的半监督学习算法

基于图的半监督学习算法把问题定义成一个图，其中顶点表示样本，边(可能带权值)则表示样本的相似性。图方法是无参数的判别方法，大部分的基于图的方法是求解一个关于图的函数，该函数必须同时满足两个条件：(1)在标记样本点上必须近似等于给定的类标；(2)在整个图上该函数是光滑的。然后，定义所需优化的目标函数并使用决策函数在图上的光滑性作为正则化项来求取最优模型参数。文献[39]提出把半监督学习问题等价为图最小分割问题，而最小分割仅给出了硬分类，并没有计算边缘概率。针对这个问题，文献[40]中试着计算离散马尔可夫随机场的边缘概率，而高斯随机场和调和函数[12]实质上是将离散马尔可夫随机场连续化。基于图的半监督学习方法可等价为优化一个带调整项的损失函数。所以很多这类方法的差别实质上是损失函数或调整函数的不同，于是出现了局部和全局一致性[29]，Tikhonov 正规算法[41]和流形正规算法[42]。Wang 等人[38]提出了一种有效的线性邻域传播 LNP 算法，并用它来解决人脸识别、物体识别和图像分割等问题，文章将在第四章对此算法进行详细的介绍。下面介绍一种基于图从局部和全局进行的半监督学习算法[29]。

设 $X=\{x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_m\}\subset R^d$ 是 m 个样本的集合，每个样本的维数是 d，$L=\{L_1,L_2,\cdots,L_c\}$ 为样本点的 c 种类别。记 $(x_1,y_1),\ldots,(x_l,y_l)$ 为给定的 l 个已标记样本，$y_i\in L$ 为 x_i 的标记，$(x_{l+1},\ldots,x_{l+u})$ 为 u 个未标记样本，其中 $m=l+u$，一般情况下 $l\ll u$。

设 $\mathcal{F}$ 为非负 $n\times c$ 矩阵，数据集 $F=[F_1,\ldots,F_n^T]\in\mathcal{F}$ 是每个样本点 x_i 对应标记的集合。定义一个 $n\times c$ 矩阵 $Y\in\mathcal{F}$，若 x_i 为已标记点且其标记为 $y_i=j$，那么 $Y_{ij}=1$，否则 $Y_{ij}=0$；若 x_i 为未标记点，则 $Y_{ij}=0$（$1\le j\le c$）。可以看出 Y 和初始标记有关，下面是算法的步骤：

(1) 计算权矩阵 W，若 $i\neq j$，$W_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$，否则 $W_{ii}=0$。

(2) 构造矩阵 $S=D^{-1/2}WD^{-1/2}$，其中 D 是对角矩阵，对角线元素 (i,i) 为 W 的第 i 行的元素之和。

(3) 迭代 $F(t+1)=\alpha SF(t)+(1-\alpha)Y$ 直至收敛，这里 α 为 $(0,1)$ 之间的一个参数。

(4) 设 F^* 为数列 $\{F(t)\}$ 的收敛点，则每个样本点 x_i 标记为 $y_i=\arg\max_{j\le c} F^*_{ij}$。

可以证明此数列收敛且收敛到 $F^*=(1-\alpha)(I-\alpha S)^{-1}Y$。

对上面的迭代算法进行推导可形成一个规则化框架。与 F 有关的损失函数定义为：

$$Q(F)=\frac{1}{2}\left(\sum_{i,j=1}^{n}W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}F_i-\frac{1}{\sqrt{D_{jj}}}F_j\right\|^2+\mu\sum_{i=1}^{n}\|F_i-Y_i\|^2\right) \tag{2-12}$$

这里 $\mu>0$ 为规则化因子，则分类函数为

$$F^*=\arg\min_{F\in\mathcal{F}}Q(F) \tag{2-13}$$

损失函数中右侧第一项为光滑约束，这意味着一个好的分类函数在邻点间改变不能太大。第二项为适应约束，它意味着一个好的分类函数与初始标记设置相比改变太大，它不仅包含标记样本而且包含未标记样本。参数 μ 用来均衡这两个约束。

对 $Q(F)$ 关于 F 求导，得到

$$\frac{\partial Q}{\partial F}|_{F=F^*}=F^*-SF^*+\mu(F^*-Y)=0 \tag{2-14}$$

可以转化为

$$F^*-\frac{1}{1+\mu}SF^*-\frac{\mu}{1+\mu}Y=0 \tag{2-15}$$

设 $\alpha=\frac{1}{1+\mu}$，则 $(I-\alpha S)F^*=(1-\alpha)Y$。由于 $I-\alpha S$ 可逆，所以得到：

$$F^*=(1-\alpha)(I-\alpha S)^{-1}Y \tag{2-16}$$

2.3.4 直推式支持向量机

直推式支持向量机 TSVM（Transductive Support Vector Machine）是标准 SVM 算法在未标记样本上的一种扩展，对二分类问题而言，标准 SVM 是仅使用了标记样本，它的目标是在重构核希尔伯特空间中寻找一个个最优超平面使两类样本间

的分类间隔最大。而 TSVM 则同时使用已标记样本和未标记样本（经 TSVM 学习后其标记将变为已知），新找到的最优分类边界应该满足对原始的未标记样本的分类具有最小的泛化误差。图 2.2 给出了 SVM 和 TSVM 在半监督学习问题中的最优分类边界示意图。

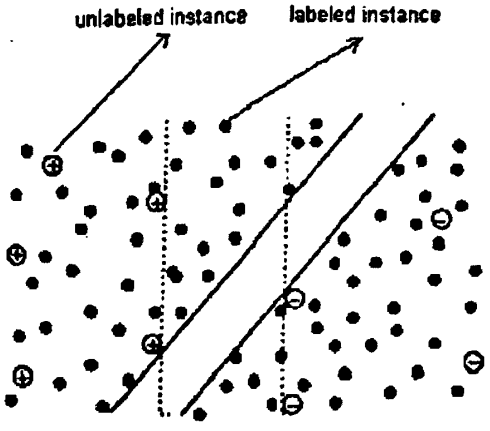

图 2.2 TSVM 和 SVM 算法最优分类边界示意图

图中的虚线表示标准 SVM 对原始的已标识样本学习后得到的分类边界，实线则表示 TSVM 对原始的已标识样本和未标识样本学习后得到的分类边界。图中实心黑点表示未标记样本，圆圈表示已标记样本，虚线上的圆圈表示已标记样本中的支持向量，实线上的黑点则表示未标记样本中的支持向量给定一组独立同分布的有标记训练样本点。

给定一组独立同分布的有标记训练样本点

$$(x_1,y_1),\ldots,(x_l,y_l),x_i\in R^m,y_i\in\{-1,+1\}$$

和另一组来自同一分布的无标记训练样本点

$$x_1^*,x_2^*,\ldots,x_k^*$$

在一般的线性不可分条件下，TSVM 的训练过程可以描述为以下的优化问题:

$$\begin{aligned}&\min_{y,w,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{k}\xi_j^*\\&s.t.\quad y_i[w\cdot x_i+b]\geq 1-\xi_i\quad \xi_i\geq 0\quad i=1,\ldots,l\\&\qquad y_j[w\cdot x_j^*+b]\geq 1-\xi_j^*\quad \xi_j^*\geq 0\quad j=1,\ldots,k\end{aligned}\tag{2-17}$$

其中参数 C 和 C^* 为用户指定和调节的参数，参数 C^* 是未标记样本在训练过程中的影响因子，$C^*\xi_j^*$ 称为未标记样本 x_i 在目标函数中的影响项。

该训练算法大致可以分为以下几个步骤：

(1) 指定参数 C 和 C^*，使用归纳式学习对标记样本进行一次初始学习，得到一

个初始分类器。并按照某个规则指定一个未标记样本中的正标签样本数 N。

(2) 用初始分类器对未标记样本进行分类，根据对每一个未标记样本的判别函数输出，对输出值最大的 N 个未标记样本暂时赋正标记值，其余的赋负标记值.并指定一个临时影响因子 C^*_{tmp}。

(3) 对所有样本重新训练，对新得到的分类器，按一定的规则交换一对标签值不同的测试样本的标签符号，使得优化问题(2-17)中的目标函数值获得最大下降。这一步骤反复执行，直到找不出符合交换条件的样本对为止。

(4) 均匀地增加临时影响因子 C^*_{tmp} 的值并返回到步骤(3)，当 $C^*_{tmp} \geq C^*$ 时，算法结束，并输出结果。

步骤(3)中的标记符号交换保证了交换后的解优于交换前的解，而步骤(4)中的临时影响因子由小到大逐步递增，它试图通过逐渐增加无标记样本对算法的影响来尽可能减小未标记样本的分类误差，由于步骤(1)中指定的 C^* 是有限的，由步骤(4)中的结束准则可知，算法能够在有限次执行后终止并输出结果。

这个算法比单一使用有标记样本训练的分类器有较大的性能改进,但存在如下的缺点：在算法执行之前必须人为指定待训练的未标记样本中的正标记样本数 N，而在一般情况下 N 值是很难做出比较准确的估计的，在 TSVM 算法中采用了一种简单的方法，即根据有标记样本中的正标记样本所占比例来相应估计未标签样本中的正标签样本比例，进而估计出 N 值。不难看出，这一方法在有标记样本数较少的情况下很容易导致较大的估计误差，一旦事先设定的 N 值和实际上的正标记样本数相差较大，就会导致学习机性能的迅速下降。为了提高分类准确率该算法往往要进行多次回溯式的学习，但这样又会造成训练时间的增加。陈毅松等人提出了一种改进算法 PTSVM[43]，该算法通过成对标记和标记重置的办法改进了 TSVM 的性能,但只适合于未标记样本较少的情况。

2.4 本章小结

本章首先从半监督学习的有效性以及基本假设两个方面介绍了半监督学习的基本知识，然后介绍了几种已有分类算法的半监督学习框架，它们是在已有的监督或无监督学习算法的基础上提出的。最后介绍了现今研究的比较多的半监督学习算法，它们分别为：基于 EM 的半监督学习算法，增量半监督学习算法，基于图的半监督学习算法以及直推式支持向量机。

第三章 维数约减算法

很多信息处理领域都面对着大量的高维数据，例如机器学习、数据压缩、模式识别、特征提取和数据可视化等。在日常生活中，我们的大脑也同样需要处理大量的高维输入信息：约 3 万个听觉神经纤维输入和 100 万个视觉神经输入。比如当我们在处理人脸数据时，我们需要同时研究不同姿态、不同光照等条件下的一个人的脸部信息，每张图片的像素都可以看成是一个上千甚至上万维的向量。我们如何从这些纷杂的输入中提取有用的信息，然后又怎样根据这些信息进行判断呢？在科学研究中，发现隐藏在高维数据观测值中的有价值的低维结构信息是一个必须解决的问题，这个过程称为降维，也叫“维数约简”。

一般的，维数约简是指将样本从高维观测空间，通过线性或非线性映射投影到一个低维特征空间，从而找出隐藏在高维观测数据中有意义的低维结构。降维的目的就是获得原始数据的一个更精简的表示，为进一步的处理，如分类、降噪、插值和可视化等做准备。根据对样本点的标记的利用情况，本章从无监督维数约减，监督维数约减和半监督维数约减三个方面来进行常见算法的介绍。

3.1 无监督维数约减算法

3.1.1 主成份分析 PCA

主成份分析 PCA（Principle Component Analysis）作为一种降维算法，它并不需要判别信息，是最小均方意义下最能代表原始数据的投影方法，是无监督学习的一个典型代表。PCA 是一种常用的线性维数约减的方法，一般认为，在线性维数约减算法中，还没有一种算法超越 PCA。

PCA 拥有欧氏空间向量基最小的特性。简单来说就是，当我们用 PCA 把数据降维后，从传统的向量范数来看，误差会是最小的，因此它更适合做资料压缩。

PCA 算法可以从两个不同的角度进行定义。一方面，PCA 可定义成数据在低维线性空间或主子空间上的正交投影，使得投影后的数据的方差最大。另一方面，PCA 也可以定义成一个线性投影，使得平均投影损失(即数据点和其投影之间的均值平方距离)最小。这里从最大化投影数据方差这个角度来介绍 PCA，我们假设 n 个数据，每个维数为 p，投影后的数据维数为 $d(d<p)$ 维。先来考虑 $d=1$ 的情况，我们使用一个 p 维的列向量 u_1 来定义投影方向，那么每个数据 x_i 被投影成一个标量值 $u_1^T x_i$，而原始数据投影后的方差为

$$\frac{1}{n}\sum_{i=1}^{n}(u_1{}^T x_i - u_1{}^T \bar{X}) = u_1{}^T S u_1 \tag{3-1}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T \tag{3-2}$$

其中 $\bar{X}$ 是总样本均值，不失一般性，我们仅考虑投影向量的方向而不考虑其大小，令 $u_1{}^T u_1 = 1$，并在此约束条件下，关于投影向量 u_1 来最大化投影方差。这里使用拉格朗日算子，可得到

$$Su_1 = \lambda_1 u_1 \tag{3-3}$$

这里 u_1 是 S 的特征向量。由于 $u_1{}^T u_1 = 1$，又可以得到

$$u_1{}^T S u_1 = \lambda_1 \tag{3-4}$$

所以当我们选取投影向量等于最大特征值 λ_1 所对应的特征向量时，降维后数据的方差将最大，这个向量也被称为第一主成份。同样的，当 $d > 2$ 时，最优的线性投影方向也就是投影方差矩阵的 $d(d < p)$ 个最大的特征值所对应的特征向量。

下面是具体的计算步骤：

(1) 计算协方差矩阵 $S = \mathrm{cov}(X - \bar{X})$；

(2) 计算特征值与特征向量：解特征方程 $|\lambda I - R| = 0$，求特征值并使其按大小顺序排列 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，然后分别求出对应于特征值 λ_i 的特征向量 u_i，这里要求 $\|u_i\| = 1$。

(3) 确定主成分个数：当累计贡献率 $\eta_d = (\lambda_1 + \lambda_2 + \ldots + \lambda_d)/(\lambda_1 + \lambda_2 + \ldots + \lambda_p)$ 大于某个域值时，可认为主成分数目为 d。

(4) 求主成分得分：新的变量值 $Y_{n\times d} = X_{n\times p} U_{p\times d}$

可以看出 PCA 将原来众多具有一定相关性的指标重新组合，用新的相互无关的综合指标来代替原来指标，从而具有重要的优良性质。在对原始数据集进行 PCA 变换后，特征根据信息量大小排序，信息量大的特征排在数据集的前面部分，数据集后面部分特征对整个数据集信息量很小。去除这些信息量小的特征能减少算法运算的复杂度，提高该数据集的质量，还能提高算法的预测精度。

3.1.2 局部线性嵌入 LLE

由于直接分析分布在高维空间中的数据是比较困难的，而学习数据集分布的

流形有助于发现高维数据集分布的内在规律及实现数据的可视化，因此近几年来对流形学习算法的研究越来越流行。流形学习的目的就是寻找一个能够保持高维观察数据的局部几何特性的低维简洁表示。

LLE 算法就是流形学习中典型的一种，是一种无监督降维算法，它能够找到嵌入在高维数据空间中的低维光滑流形。虽然其名字中带有“线性”，但事实上它是一种非线性降维方法，它主要利用局部线性来逼近全局非线性，保持局部的几何结构不变，通过相互重叠的局部邻域来提供整体的信息，从而保持整体的几何性质。

LLE 算法是基于几何直觉的，把高维空间中的点按维数映射到低维嵌入空间，即 $x_i \to y_i$。它首先寻找数据点 x_i 的邻近数据点，然后计算该数据点的局部重构权值矩阵 W_{ij}，最后由局部重构权值矩阵及其邻域内的数据点计算低维输出向量 y_i。其具体实现如下：

(1) 寻找数据点 x_i 的邻近数据点

利用欧氏距离作为测度寻找每个数据点 x_i 的 k 个最近邻数据点，其组成的集合用 $N(x_i)$ 来表示。则每个数据点的重构误差用成本函数来衡量，可表示为

$$\min \varepsilon_i = \min_{W_i} \left\| x_i - \sum_{x_j \in N(x_i)} W_{ij} x_j \right\|^2 \tag{3-5}$$

(2) 计算权值矩阵

式(3-5)中的权值 W_{ij} 表明第 j 个数据点对重构第 i 个数据点的权值。当两个数据点越近或越相似时，这个权值也就越大。为了得到合适的权值，需要两个限制条件：

①每个数据点只能通过它邻域内的数据点来构造，并且当某个数据点不属于此邻域内时 $W_{ij} \approx 0$。

②权值矩阵每一行的所有元素之和等于 1，即 $\sum_{x_j \in N(x_i)} W_{ij} = 1$，最小化成本函数 $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$，求得最优权值 W_{ij}。

在上述两个限制条件下最小化重构误差而得到的最优权值遵循对称特性：对于具体的数据点，其本身和其邻近数据点在旋转、缩放、平移操作时都将保持其原有的性质不变。旋转和缩放不变性可以从式(3-5)得到，而平移不变性则出条件

②得到保证。

(3) 计算低维向量 y_i

高维观察值 x_i 被映射为低维向量 y_i，正好反映了真正的维数。d 维向量 y_i 可以通过最小化嵌入成本函数得到

$$\tilde{\varepsilon} = \sum_{i=1}^{n} \left\| y_i - \sum_{x_j \in N(x_i)} W_{ij} y_i \right\|^2 \tag{3-6}$$

该成本函数以局部线性重构误差为基础。式(3-6)中的嵌入成本函数是向量 y_i 的一个二次方的形式，为简化计算，可通过求解稀疏矩阵的特征向量求解其最小值。由于该算法只有一个参数，即每个数据点的近邻点数 k，一旦 k 值选定，利用线性代数相关知识就可以计算出最优权值和每个数据点对应的低维空间的数据点。

LLE 方法可以学习任意维的局部线性的低维流形。只有两个待定系数 k 和 d。同时由重构代价函数最小化得到的最优权值遵循对称特性，每个点的近邻权值在平移、旋转、伸缩变化下保持不变。LLE 方法有解析的全局最优解，不需要迭代，低维嵌入的计算归结为稀疏矩阵特征值的计算，这样计算复杂度相对较小。然而，LLE 方法要求所学习的流形只能是不闭合的且在局部是线性的，还要求样本在流形上是稠密采样的。另外，该方法的参数选择不确定，对样本中的噪声很敏感。

3.2 监督维数约减算法

在一些情况下，人们可以获得样本的标签，而在更多的情况下，人们往往不能明确得知某一样本的具体类别标签，所知的是某两个样本是否属于同一类别(类别标签未知)的成对约束信息，这种成对约束信息称为“边信息”。边信息包括两种，一种是正约束，表示两个样本属于同一类，另一种是负约束，表示两个样本不属于同一类别。边信息是一种比标记信息更一般的信息，因为边信息可以从标签信息中得到，反之则不行。一些监督维数约减技术就是充分利用这些边信息的。

3.2.1 线性判别分析 LDA

线性判别分析 LDA(Linear Discriminant Analysis)，也称为 Fisher 线性判别 FLD (Fisher Linear Discriminant)，也是模式识别中的一种经典的特征提取和降维方法，已经被广泛地应用于语音识别、计算机视觉、图像检索以及人脸识别等问题中。LDA 应用于特征提取时，它充分考虑了类别之间的信息，能使得样本在降维后的低维特征空间中，不同类样本尽可能远，同类样本尽可能近。这个准则下的分类

器比简单的线性分类器性能要强。

LDA 的具体算法步骤为：

(1) 初始的 m 个 s 维的特征向量构建矩阵 $X_{s\times m}$；

(2) 计算各样本集的均值和整个样本集的均值：

μ_i 表示样本集 i 的均值，μ 表示整个样本集的均值 $\mu=\sum_j p_j\times\mu_j$，其中 p_j 是类 j 的先验概率，为第 j 类样本数目的倒数。

(3) LDA 中，类内距离和类间距离作为分类的标准：

类内距离 S_w 表征了每个类内部的分布结构：$S_w=\sum_j p_j\times(\mathrm{cov}_j)$，其中 $\mathrm{cov}_j=\sum_k(x_{jk}-\mu_j)(x_{jk}-\mu_j)^T$；

类间距离的计算公式为：$S_b=\sum_j(\mu_j-\mu)(\mu_j-\mu)^T$；

LDA 基于类的投影中，各类别均有一个优化标准，各类的投影法则是：$criterion_j=inv(\mathrm{cov}_j)S_b$。而与类无关的投影中，只有一个投影法则：$criterion=inv(S_w)S_b$；

(4) 计算每类的 $criterion_j$ 以及与类无关的 $criterion$ 的特征值和特征向量：

所有特征值和特征向量按照特征值大小顺序排列 $(\lambda_1,e_1),(\lambda_2,e_2),\ldots,(\lambda_s,e_s)$，两种投影算法分别取前 t 个特征向量组成 $s\times t$ 投影矩阵 W_j 和 W。

(5) 初始特征向量经投影矩阵映射到目标子空间中：

基于类的投影中：$Y_j=(W_j)^T X_j$

与类无关的投影中：$Y=(W)^T X$

在很多的分类问题中，LDA 得到了很好的应用。然而，对于图像而言，它的维数可能会很大，所以在求取矩阵特征值的过程中计算量会很大，甚至矩阵是奇异的。为了解决上述问题，许多关于特征提取的研究都使用 PCA+LDA 的方法，即先用 PCA 进行降维，然后再应用 LDA。

3.2.2 有监督的局部线性嵌入 SLLE

LLE 算法并没有考虑到类别信息在分类中的作用，Dick 和 Robert 提出一种针

对有监督的 LLE 算法 SLLE[44](Supervised linear locally embedding)。传统的 LLE 算法在第一步时是根据样本点间的欧氏距离来寻找 k 个近邻点的，而 SLLE 在处理这一步时增加了样本点的类别信息，其余步骤同 LLE 算法一致。它在计算点与点之间的距离时有两种方法：

SLLE1：这种方法是采用式(3-7)来修正点与点之间的距离

$$D' = D + \alpha \max(D)\Delta \tag{3-7}$$

其中 D' 是计算后的距离，D 是最初采用的距离，$\max(D)$ 表示类与类之间的最大距离，Δ 取 0 或者 1，当两点属于同类时 Δ 取为 0，否则取 1，$\alpha \in [0,1]$ 是控制点集之间的距离参数，当 α 取 0 时 SLLE 和 LLE 算法相同。这个方法中 α 是一个经验参数，对实验的结果有很大的影响。

SLLE2：求解点与点之间的距离，目的在于寻找样本点的近邻点。SLLE2 的方法在寻找近邻点时，不是在全局样本中寻找近邻点，而是在每个点所在类的样本中寻找。也就是说，在类内中寻找样本点的近邻点。这个方法没有采用参数的方法，但是如果某一类的样本个数小于 k，那么这种方法将失败，所以，在每个类的样本个数相当多的情况下可以采用这种方法。

3.3 半监督维数约减算法

半监督维数约减算法是利用少量的约束信息和大量的未标记样本来进行降维了一类算法，本节将介绍几种半监督维数约减算法。

3.3.1 半监督典型相关分析 Semi-CCA

首先简要介绍一下典型相关分析 CCA(canonical correlation analysis)，它侧重于多模态的识别，即利用互补原理最大化不同模态数据之间的相关性，减少数据之间的不确定性，从而达到增强识别能力的目的，这里的多模态既可以是多种信息来源(如声音和图像)，也可以是从同一来源的信息中抽取的不同特征。

给定一批成对的观察样本集 $(x_i, y_i) \in R^p \times R^q$，$i = 1,\ldots,n$，其中 x_i 和 y_i 分别由不同信息渠道获得，记 $X = [x_1,\ldots,x_n] \in R^{p\times n}$ 和 $Y = [y_1,\ldots,y_n] \in R^{q\times n}$，记 (x, y) 为样本集中任一对样本，并设样本已经中心化，即 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 0$，，则 CCA 的目标是分别为样本集 X 和 Y 寻找两组基向量 $w_x \in R^p$ 和 $w_y \in R^q$，使得随机变量

$x = w_x^T x$ 和 $y = w_y^T y$ 之间的相关达到最大，即求如下相关系数的最大值问题：

$$\rho = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} \cdot w_y^T C_{yy} w_y}} \tag{3-8}$$

其中 $C_{xx} = E[xx^T] = XX^T \in R^{p\times q}$ 和 $C_{yy} = E[yy^T] = YY^T \in R^{p\times q}$ 表示集合内协方差矩阵，$C_{xy} = E[xy^T] = XY^T \in R^{p\times q}$ 表示集合间协方差矩阵，且有 $C_{xy} = E[yx^T] = C_{xy}^T$。

Semi-CCA 加入了样本间成对约束信息，设有 n 对样本 $(x_i, y_i) \in R^p \times R^q$，$i = 1,\ldots,n$，Semi-CCA 的目标是寻找一组投影向量 $w_x \in R^p$ 和 $w_y \in R^q$，使得抽取的同类样本特征之间的相关最大化，同时使得不同类样本特征之间的相关最小化。具体可表述为如下优化问题：

$$\rho = \frac{w_x^T \tilde{C}_{xy} w_y}{\sqrt{w_x^T C_{xx} \cdot w_y^T C_{yy} w_y}} \tag{3-9}$$

其中

$$\begin{aligned} \tilde{C}_{xy} &= XY^T + \sum_{(x_i,y_j)\in M} (x_i y_j^T + x_j y_i^T) - \sum_{(x_i,y_j)\in C} (x_i y_j^T + x_j y_i^T) \\ &= XEY^T + XMY^T - XCY^T \\ &= X(E + M - C)Y^T \\ &= XSY^T \end{aligned} \tag{3-10}$$

E 是单位矩阵，M 是表示所有正约束的一个集合，C 是表示所有负约束的一个集合。若将 M 和 C 设为矩阵，则 $M \in R^{p\times q}$，$C \in R^{p\times q}$，设初值时，M 和 C 都为零矩阵。由成对约束信息知，两个样本 x_i 和 y_j 属于同一类时，相应的 M_{ij} 和 x_{ji} 为 1；不属于同一类时，相应的 C_{ij} 和 x_{ji} 为 -1。

由于相关系数 ρ 与 w_x 和 w_y 的尺度无关，故 Semi-CCA 的求解问题可表述为如下优化形式：

$$\begin{aligned} &\max_{w_x, w_y} \ w_x^T \tilde{C}_{xy} w_y \\ &s.t. \ \ w_x^T XX^T w_x = 1 \\ &\qquad w_y^T YY^T w_y = 1 \end{aligned} \tag{3-11}$$

利用 Lagrange 乘子法求解此优化问题求得特征向量 w_x，w_y 后，对任一对样本 (x,y) 即可用如下方式进行特征组合：

$$w_x^T x + w_y^T y \tag{3-12}$$

$$\begin{bmatrix} w_x^T x \\ w_y^T y \end{bmatrix} \tag{3-13}$$

其中 $w_x = [w_{x1},\ldots,w_{xd}] \in R^{p\times q}$，$w_y = [w_{y1},\ldots,w_{yd}] \in R^{p\times qT}$，$d \le \min(p,q)$。基于式(3-12)和式(3-13)的特征组合方式分别简称为“并行组合”与“串行组合”方式利用组合的特征。

3.3.2 半监督维数约减 SSDR

给定样本集 $X = \{x_1, x_2, \ldots, x_n\} \subset R^D$ 以及正约束对集 M 和负约束对集 C，所期望的结果是从上述给定的条件中学习到线性变换矩阵 $W = [w_1, w_2, \ldots, w_d] \in R^{D\times d}$，使得原数据经由变换矩阵所得的低维投影为 $y_i = W^T x_i$，能够保持原始数据的结构以及 M 和 C 的约束特性。定义目标函数为最大化 $J(W)$：

$$\begin{aligned} J(W) &= \frac{\alpha}{2n_C}\sum_{(x_i,x_j)\in C}(y_i - y_j)^2 - \frac{\beta}{2n_M}\sum_{(x_i,x_j)\in M}(y_i - y_j)^2 \\ &= \frac{\alpha}{2n_C}\sum_{(x_i,x_j)\in C}(W^T x_i - W^T x_j)^2 - \frac{\beta}{2n_M}\sum_{(x_i,x_j)\in M}(W^T x_i - W^T x_j)^2 \end{aligned} \tag{3-14}$$

其中 $W^T W = E$，n_C 和 n_M 分别为约束对的量，α 和 β 均衡负约束和正约束两项的贡献。

式(3-14)的意义在于使得不属于同一类的数据离的越远越好，而属于同一类的数据靠的越近越好，但它仅考虑了约束对，下面引入剩余的未标记数据：

$$\begin{aligned} J(W) &= \frac{1}{2n^2}\sum_{i,j}(W^T x_i - W^T x_j)^2 + \frac{\alpha}{2n_C}\sum_{(x_i,x_j)\in C}(W^T x_i - W^T x_j)^2 \\ &\quad - \frac{\beta}{2n_M}\sum_{(x_i,x_j)\in M}(W^T x_i - W^T x_j)^2 \end{aligned} \tag{3-15}$$

式(3-15)的第一项是变换空间内所有的样本的均方距离，由于包含了大量的未标记样本，所以当约束对信息较少时能有效的强化性能。式(3-15)等价于：

$$J(W) = \frac{1}{2}\sum_{i,j}(W^T x_i - W^T x_j)^2 S_{ij} \tag{3-16}$$

$$\text{其中 } S_{ij}=\begin{cases}\dfrac{1}{n^2}+\dfrac{\alpha}{n_C} & \text{若}(x_i,x_j)\in C\\ \dfrac{1}{n^2}-\dfrac{\beta}{n_M} & \text{若}(x_i,x_j)\in M\\ \dfrac{1}{n^2} & \text{其他}\end{cases}。$$

由式(3-16)可得

$$\begin{aligned}J(W)&=\frac{1}{2}\sum_{i,j}(W^Tx_i-W^Tx_j)^2S_{ij}\\&=\frac{1}{2}\sum_{i,j}(W^Tx_ix_i^TW-W^Tx_jx_j^TW+2W^Tx_ix_j^TW)S_{ij}\\&=\sum_{i,j}W^Tx_iS_{ij}x_i^TW-\sum_{i,j}W^Tx_iS_{ij}x_j^TW\\&=\sum_{i}W^Tx_iD_{ii}x_i^TW-W^TXSX^TW\\&=W^TX(D-S)X^TW\\&=W^TXLX^TW\end{aligned}\tag{3-17}$$

这里 D 是对角矩阵，对角线元素 $D_{ii}=\sum_j S_{ij}$，$L=D-S$ 是拉普拉斯矩阵。

由上面的推导可将此问题表述为：

$$\begin{aligned}&\max J(W)=W^TXLX^TW\\&s.t.\quad W^TW=E\end{aligned}\tag{3-18}$$

很明显这是一个典型的特征值问题，其解可以通过求解矩阵 XLX^T 的特征值所对应的特征向量得到。易知，如果 $d>1$，那么取前 d 个最大非零广义特征值所对应的特征向量即可组成变换矩阵 W。求出变换矩阵 W 后,对于一个新来样本 x 可以直接得到其低维投影 y,即 $y=W^Tx$。

3.4 改进的半监督维数约减算法

3.4.1 ISSDR 算法介绍

在 SSDR 算法中，式（3-15）中引入剩余的未标记数据，保持了数据集的全局结构，但是它并没有更有效的利用未标记数据中的潜在信息，即没有利用数据集的局部信息。

为了能够利用大量未标记数据中的潜在信息，我们提出 ISSDR 算法。假设：

高维空间中互相靠近的点在低维投影空间中也是互相靠近的。基于此，我们可以利用最近邻图来表示这种邻接关系。具体来说，如果点 x_i 是离点 x_j 最近的 k 个点中的一个，那么就用一条边把节点 i 和节点 j 连接起来，这样形成的图就是 k-最近邻图，记为 G。有了上述定义,我们定义新的目标函数为最大化 $J(W)$：

$$\begin{aligned} J(W)=\frac{1}{2}(\alpha \sum_{x_i\notin N_k(x_j)orx_j\notin N_k(x_i)}(W^Tx_i-W^Tx_j)^2-\beta \sum_{x_i\in N_k(x_j)orx_j\in N_k(x_i)}(W^Tx_i-W^Tx_j)^2 \\ +\sum_{(x_i,x_j)\in C}(W^Tx_i-W^Tx_j)^2-\sum_{(x_i,x_j)\in M}(W^Tx_i-W^Tx_j)^2) \end{aligned} \tag{3-1}$$

其中 $W^TW=E$，α 和 β 分别用来调节邻近点和非邻近点的贡献度，均衡两项的贡献。

式(3-19)等价于：

$$J(W)=\frac{1}{2}(\sum_{i,j}(W^Tx_i-W^Tx_j)^2(S_{ij}+S_{ij}')) \tag{3-20}$$

其中 $S_{ij}=\begin{cases}1 & 若(x_i,x_j)\in C\\ -1 & 若(x_i,x_j)\in M\\ 0 & 其他\end{cases}$，$S_{ij}'=\begin{cases}\alpha & 若x_i\notin N_k(x_j)orx_j\notin N_k(x_i)\\ -\beta & 若x_i\in N_k(x_j)orx_j\in N_k(x_i)\\ 0 & 其他\end{cases}$。

由式(3-17)的推导类似的推导可得：

$$\begin{aligned} J(W)&=W^TX(D-S-S')X^TW\\ &=W^TXLX^TW \end{aligned} \tag{3-21}$$

这里 D 是对角矩阵，对角线元素 $D_{ii}=\sum_j(S_{ij}+S_{ij}')$，$L=D-S-S'$ 是拉普拉斯矩阵。此时式(3-21)和式(3-17)具有相同的形式，接下来可以通过求解矩阵 XLX^T 的特征值所对应的特征向量得到 W。

3.4.2 实验分析

为了验证 ISSDR 算法的有效性，通过两个实验将该算法与其他算法相比较，比较的指标是降维后的低维投影在分类问题上的效果好坏。假设在分类时知道所有训练样本的标记，所使用的分类方法为最近邻分类法(1-NN)。与之相比较的算法包括 PCA 维数约减后用最近邻分类，SSDR 维数约减后用最近邻分类，实验所用数据集是 UCI 数据集中的两组数据 Iris 和 Wine。在下面的实验中，边信息都是

通过从训练样本中随机选取样本点对来获取，如果某一样本点对的两个样本属于同一类，则把该点对放入正约束中，反之则放入负约束中。在实验过程中，假设只知道边信息而不知道训练样本的标记信息，参数α和β都设置为 0.05，实验结果是不同算法在不同约束对数量情况下的准确度，均为 200 次不同边信息情况下的平均值。

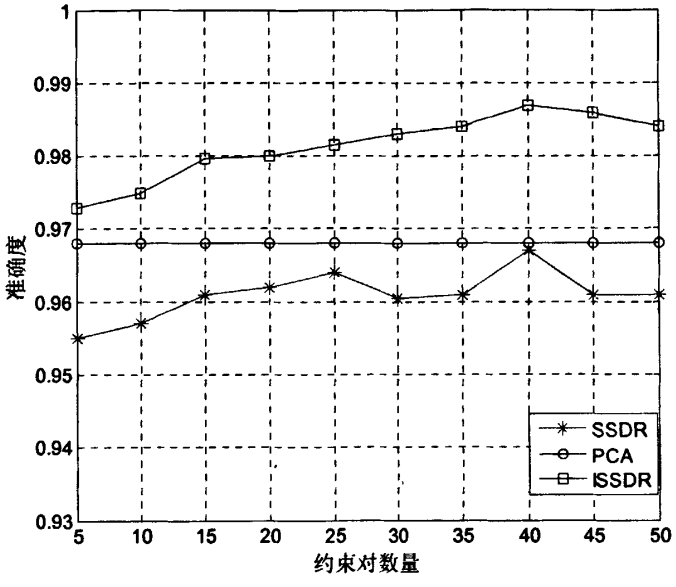

图 3.1 Iris 实验结果

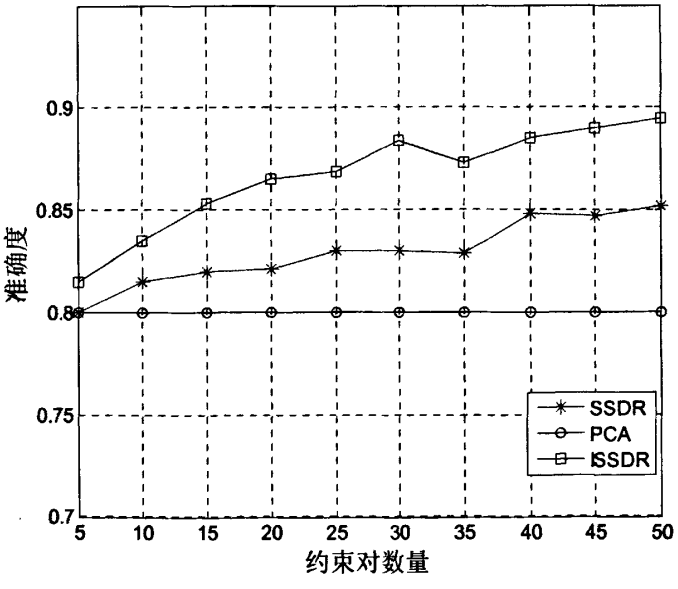

图 3.2 Wine 实验结果

Iris 数据集分为 3 类，每个样本 4 维，随机选择其中的 120 个样本作为训练集，剩下的 30 个样本作为测试集，约减后的数据维数为 2，邻近点个数为 3，实验结果如图 3.1 所示。Wine 数据集分为 3 类，每个样本 13 维，随机选择其中的 148 个样本作为训练集，剩余的 30 个样本作为测试集，约减后的数据维数为 4，邻近点个数为 5，实验结果如图 3.2 所示。

从图中可以看出，由于 PCA 算法并没有利用约束对信息，所以随着约束对数量的变化，PCA 算法的准确度保持不变。ISSDR 算法无论是在少量约束还是大量约束数量条件下，都能够得到比 PCA 算法以及 SSDR 算法更好的分类精度。ISSDR 算法比 PCA 算法更好是因为 ISSDR 利用了约束信息和先验知识即流形假设，而 PCA 则纯粹是一种无监督的算法。ISSDR 算法比 SSDR 算法更好是因为 ISSDR 在利用约束信息的同时保持了数据集的局部以及全局结构，而 SSDR 在利用约束信息的同时仅保持了数据集的全局结构，这说明局部结构的保持对算法性能的提升有着重要的作用。另外，SSDR 算法在 Iris 数据集上的性能是最差的，甚至不如 PCA 算法，这说明仅仅利用边信息和保持全局信息是不够的，可能导致降维效果不如无监督算法，这也从另一个侧面说明了保持数据集局部结构的重要性。

3.5 本章小结

本章首先分别从无监督、监督、半监督三个方面介绍了一些常见的维数约减算法，然后基于 SSDR 算法提出了一种新的半监督维数约减算法 ISSDR，并对其进行了详细描述。ISSDR 算法在利用正负约束对信息的同时，将大量的未标记数据对分为邻域内和邻域外两部分，从而既能够保持数据集的全局结构又能保持局部结构，实验证明了该算法的优越性。

第四章 基于集成学习的半监督学习算法

4.1 集成学习

4.1.1 集成学习的理论基础

集成学习是一种新的机器学习范式，它使用多个(通常是同质的)学习器来解决同一个问题，能够显著地提高学习系统的泛化能力。因此从20世纪90年代开始，对集成学习的理论和算法的研究成为了机器学习领域的一个热点。1988年Kearns和Valiant指出[45]，在PAC学习模型[46]中，若存在一个多项式级的学习算法来辨别一组概念，并且辨别正确率很高，那么这组概念是强可学习的；而如果学习算法辨别一组概念的正确率仅比随机猜测略好，那么这组概念是弱可学习的。Kearns和Valiant提出了弱学习算法与强学习算法的等价性问题，即是否可以将弱学习算法提升成强学习算法。如果两者等价，那么在学习概念时，只需找到一个比随机猜测略好的弱学习算法，就可以将其提升为强学习算法，而不必直接去寻找通常情况下很难获得的强学习算法。1990年，Schapire通过一个构造性方法对该问题作出了肯定的证明[47]，证明如果一个概念是弱可学习的，充要条件是它是强可学习的，即多个弱学习器可以集成为一个强学习器。集成机器学习的理论基础由此奠定。

狭义地说，集成学习是指利用多个同质的学习器来对同一个问题进行学习，这里的“同质”是指所使用的学习器属于同一种类型，例如所有的学习器都是决策树、都是神经网络等等。广义地来说，只要是使用多个学习器来解决问题，就是集成学习。在集成学习的早期研究中，狭义定义采用得比较多，而随着该领域的发展，越来越多的学者倾向于接受广义定义。例如，以神经网络集成为例，P.Sollich和A.Krogh在1996年给出的定义就采用了狭义定义[48]，而Opitz和Maclin在1999年给出的定义则采用了广义定义[49]。采用广义定义有一个很大的好处，就是以往存在的很多名称上不同、但本质上很接近的分支，例如多分类器系统、多专家混合、基于委员会的学习等，都统一地归属到集成学习之下。所以在广义的情况下，集成学习已经成为了一个包含内容相当多的、比较大的研究领域。

集成学习是通过训练多个学习器，将各自的输出结果进行某种方式的结合，来对新的实例进行分类或预测。在学习阶段，先由原始训练集T产生s个训练子集，由每一个训练子集$T_i(i=1,\ldots,s)$产生对应的学习器$h_i(i=1,\ldots,s)$。在实际应用时，个体学习器以某种方式结合在一起组成$h^* = F(h_1,\ldots,h_s)$。测试样本x由集成学习

系统h^*进行识别。可以看出，集成学习研究的重点集中在两个方面一是如何产生个体学习器，二是如何把产生的学习器进行适当的结合，以下两节将分别讨论这两方面的内容。

由于集成学习的基本思想是用多个模型解决方案来解决同一个问题，它的作用主要体现在以下四个方面：

(1) 提高预测结果的准确性

机器学习的一个重要目标就是对新的测试样本尽可能给出最精确的估计。构造单个高精度的学习器是一件相当困难的事情，然而产生若干个只比随机猜想略好的学习器却很容易。研究者们在应用研究中发现，将多个学习器进行集成后得到的预测精度明显高于单个学习器的精度，甚至比单个最好的学习器的精度更高。

(2) 提高预测结果的稳定性

有些学习算法单一的预测结果时好时坏，不具有稳定性，不能一直保持高精度的预测。就拿全球地震预测系统来说，一个模型对于某一个地区的地震可能预测得很好。但是，对于其他地区，可能别的模型会预测得比较好。通过模型的集成，可以在全球范围内以较高的概率普遍取得很好的结果。

(3) 解决过适应问题

在对已知的数据集合进行学习的时候我们常常选择适应度值最好的一个模型作为最后的结果。也许我们选择的模型能够很好的解释训练数据集合，但是却不能很好的解释测试数据或者其他数据，也就是说这个模型过于精细的刻画了训练数据，对于测试数据或者其他新的数据泛化能力不强，这种现象就称为过适应。为了解决过适应问题，按照集成学习的思想，可以选择多个模型作为结果，对于每个模型赋予相应的权重，从而集合生成合适的结果，提高预测精度。

(4) 改进参数选择

对于一些算法而言如神经网络遗传算法在解决实际问题的时候需要选择操作参数，但是这些操作参数的选取没有确定性的规则可以依据，只能凭借经验来选取，对于非专业的一般操作人员会有一定的难度，而且参数选择不同，结果会有很大的差异。通过建立多个不同操作参数的模型，可以解决选取参数的难题，同时将不同模型的结果按照一定的方式集成就可以生成我们想要的结果。

4.1.2 个体生成方法

产生个体学习器的方式有很多种，下面从三个方面进行介绍。

a. 基于训练集处理的方法

对训练集进行处理的方法是一种被广泛使用的个体生成方法，该方法让学习算法运行多次，每一次使用一种不同的训练集产生一个学习器。当然，各次使用

的训练集是从原始数据集构造出来的，构造方法常用的是随机抽取子集。值得注意的是，这种个体生成方法特别适用于不稳定的学习算法，学习算法不稳定是说算法产生的学习器受训练集影响很大，训练集微小的变化也会导致产生一个不同的学习器。决策树、神经网络、规则学习算法都是不稳定的，而线性回归、最近邻等算法则是稳定的。这种个体生成方法里有两个代表性的算法 Bagging[50]和 Boosting[47]。

(1) Bagging

Bagging 算法产生T个样本集$B_1,B_2,\ldots,B_T$，在每个样本集B_i上训练得到分类器C_i，最后通过分类器$C_1,C_2,\ldots,C_T$的相对多数投票集成得到分类器C^*。

在等概率的重复采样产生的 Bootstrap 样本集中，原始训练集中某些样本可能在新的训练集中出现多次，而另外一些样本则可能一次也不出现。假设有m个样本，当可重复采样时，一个样本至少被选中一次的概率为$1-(1-1/m)^m$，当m较大时，其值接近于$1-1/e=63.2\%$，也就是说每轮重复取样平均包含原训练样本集的63.2%样本[51]。Bagging 方法通过重新选取训练集增加了集成个体间的差异度，从而提高了泛化能力。稳定性是 Bagging 能否发挥作用的关键因素，Bagging 能提高不稳定学习算法的预测精度，而对稳定的学习算法效果不明显，有时甚至使预测精度降低。目前 Bagging 也有很多种变体，例如在扰动训练集时不进行重取样，而是对各样本的权加入零均值高斯噪音。

(2) Boosting

PAC 学习模型中提出了强学习算法与弱学习算法的等价性问题，该问题可看作是 Boosting 系列算法的出发点。1990 年，Schapire 通过一个构造性方法对该问题作出了肯定的证明[47]，其构造过程称为 Boosting，并由 Freund 进行了改进[53]，和 Bagging 方法不同，Boosting 方法的个体分类器是串行产生的，即第n个分类器的训练集依赖于前$n-1$个分类器在训练样本集上的性能。但是，Schapire 和 Freund 的算法在解决实际问题时有一个缺陷，即它们都要求事先知道弱学习算法学习正确率的下限，这在实际问题中很难做到。1997 年，Schapire 和 Freund 提出了 AdaBoost(Adaptive Boost)算法[54]，该算法的效率与 Freund 算法很接近，却可以非常容易地应用到实际问题中。因此，该算法已成为目前最流行的算法 Boosting。它维持训练样本集的一个权值分布，训练样本的初始权值均为 1，然后训练得到分类器，根据分类器对训练样本分类的正误以及本轮训练集上的加权错误率更新样本权值，使得被错分的样本权值增加，从而使下一轮的分类器训练时努力使分类错误的样本分类正确。最后，集成分类器通过分类器集合的加权投票得到，训练错误率低的个体分类器在最后投票中占较高的权重。

虽然 Boosting 方法能够增强集成算法的泛化能力，但是同时也有可能使集成过分偏向于某几个特别困难的样本。因此，该方法不太稳定，有时能起到很好的

作用，有时却没有效果，甚至会发生加入新的个体分类器，集成分类器准确率下降的情况。

b. 基于特征选择的方法

构造个体学习器的另一类方法是通过选择不同的特征子集实现的，不同的特征子集构成了不同的输入训练样本集。和上节基于训练样本处理的集成学习方法比较而言，这种基于特征选择的集成学习方法的研究及应用相对较少，这是由于研究者认为这种方法只适合特证间高度冗余的情况。文献[55]中采取保留那些和某类别相关度较大的特征的方法进行集成，该方法在 Sonar 数据上的实验表明，即使只去掉很少的特征也会导致单个分类器性能的下降从而使得集成分类器的性能降低。文献[56]中提出的 AB(Attribute Bagging)算法，它是基于特征选择的集成学习的代表性算法，该算法通过对属性进行随机扰动，可以得到较强的泛化能力。

c. 基于随机扰动的方法

随机扰动就是对学习算法的初始值、执行参数等可以设置与变动的地方加入随机性的改变，以此产生出不同的学习器结果。神经网络算法在使用完全相同的训练集的情况下，改变网络的初始权值就能训练出不同的 BP 神经网络。文献[57]中对 C4.5 的节点属性测试采用了等概率随机扰动选择，把它和单个 C4.5 与进行了比较，结果表明 C4.5 集成随机扰动的集成取得了最好性能。另外，也有学者把 MCMC(Markov Chain Monte Carlo)方法应用到神经网络和决策树，以达到同样的随机扰动效果[58]。

4.1.3 结论生成方法

当我们已获得个体学习器时，怎样结合个体学习器的输出就是集成系统结论生成方法需要解决的问题。当集成学习器用于分类时，集成的输出通常由个体学习器的输出投票产生。一种方法是采用绝对多数投票法，即某分类成为最终结果当且仅当有超过半数的学习器输出结果为该分类，另一种方法是相对多数投票法，即某分类成为最终结果当且仅当输出结果为该分类的学习器的数目最多。

1990 年，Hansen 和 Salamon 证明[59]，对神经网络分类器来说，采用集成方法能够有效提高系统的泛化能力。假设集成由 N 个独立的分类器构成，采用绝对多数投票法，再假设每个分类器以 $1-p$ 的概率给出正确的分类结果，并且分类器之间错误不相关，则该集成学习发生错误的概率为：

$$p_e = \sum_{k>N/2}^{N} \binom{N}{K} p^k (1-p)^{N-k} \tag{4-1}$$

在 $p<1/2$ 时，p_e 随 N 的增大而单调递减。因此，如果每个分类器的预测精度都高于 50%，并且各分类器之间错误不相关，则分类器集成中的分类器数目越多，集

成的精度就越高。当N趋向于无穷，集成的错误率趋向于0。在采用相对多数投票法时，分类器集成的错误率比式(4-1)复杂得多，但是Hansen和Salamon的分析表明，采用相对多数投票法在多数情况下能够得到比绝对多数投票法更好的结果。因此，在对分类器进行集成时，目前大多采用相对多数投票法。

在实际应用中，由于各个独立的分类器并不能保证错误不相关，因此，分类器集成的效果与理想值相比有一定的差距，但其提高泛化能力的作用仍相当明显。

4.2 线性邻域传播算法 LNP

4.2.1 LNP 算法描述

线性邻域传播算法 LNP[38]是一种基于图正则化的半监督学习算法，实质上是借用了无监督降维算法 LLE 的思想，再加上少量的数据类别标记信息作为约束条件来提高算法性能。它基于如下假设：任意点的类标可以由其邻近点的类标线性重构(线性表示)。

设 $X=\{x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_m\}\subset R^d$ 是 m 个样本的集合，每个样本的维数是 d，$L=\{L_1,L_2,\cdots,L_c\}$ 为样本点的 c 种类别。记 $(x_1,y_1),\ldots,(x_l,y_l)$ 为给定的 l 个已标记样本，$y_i\in L$ 为 x_i 的标记，$(x_{l+1},\ldots,x_{l+u})$ 为 u 个未标记样本，其中 $m=l+u$，一般情况下 $l\ll u$。

它的实现过程可分为下面两步：

(1) 求解每个数据点的邻域结构(即权值)

$$\begin{aligned}\min\quad & \varepsilon_i=\min_{W_i}\left\|x_i-\sum_{x_j\in N(x_i)}W_{ij}x_j\right\|^2\\ s.t.\quad & W_{ij}=0,\forall x_j\notin N(x_i)\\ & \sum_{x_j\in N(x_i)}W_{ij}=1\end{aligned}\tag{4-2}$$

使用某一测度求得其 k (邻域尺度参数)个最近邻，并将其所组成的子集用 $N(x_i)$ 来表示。而重构权值 W_{ij} 表示点 x_j 在线性重构点 x_i 时所作的贡献，当 W_{ij} 值越大时，表示两个点的相似性越大，反之则越小。如式(4-2)所示，当 $W_{ij}=0$ 时，表示点 x_j 不在点 x_i 的 k 邻域范围内。

很容易推导得出：

$$\begin{aligned}\varepsilon_i &= \left\| x_i - \sum_{j:x_j \in N(x_i)} W_{ij} x_j \right\|^2 \\ &= \left\| \sum_{j:x_j \in N(x_i)} W_{ij}(x_i - x_j) \right\|^2 \\ &= \sum_{j,k:x_j,x_k \in N(x_i)} W_{ij} W_{ik} (x_i - x_j)^T (x_i - x_k) \\ &= \sum_{j,k:x_j,x_k \in N(x_i)} W_{ij} G^i_{jk} W_{ik}\end{aligned} \tag{4-3}$$

其中 G^i_{jk} 是大小为 $K \times K$ 的 Gram 矩阵，G^i 的元素 $G^i_{jk} = (x_i - x_j)^T (x_i - x_k)$，这样，每个样本点的权值可通过求解以下标准二次规划问题而得：

$$\begin{aligned} \min_{W_i} \quad & \frac{1}{2} W_i^T G^i W_i \\ s.t. \quad & W_i^T e = 1, \\ & W_i \geq 0 \end{aligned} \tag{4-4}$$

这里 $W_i = (W_{i1}, W_{i2}, \cdots W_{iK})$，$e$ 为 $K \times 1$ 的全一矩阵。

求解二次规划问题的方法很多，典型的有 Lagrange 方法、积极集方法、跟踪路径法以及 Lemke 方法等[60]。其中 Lemke 方法实现简单，通过求解线性互补问题解二次规划，它由一个准互补基本可行解出发，通过转轴方法（即主元消去）求出一个新的准互补基本可行解，直至得到满足互补条件的解。它对中小规模凸二次规划效果较好，本文将采用 Lemke 方法求解二次规划(4-4)。

(2) 使用已构造的图将已标记样本点的标记向未标记样本点传播

迭代过程中，使每个样本点的标记综合邻点信息（第一项）以及初始信息（第二项）对样本进行预测：

$$F^{t+1} = \alpha W F^t + (1-\alpha) Y \tag{4-5}$$

Y 是初始的标记矩阵，若 x_i 为已标记点且其标记为 $L_j \in L$，那么 $Y_{ij} = 1$，否则 $Y_{ij} = 0$；若 x_i 为未标记点，则 $Y_{ij} = 0$（$1 \leq j \leq c$），F^t 是第 t 次迭代的结果，初始化时 $F^0 = Y$。$0 < \alpha < 1$ 为样本点 x_i 从邻域结构中得到信息的比例。

可以证明数列 $\{F^t\}$ 收敛且收敛到

$$F^* = (1-\alpha)(I - \alpha W)^{-1} Y \tag{4-6}$$

不失一般性，这里假设样本集 X 中前 $l(l < n)$ 个样本是标记样本，并假设其类标为 $y_i \in L(1 \leq i \leq l)$，其中 $L = \{1, -1\}$ 表示两类问题的类标集合。f_i 为预测函数，而

最后未标记样本的类标为 $sign(f_i)(l+1 \le i \le N)$，其中，$sign(\cdot)$ 是符号函数。如图 4.1 所示，它能将有限个点(左图红色三角形、蓝色矩形所标注)的类标传播到附近未标记的点直到整个点集都标记完，这也形象解释了线性邻域传播这个名字的由来。

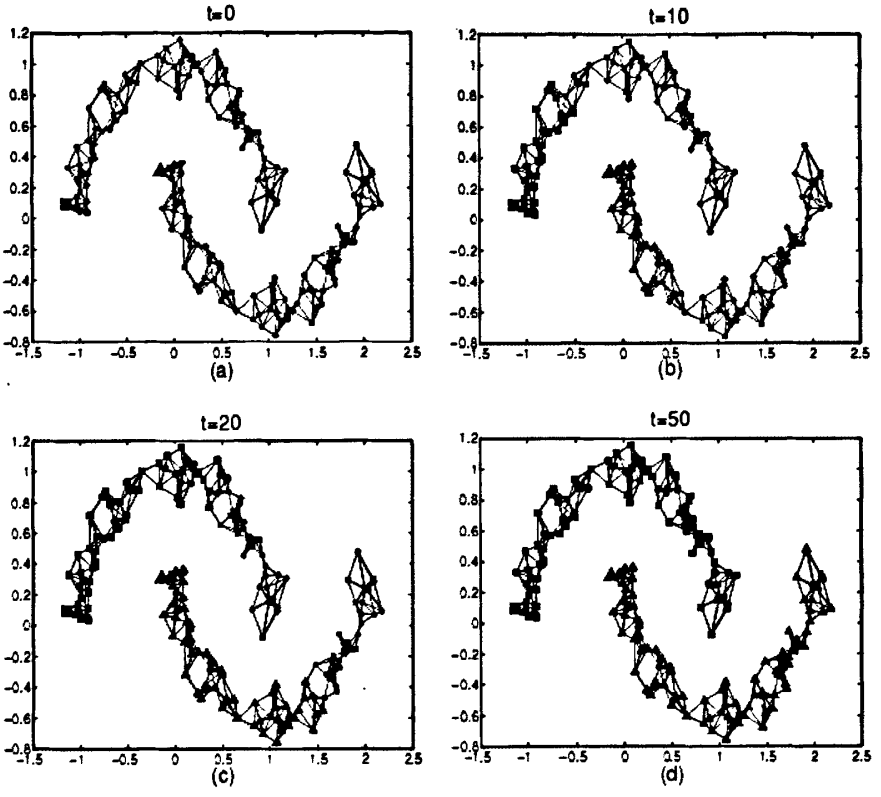

图 4.1 LNP 示意图（two-moons 数据库）

4.2.2 LNP 算法分析

LNP 算法具有 3 个很好的优点：

(a) 很好的实用性，因为它仅有一个邻域尺度参数 k 需要调整；

(b) 由 LNP 得到的类标预测函数 $f(\cdot)$ 对标记样本和未标记样本都足够光滑；

(c) 它可以很容易推广到用来标记样本外新来的数据。当新来一个样本 x_t 时，可以通过下面两个步骤方便地求得其类标：

(1) 在样本集 X 中寻找 x_t 的 k 近邻；

(2) 最小化成本函数

$$\min_f \tilde{\eta} = \min \sum_{i=1}^{N} \left\| f_i - \sum_{j:x_j \in N(x_i)} W_{ij} f_j \right\|^2 \quad (4\text{-}7)$$
$$s.t.\ f_i = y_i (1 \le i \le l)$$

从而得到新样本的类标预测函数为

$$f_t = \sum_{j:x_j \in N(x_t)} W_{tj} f_j \quad (4\text{-}8)$$

4.3 集成算法在半监督学习中的应用

4.3.1 算法描述

LNP 算法所需的存储空间比一般的基于图的算法要小很多，并且对于参数的选择具有很好的稳定性，分类精度高。为了进一步提高半监督学习的分类精度，我们将 LNP 作为子分类器，提出基于集成的半监督学习算法 E-LNP。

对于高维数据，半监督学习有可能受到数据冗余特征的影响，我们将对其进行实验说明。冗余特征的处理方法有很多种，用主成份分析 PCA 进行数据变换是一种经典的方法，它是将分散在一组变量上的信息集中到某几个综合指标(主成份)上的探索性的统计分析方法。由于它把原来众多具有一定相关性的指标重新组合成一组新的相互无关的综合指标来代替原来指标，用 PCA 对样本集进行处理后，其特征根据信息量大小排序，信息量大的特征排在样本集的前面部分，从而具有重要的优良性质。

算法 1.E1-LNP 算法

首先对样本集进行 PCA 维数约减，在经过 PCA 变换的基础上保留前一部分特征，然后从剩余的特征中随机选择一部分特征共同构成新的特征集，每次对确定的样本集用 LNP 算法进行学习，产生 t 个分类器。这样，保证了每次学习既能充分利用有用信息，又能使不同的分类器具有很好的差异性。对于每个样本点，t 个分类器分别产生 t 个预测值，用相对多数投票法得到最终的预测结果。

输入：每一行代表一个样本的样本矩阵 X；$L=\{L_1, L_2, \cdots, L_c\}$ 为样本点的 c 个类别；初始的标记矩阵 Y；固定的最近邻点值 $\bar{K}$；特征选择 PCA 保留特征的个数 n，保持不变的特征个数 n_1，在剩余特征中选择的特征个数 n_2；所要集成的子分类器个数 t；常量 α。

下面是算法的具体步骤：

(1) 对样本矩阵进行 PCA 维数约减，保留前 n 个特征，得到新的样本矩阵 X'；

(2) 对 t 个子分类器分别进行：

①保持 X' 中每个样本的前 n_1 个特征不变，再从剩余的特征中随机选择 n_2 个，

构成 X''；

②求解由 X'' 和 $\bar{K}$ 所确定的二次规划问题(4-4)，得到权矩阵 W；

③由迭代公式(4-6)得到收敛点，用其对样本点进行预测；

(3) 对 t 个子分类器的预测值分别进行投票，将出现最多的那个预测值作为样本 x_i 的最终预测标记；

(4) 输出所有样本的最终预测结果。

算法 2. E2-LNP 算法

首先对样本集进行 PCA 维数约减构成新的样本集，然后选择不同的学习参数，即给定一个包含不同的最近邻点数的序列集合 K，子分类器每次从中随机选择不同的邻点数进行学习。t 个分类器的分别随机从 K 中选择一个最近邻点值 K_i，与之一起来确定权矩阵 W。对于每个样本点，t 个分类器分别产生 t 个预测值，用相对多数投票法得到最终的预测结果。

输入：每一行代表一个样本的样本矩阵 X；$L=\{L_1,L_2,\cdots,L_c\}$ 为样本点的 c 个类别；初始的标记矩阵 Y；最近邻点数的序列集合 K；特征选择 PCA 保留特征的个数 n；所要集成的子分类器个数 t；常量 α。

下面是算法的步骤：

(1) 对样本矩阵进行 PCA 维数约减，保留前 n 个特征，得到新的样本矩阵 X'；

(2) 对 t 个子分类器分别进行：

①随机从参数序列集 K 中选择一个最近邻点值 K_i；

②求解由 X' 和 K_i 所确定的二次规划问题(4-4)，得到权矩阵 W；

③由迭代公式(4-6)得到收敛点，用其对样本点进行预测；

(3) 对 t 个子分类器的预测值分别进行投票，将出现最多的那个预测值作为样本 x_i 的最终预测标记；

(4) 输出所有样本的最终预测结果。

算法 3.E-LNP 算法

算法 1 和算法 2 是从两个不同的角度来增加分类器的差异性，我们对其进行综合从而得到了 E-LNP 算法。此算法中，我们首先对样本集进行 PCA 维数约减，然后对新的特征集分别做 t 次随机处理并选择不同的学习参数，每次对确定的样本集和学习参数用 LNP 算法进行学习，产生 t 个分类器。对于每个样本点，t 个分类器分别产生 t 个预测值，用相对多数投票法得到最终的预测结果。

输入：每一行代表一个样本的样本矩阵 X；$L=\{L_1,L_2,\cdots,L_c\}$ 为样本点的 c 个类别；初始的标记矩阵 Y；最近邻点数的序列集合 K；特征选择 PCA 保留特征的个数 n，保持不变的特征个数 n_1，在剩余特征中选择的特征个数 n_2；所要集成的

子分类器个数t；常量α。

下面是算法的步骤：

(1) 对样本矩阵进行PCA维数约减，保留前n个特征，得到新的样本矩阵X'；

(2) 对t个子分类器分别进行：

①保持X'中每个样本的前n_1个特征不变，再从剩余的特征中随机选择n_2个，构成X''；

②随机从参数序列集K中选择一个最近邻点值K_i；

③求解由X''和K_i所确定的二次规划问题(4-4)，得到权矩阵W；

④由迭代公式(4-6)得到收敛点，用其对样本点进行预测；

(3) 对t个子分类器的预测值分别进行投票，将出现最多的那个预测值作为样本x_i的最终预测标记；

(4) 输出所有样本的最终预测结果。

由于以上三种学习算法均采用了集成投票的方法，所以有可能减少分类误差，获得对问题空间模型更加准确的表示，从而提高分类器的分类精度和稳定性。

4.3.2 实验分析

实验使用美国国家邮政局数据库USPS(US Postal Service Database，USPS)收集的手写体数字，如图4.2所示。该数据库中包括10类数据共9298个样本，分别是数字0～9，如图所示。每个样本经过简单的预处理并归一化为16×16(=256)像素的灰度图，即包含256个输入特征。我们从每类数据中选择300个，总共3000个样本进行实验。

图4.2 USPS手写数字部分样本

实验中，每次从各类数据中随机选取k（$k=1\sim14$）个样本作为已标记样本，剩余样本作为未标记样本，重复10次。基本算法LNP中选择最近邻点$K=5$，常量$\alpha=0.99$。

表 4.1 PCA 特征约减结果

k	无特征约减	保留 200 个特征	保留 100 个特征	保留 50 个特征
1	58.74%	58.76%	58.99%	58.69%
2	69.96%	70.00%	69.98%	68.76%
3	74.60%	74.61%	74.83%	73.54%
4	78.09%	78.12%	78.17%	77.47%
5	79.78%	79.82%	79.95%	78.91%
6	80.86%	80.88%	80.89%	79.94%
7	81.87%	81.88%	81.87%	81.22%
8	82.44%	82.45%	82.51%	81.85%
9	82.44%	82.45%	82.50%	81.68%
10	83.49%	83.51%	83.59%	82.99%
11	84.05%	84.07%	84.06%	83.34%
12	84.43%	84.44%	84.53%	83.93%
13	84.77%	84.78%	84.78%	84.10%
14	85.57%	85.58%	85.59%	84.92%

实验一：

为了说明冗余特征对分类结果的影响，我们用PCA对特征集进行处理，保留特征个数分别为200，100，50，然后用LNP算法进行学习，实验进行100次求得平均值，结果如表4.1所示，表中无特征约减即是保持原来的256个特征不变。

从表 4.1 中可以看出，用 PCA 变换对样本集进行处理后，当保留特征个数为 200 和 100 时，分类精度都有所提高，说明了冗余特征对算法 LNP 有一定的影响。但当特征个数为 50 时，分类精度有所降低，这是因为此时所保留的特征已经不足以描述样本点的属性了。

实验二：

我们将最近邻点算法 1-NN 作为基本的对比算法，对 E-LNP 算法，E1-LNP 算法，E2-LNP 算法，以及 LNP 算法进行实验。其中 PCA 特征选择要保留特征的个数为 $n=100$，E-LNP 和 E2-LNP 中最近邻点数序列集合 $K=\{5,6,7,8\}$，E1-LNP 选择最近邻点 $K=5$，通过大量实验 E-LNP 和 E1-LNP 中保持不变的特征个数 $n_1=25$，在剩余特征中选择特征个数 $n_2=25$，三种集成算法子分类器个数均为 $t=100$，常量 $\alpha=0.99$，实验结果如图 4.3 所示。

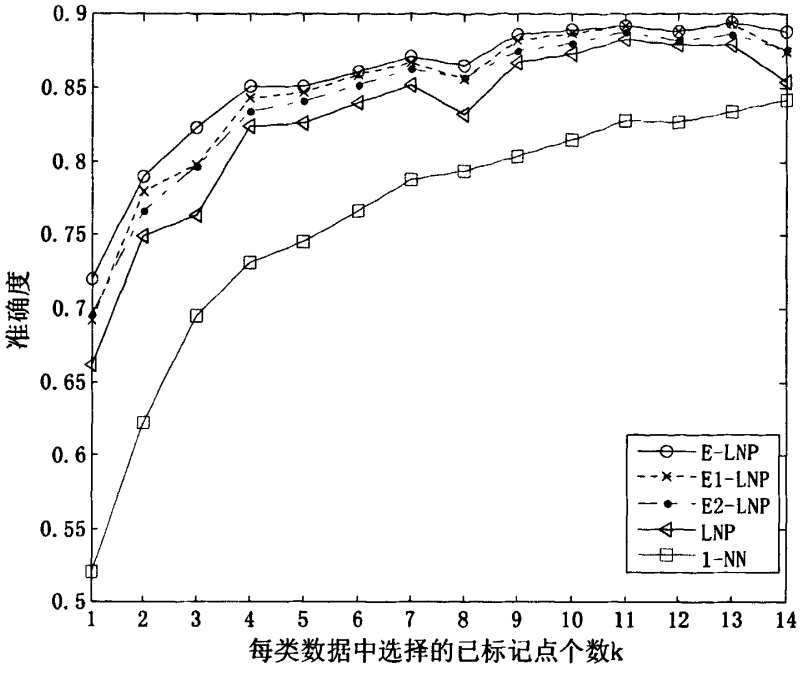

图 4.3 USPS 实验结果

由图 4.3 可以看出，子算法 LNP 明显优于 1-NN 算法。实验结果中，E-LNP 算法准确度最高，且当每类数据中给定的已标记点个数比较少时，E-LNP 算法相对于 LNP 算法的准确度提高的最明显，最多能提高 5.83%。E1-LNP 和 E2-LNP 两种算法所得的实验准确度也均比 LNP 算法的结果高，这说明了分别从两个不同的角度对算法进行集成是有效的，但两种算法均比 E-LNP 算法稍差一些，这说明子分类器的正确率和差异性越高时，集成的效果越明显。

4.4 本章小结

本章首先介绍了集成学习的知识，它将多个不同的单个模型组合成一个模型，其目的是利用这些单个模型之间的差异，来改善模型的泛化性能。提出了一种新的集成算法 E-LNP，它选用一种基于图的半监督学习算法 LNP 作为子分类器，每个子分类器分别选择不同的特征个数以及学习参数，进行多次训练预测，然后将其预测结果按相对多数投票方式进行集成，得到最终的学习结果。实验结果表明，该集成多个分类器的算法比仅使用一个分类器的分类精度有明显提高。

结束语

在机器学习中，监督学习需要大量带标记样本组成训练集以保证泛化能力。但在文件处理、生物信息学和网页分类等实际应用中，对数据进行人工标记的代价很高，容易获取的是大量无标记数据。无监督学习属于无任何监督信息的自动学习，虽然不需要对学习样本做类别标记，但是在不提供监督信息的情况下学习得到的模型是不够精确的。

半监督学习是介于两者之间的学习方式，即学习样本既包括已标记类别的样本也包括未标记类别的样本，在已标记类别的监督信息的指导下学习其他未标记样本。半监督学习方式事实上是假设了同类别的未标记样本与已标记样本在特征空间上的“距离”比较近。由于只需提供少量的标记样本，而通过全部样本学习又可以获得相对于非监督学习更好的学习效果，因此，将少量带标记数据和大量无标记数据结合的半监督学习得到了广泛的研究和应用。

本文主要对半监督学习中的降维和分类进行了研究。通过分析一些现有的较典型的降维方法和半监督分类方法，提出了新的半监督算法，本文的主要研究成果如下：

1. 基于半监督维数约减算法 SSDR 提出了一种改进的的算法 ISSDR。它不仅能够利用给出的正负约束对信息，使得属于负约束对集中的数据在低维空间中尽量远离，而属于正约束对集的数据尽量靠近，而且还引入剩余的未标记数据，基于流形假设充分挖掘出隐藏在未标记数据中的潜在信息，保持了数据集的局部以及全局结构。用最近邻分类 NN 在 Iris 数据集和 Wine 数据集上验证该算法的性能，实验证明了该算法的优越性。

2. 提出了一种集成的半监督分类算法 E-LNP，它选用线性邻域传播算法 LNP 作为基分类器。首先选择不同的样本特征以及学习参数，利用 LNP 算法训练出多个半监督分类器，然后将得到的预测结果按相对多数投票方式进行集成，从而确定最终的分类结果。用 USPS 数据集中的十类数据对该算法进行验证，表明 E-LNP 算法可以有效的提高半监督学习的分类精度，比仅使用单一的半监督分类器有更强的泛化能力。

由于半监督分类算法 E-LNP 实验的运行比较耗时，所以下一步考虑选用其他的半监督学习算法作为基分类器，在保证分类性能的情况下减少运行时间。在本文提出的两种算法中，可以考虑进一步选用更多的数据库来进行测试，从而更好的验证算法的有效性和优越性。

致谢

两年半的研究生涯转眼就过去了，回想两年多的日子，在盘点自己的得与失，憧憬未来的日子之时，深感过去日子里尽管有着自己艰辛的努力，同样有着指导老师的耐心指导和言传身教，有着实验室同学的帮助与鼓励，有着父母的关心和默默付出。在研究生生活结束之际，我想对他们道出真诚的谢意。

感谢我的指导老师周水生副教授。在这两年的研究生生涯里，他给我很多关心和精心的指导。周老师渊博的知识、严谨的治学风范、永不松懈的治学精神都给我留下了深刻的印象。我觉得自己很幸运有这样的指导老师，从他那里收获的东西对于我日后学习、研究、工作来说，都是非常珍贵的。在此，向周老师表示我诚挚的谢意和深深的敬意！

感谢舍友姜文丽、王雪岩和张春丽给我生活上的关照以及学习上的帮助！感谢同导师的王文丽，刘明，吴慧，以及实验室的师妹们，这段日子与我一起经历了求学路上的所有酸甜苦辣！感谢两年多来与她们一起度过了许多难忘的时光。

另外，我特别要感谢我的父母和亲人，没有他们不遗余力的强力支持，我不可能那么顺利的完成我的学业。是他们多年来一贯的支持和关怀，使我在学业上可以没有后顾之忧。他们是我坚强的精神后盾，正是他们的理解、信任、鼓励和宽容才使我能够坦然面对生活和学习中所遇到的各种困难和挑战，不断进取，并最终如期完成硕士研究生阶段的学业。

总之，我要感谢所有给过我帮助和关爱的人，祝他们永远平安、健康、幸福！

参考文献

[1] Huang L. A surver on web information retrieval technologies. Technical Report ECSL-TR-120,2000.

[2] Deerwester S, Dumails S, George W, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990, 41:391-407.

[3] 李晓黎, 刘继敏, 史忠植. 概念推理网及其在文本分类中的应用.计算机研究与发展.2000, 37(9):1032-1038.

[4] Nigam K, McCallum A, Thrun S, et al. Learning to classify text from labeled and unlabeled documents. Proceedings of 15the National Conference on Artificial Intelligence. 1998: 792-799.

[5] Tony J. Discriminative, Generative and Imitative Learning. PhD thesis, Massachusetts Inst. of Technology Media laboratory, Dec 2001.

[6] Schafer J. L. Analysis of Incomplete Multivariate Data. Chapman & Hall, 1997.

[7] McCallum A, Rosenfeld R, Mitchell T, et al. Improving text classification by shrinkage in a hierarchy of classes. In Proc. 15th Intl. Conf. on Machine Learning (ICML)[ICML98]. 1998: 359~367.

[8] Bennett K. P, Demiriz A. Semi-supervised support vector machines, Advances in Neural Information Processing Systems, Cambridge, MA, 1998,10: 368~374.

[9] Aharon B, Tomer H, Noam S, et al. Learning distance functions using equivalence relations. In Proc. of 20th International Conference on Machine Learning, 2003: 11~18.

[10] Kamal N, Andrew M, Sebastian T, et al. Text Classification from Labeled and Unlabeled Documents using EM, Machine Learning, 2000, 39:103~134.

[11] Chapelle O, Weston J, Scholkopf B. Cluster Kernels for Semi-Supervised Learning. Advances in Neural Information Processing Systems 15.

[12] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. Proceedings of the 20th International Conference on Machine Learning (ICML'03), Washington, DC. 2003: 912-919.

[13] Blum A, Chawla S. Learning from Labeled and Unlabeled Data using Graph Mincuts. In Proceedings of ICML 2001.

[14] Belkin M, Niyogi P. Using Manifold Structure for Partially Labeled Classification. Advances in Neural Information Processing Systems 15 (NIPS 2002), 2002.

[15] Zhu Xiaojin. Semi-Supervised Learning Literature Survey. Tech-nich report 1530,

Department of computer sciences. University of Wisconsin at Madison, 2008.

[16] Eric P, Andrew Y, Michael I, et al. Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems. The MIT Press.

[17] d'Alche F, Grandvalet Y, Ambroise C. Semi-supervised marginboost. Advances in Neural Information Processing Systems 14. MIT Press, 2002.

[18] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998:92~100.

[19] Fabio C, Ira C. Marcelo C. Semi-Supervised Learning of Mixture Models. In Proc. Of ICML2003.

[20] Nigam K, Ghan R. I. Analyzing the effectiveness and applicability of Co-training. Proceedings of Information and Knowledge Management, 2000: 86~93.

[21] Ion M, Steven M, Craig A. Active + Semi-Supervised Learning =Robust Multi-View Learning, ICML2002.

[22] Martin S, Tommi J, Tomaso P. Learning from Partially Labeled Data, Artificial Intelligence Laboratory and the Center for Biological and Computational Learning. Massachusetts Institute of Technology Cambridge, Massachusetts 02139, http://www.ai.mit.edu.

[23] Shahshahani B, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[24] Miller D. J, Uyar H. S. A mixture of experts classifier with learning based on both labeled and unlabelled data. Advances in Neural Information Processing Systems 9, Cambridge, MA: MIT Press. 1997: 571-577.

[25] Zhang T, Oles F. J. A probability analysis on the value of unlabeled data for classification problems. In: Proceedings of the 17th International Conference on Machine Learning (ICML'00), San Francisco, CA, 2000: 1191-1198.

[26] Joachims T. Transductive inference for text classification using support vector machines. Proceedings of the 16th International Conference on Machine Learning (ICML'99), Bled, Slovenia, 1999:, 200-209.

[27] Lawrence N. D, Jordan M. I. Semi-supervised learning via Gaussian processes. Advances in Neural Information Processing Systems 17, Cambridge, MA: MIT Press. 2005: 753-760.

[28] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. Advances

in Neural Information Processing Systems 17, Cambridge, MA: MIT Press, 2005: 529-536.

[29] Zhou D, Bousquet O, Lal T, et al. Learning with local and global consistency. Advances in Neural Information Processing Systems 16, Cambridge, MA: MIT Press, 2004: 321-328.

[30] Glenn F, Mangasarian O.L. Semi-Supervised Support Vector Machines for Unlabeled Data Classification. Optimization Methods and Software, 2001, 15:29~44.

[31] Chapelle O, Weston J, Sch(o)lkopf B. Cluster Kernels for Semi-Supervised Learning. Advances in Neural Information Processing Systems 15.

[32] Seeger M. Covariance kernels from Bayesian generative models. Advances in Neural Information Processing Systems, 2001.

[33] Jaakkola T, Haussler D. Exploiting generative models in discriminative classifiers. Advances in Neural Information Processing. The MIT Press, 1998, 11:487~493.

[34] Kiri W, Claire C, Seth R, et al. Constrained K-means Clustering with Background Knowledge. ICML-2001: 577~584.

[35] Sugato B, Arindam B, Raymond J. Mooney: Semi-supervised Clustering by Seeding. ICML 2002:19~26.

[36] Dempster A P, Laid N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc, Ser B, 1977, 39: 1~38

[37] Zhang R, Rudnicky A. A new data selection approach for semi-supervised acoustic modeling. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2006, 1.

[38] Wang F, Wang J, Zhang C, et al. Semi-Supervised classification using linear neighborhood propagation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2006, 1: 160-167.

[39] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. Proceedings of 18th International Conference on Machine Learning, 2001: 19-26.

[40] Zhu X, Ghahramani Z. Towards semi-supervised classification with Markov random fields. Technical Report CMU-CALD-02-106. Carnegie Mellon University, 2002.

[41] Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs. Conference on Learning Theory. 2004, 3120: 624-638.

[42] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from examples. Technical Report TR-2004-06. University of Chicago, 2004.

[43] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习算法[J].软件学报,2003,14(3):451~460.

[44] Dick de R Olga K, Oleg O, et al. Supervised locally linear embedding. Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP 2003 Proceedings, Lecture Notes in Computer Science 2714, Springer: 333-341.

[45] Kearns M, Valiant L. Learning Boolean formulae or factoring. Aiken Computation Laboratory. Harvard University, Cambridge, MA, Technical Report: TR-1488, 1988.

[46] Kearns M, Valiant L. Crytographic limitation on learning Boolean formulae and finite antomata. Proceedings of the 21St Annual ACM Sysposium on Theory of Computing, New York, NY: ACM press, 1989: 433-444.

[47] Schapire R. The strength of weak learn ability [J]. Machine Learning. 1990, 5(2): 197-227.

[48] Sollich P, Krogh A. Learning with ensembles: how over-fitting can be useful. Advances in Neural Information Processing Systems 8, Cambridge, MA: MIT Press, 1996:190-196.

[49] Opitz D, Maclin R. Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research, 1999, 11:169-198.

[50] Breiman L.Bagging predictors. Machine Learning. 1996,24(2):123-140.

[51] Dietterich T. Machine learning research: four current directions. AI Magazine. 1997, 18(4):97-136.

[52] Breiman L.Bagging predictors. Machine Learning. 1996,24(2):123-140.

[53] Freund Y. Boosting a weak algorithm by majority. Information and Computation. 1995, 121(2):256-285.

[54] Freund Y, Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997, 55(1):119-139.

[55] Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connection Science, 1996, 8(3-4):385-404.

[56] Bryll R, Gutierrez-Osuna R, Quek F. Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition. 2003, 36(6):1291-1302.

[57] Dietterch T.G, Kong E.B. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Department of Computer Science, Oregon State University, Corvallis, Oregon.1995.

[58] Mackay D. A practical Bayesian framework for backpropagation networks. Neural Computation, 1992. 4(3):448-472.

[59] Hansen L, Salamon P. Neural network ensembles. IEEE Trans Pattern Analysis and

Machine Intelligence. 1990, 12(10):993-1001.

[60] 黄红选, 韩继业.数学规划.北京:清华大学出版社.2006.03:242-318.

在读期间的研究成果

1.发表的论文：

[1]赵玲玲,周水生,王雪岩."基于集成算法的半监督学习".《信号处理》,2009,25(8A):320-323.

[2]常甜甜,赵玲玲,刘红卫,周水生."多模型扰动模型动态加权 SVM 集成研究".《计算机工程与应用》.已录用.

[3]吴慧,周水生,赵玲玲."一种新的支持向量机增量学习算法".《计算机工程》.已录用.

2.参加的科研项目：

[1]国家自然科学基金"多核学习研究",编号 60603098.