半监督学习方法

刘建伟 刘 媛 罗雄麟
(中国石油大学(北京)自动化研究所 北京 102249)

摘 要 半监督学习研究如何同时利用有类标签的样本和无类标签的样例改进学习性能,成为近年来机器学习领域的研究热点.鉴于半监督学习的理论意义和实际应用价值,系统综述了半监督学习方法.首先概述了半监督学习的相关概念,包括半监督学习的定义、半监督学习研究的发展历程、半监督学习方法依赖的假设以及半监督学习的分类,然后分别从分类、回归、聚类和降维这 4 个方面详述了半监督学习方法,接着从理论上对半监督学习进行了分析并给出半监督学习的误差界和样本复杂度,最后探讨了半监督学习领域未来的研究方向.

关键词 半监督学习;有类标签的样本;无类标签的样例;类标签;成对约束
中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2015.01592

Semi-Supervised Learning Methods

LIU Jian-Wei LIU Yuan LUO Xiong-Lin
(*Research Institute of Automation, China University of Petroleum, Beijing* 102249)

Abstract Semi-supervised learning is used to study how to improve performance in the presence of both examples and instances, and it has become a hot area of machine learning field. In view of the theoretical significance and practical value of semi-supervised learning, semi-supervised learning methods were reviewed in this paper systematically. Firstly, some concepts about semi-supervised learning were summarized, including definition of semi-supervised learning, development of research, assumptions relied on semi-supervised learning methods and classification of semi-supervised learning. Secondly, semi-supervised learning methods were detailed from four aspects, including classification, regression, clustering, and dimension reduction. Thirdly, theoretical analysis on semi-supervised learning was studied, and error bounds and sample complexity were given. Finally, the future research on semi-supervised learning was discussed.

Keywords semi-supervised learning; labeled examples; unlabeled instances; label; pair-wise constraints

1 引 言

半监督学习(Semi-Supervised Learning,SSL)是机器学习(Machine Learning,ML)领域中的研究热点,已经被应用于解决实际问题,尤其是自然语言处理问题.SSL 被研究了几十年,国内外涌现出大量关于该领域的研究工作,研究人员在这个问题上已经取得了显著的进步,目前已经有多个文献对 SSL 领域进行了综述,例如文献[1]综述了早期 SSL 的一些进展,文献[2]对 SSL 进行了比较全面的综述,文献[3]对基于不一致的 SSL 方法进行了综述,文献[4]详细综述了协同训练风范.由于 SSL 研究的发展非常迅速,因此需要有更新的综述来对近几年

收稿日期:2013-08-12;最终修改稿收到日期:2014-08-28.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB720500)、国家自然科学基金(21006127)、中国石油大学(北京)基础学科研究基金项目(JCXK-2011-07)资助.**刘建伟**,男,1966 年生,博士,副研究员,主要研究方向为智能信息处理、复杂非线性系统分析、预测与控制、算法分析与设计.E-mail:liujw@cup.edu.cn.**刘 媛**,女,1989 年生,硕士研究生,主要研究方向为机器学习、数字图像处理.**罗雄麟**,男,1963 年生,博士,教授,主要研究领域为智能控制、复杂非线性系统分析、预测与控制.

SSL 研究的相关情况进行总结.

鉴于 SSL 的理论意义和实际应用价值，本文系统综述 SSL 方法的研究进展，为进一步深入研究 SSL 理论和拓展其应用领域奠定一定的基础. 本文第 2 节概述 SSL 的基本概念、研究历程、依赖的假设及分类；第 3 节到第 6 节分别介绍用于分类、回归、聚类、降维问题的 SSL 方法；第 7 节对 SSL 进行理论分析，综述 SSL 的抽样复杂性和误差界；第 8 节展望未来的研究方向；第 9 节对全文进行总结.

2　半监督学习概述

ML 有两种基本类型的学习任务：

(1) 监督学习(Supervised Learning，SL)根据输入-输出样本对 $\boldsymbol{L}=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$ 学习输入到输出的映射 $f:\boldsymbol{X}\to\boldsymbol{Y}$，来预测测试样例的输出值. SL 包括分类(Classification)和回归(Regression)两类任务，分类中的样例 $\boldsymbol{x}_i\in\boldsymbol{R}^m$，类标签 $y_i\in\{c_1,c_2,\cdots,c_C\}$，$c_j\in\boldsymbol{N}$；回归中的输入 $\boldsymbol{x}_i\in\boldsymbol{R}^m$，输出 $y_i\in\boldsymbol{R}$. 具有代表性的 SL 方法有线性判别分析(Linear Discriminative Analysis，LDA)、偏最小二乘(Partial Least Square，PLS)、支持向量机(Support Vector Machine，SVM)、K 近邻(K-Nearest Neighbor，KNN)、朴素贝叶斯(Naive Bayes)、逻辑斯蒂回归(Logistic Regression)、决策树(Decision Tree)和神经网络等.

(2) 无监督学习(Unsupervised Learning，UL)利用无类标签的样例 $\boldsymbol{U}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ 所包含的信息学习其对应的类标签 $\hat{\boldsymbol{Y}}_u=[\hat{y}_1\cdots\hat{y}_n]^{\mathrm{T}}$，由学习到的类标签信息把样例划分到不同的簇(Cluster)或找到高维输入数据的低维结构. UL 包括聚类(Clustering)和降维(Dimensionality Reduction)两类任务. 具有代表性的 UL 方法有 K 均值(K-Means)、层次聚类(Hierarchical Clustering)、主成分分析(Principal Component Analysis，PCA)、典型相关分析法(Canonical Correlation Analysis，CCA)、等距特征映射(Isometric Feature Mapping，ISOMAP)、局部线性嵌入(Locally Linear Embedding，LLE)和局部保持投影(Locality Preserving Projections，LPP)等.

在许多 ML 的实际应用中，如网页分类、文本分类、基因序列比对、蛋白质功能预测、语音识别、自然语言处理、计算机视觉和基因生物学，很容易找到海量的无类标签的样例，但需要使用特殊设备或经过昂贵且用时非常长的实验过程进行人工标记才能得到有类标签的样本，由此产生了极少量的有类标签的样本和过剩的无类标签的样例[5]. 因此，人们尝试将大量的无类标签的样例加入到有限的有类标签的样本中一起训练来进行学习，期望能对学习性能起到改进的作用，由此产生了 SSL[1-2]，如图 1 所示. SSL 避免了数据和资源的浪费，同时解决了 SL 的模型泛化能力不强和 UL 的模型不精确等问题.

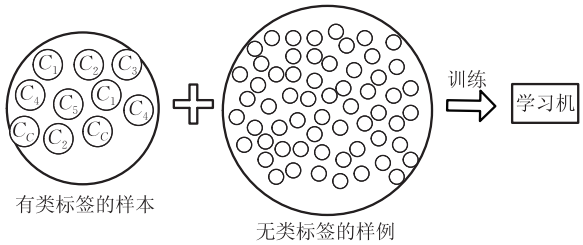


图 1　半监督学习示意

2.1　半监督学习研究的发展历程

SSL 的研究历史可以追溯到 20 世纪 70 年代，这一时期，出现了自训练(Self-Training)、直推学习(Transductive Learning)、生成式模型(Generative Model)等学习方法. Scudder[6]、Fralick[7] 和 Agrawala[8] 提出的自训练方法是最早将无类标签的样例用于 SL 的方法. 该方法是打包算法，在每一轮的训练过程中反复运用 SL 方法，将上一轮标记结果最优的样例和它的类标签一起加入到当前训练样本集中，用自己产生的结果再次训练自己. 这种方法的优点是简单，缺点是学习性能依赖于其内部使用的 SL 方法，可能会导致错误的累积. 直推学习的概念最先由 Vapnik 于 1974 年提出[1]. 与归纳学习(Inductive Learning)不同，直推学习只预测当前训练数据和测试数据中无类标签的样例的类标签，而不推断整个样本空间的广义决策规则. Cooper 等人提出的生成式模型假设生成数据的概率密度函数为多项式分布模型，用有类标签的样本和无类标签的样例估计该模型中的参数[1]. 后来，Shahshahani 和 Landgrebe 将这种每类单组分的场景拓展到每类多组分，Miller 和 Uyar 进一步将其推广[1]. 这一时期，McLachlan 等人研究用无类标签的样例估计费希尔线性判别(Fisher Linear Discriminative，FLD)规则的问题[1].

对 SSL 的研究到了 20 世纪 90 年代变得更加狂热，新的理论的出现，以及自然语言处理、文本分类和计算机视觉中的新应用的发展，促进了 SSL 的发展，出现了协同训练(Co-Training)和转导支持向量机(Transductive Support Vector Machine，TSVM)等

新方法. Merz 等人[9]在 1992 年提出了 SSL 这个术语,并首次将 SSL 用于分类问题. 接着 Shahshahani 和 Landgrebe[10]展开了对 SSL 的研究. 协同训练方法由 Blum 和 Mitchell[11]提出,基于不同的视图训练出两个不同的学习机,提高了训练样本的置信度. Vapnik 和 Sterin[12]提出了 TSVM,用于估计类标签的线性预测函数. 为了求解 TSVM,Joachims[13]提出了 SVM^{light} 方法,De Bie 和 Cristianini[14]将 TSVM 放松为半定规划问题从而进行求解. 许多研究学者研究将期望最大算法(Expectation Maximum,EM)与高斯混合模型(Gaussian Mixture Model,GMM)相结合的生成式 SSL 方法[15-16]. Blum 等人[17]提出了最小割法(Mincut),首次将图论应用于解决 SSL 问题. Zhu 等人[18]提出的调和函数法(Harmonic Function)将预测函数从离散形式扩展到连续形式. 由 Belkin 等人[19]提出的流形正则化法(Manifold Regularization)将流形学习的思想用于 SSL 场景. Klein 等人[20]提出首个用于聚类的半监督距离度量学习方法,学习一种距离度量.

研究人员通过理论研究和实验对 SSL 的学习性能进行了分析. Castelli 和 Cover[21]在服从高斯混合分布的无类标签的样例集中引入了一个新的有类标签的样本,通过理论分析证明了在无类标签的样例数量无限的情况下,可识别的混合模型的分类误差率以指数形式快速收敛到贝叶斯风险. Sinha 和 Belkin[22]从理论上研究了当模型不完善时使用无类标签的样例对学习性能产生的影响. Balcan 和 Blum[23]以及 Singh 等人[24]用概率近似正确(Probably Approximately Correct,PAC)理论和大偏差界理论分析了基于判别方法的 SSL 方法的性能,给出了说明无类标签的样例何时帮助改进学习性能的相容性函数. Balcan 等人[25]在理论上说明了在每个视图给定适当强的 PAC 学习机,仅依赖比充分冗余假设更弱的假设,也足以使协同训练迭代成功. Goldberg 和 Zhu[26]将基于图的 SSL 方法用于情绪分级问题,证明了无类标签的样例能够改进学习性能. Leskes 说明当协同训练的不同学习机在相同的给定训练数据集上得到的结果一致时,训练结果的误差减小[27].

在 SSL 成为一个热门研究领域之后,出现了许多利用无类标签的样例提高学习算法预测精度和加快速度的学习方法,因此出现了大量改进的 SSL 方法. Nigam 等人[28]将 EM 和朴素贝叶斯结合,通过引入加权系数动态调整无类标签的样例的影响提高了分类准确率,建立每类中具有多个混合部分的模型,使贝叶斯偏差减小. Zhou 和 Goldman[29]提出了协同训练改进算法,不需要充分冗余的视图,而利用两个不同类型的分类器来完成学习. Zha 等人[30]提出了一种解决多类标签问题的基于图的 SSL 方法. Zhou 和 Li[3]提出了基于差异的 SSL 方法,利用多个学习机之间的差异性来改进 SSL 性能,有效地降低了时间损耗,并且提高了学习机的泛化能力. Wu 等人[31]引入一种密度敏感的距离度量,并结合基于图的方法,显著提高了算法的聚类性能. Xing 等人[32]引入度量学习的思想进行聚类,并通过实验说明用成对约束的马氏距离度量能提高聚类的准确性. Yu 等人[33]将类标签信息引入概率 PCA 模型处理多输出问题,具有较好的可扩展性. Hwa 等人[34]将主动学习与 SSL 相结合,提出一种基于协同训练的主动半监督句法分析方法,实验结果显示该方法可以减少大量的人工标记量. Johnson 和 Zhang[35]将基于频谱分解的无监督核与基于图的方法结合,提高了预测性能. Mallapragada 等人[36]提出一种 SSL 的改进框架,提高了已有方法的分类准确性. Shin 等人[37]提出解决反向边问题的方法,提高了学习性能. Shang 等人[38]提出一种新的 SSL 方法——核归一正则化 SSL 方法(Semi-Supervised Learning with Nuclear Norm Regularization,SSL-NNR),能同时解决有类标签样本稀疏和具有附加无类标签样例成对约束的问题. Wang 等人[39]提出双变量的基于图 SSL 方法,将二值类标签信息和连续分类函数同时用于优化学习问题.

随着 SSL 技术的发展,SSL 已用于解决实际问题. 例如,Yarowsky[5]用协同训练从两个视图构造不同的分类器对词义进行消歧,其中一个分类器利用文本中该词的上下文,另一个分类器基于该文本中其他地方出现的该词的意义;Riloff 和 Jones[40]同时考虑名词及该词出现的语境,实现了对地理位置名词的分类;Collins 和 Singer[41]同时利用实体的拼写和该实体出现的上下文,完成了对命名实体的分类;Yu 等人[42]完成了对中文问题的分类;Li 和 Zhou[43]对三训方法进行了扩展,并将该方法用于乳腺癌诊断中的微钙化检测;Zhou 等人[44]将协同训练用于图像检索;Goldberg 和 Zhu[26]利用基于图的方法解决了情绪分级问题;Chen 等人[45]将标签传播法用于关系抽取;Camps-Valls 等人[46]提出基于

图的混合核分类方法，并将其应用于解决超光谱图像问题；Cheng 等人[47]提出一种基于半监督分类器的粒子群优化算法用于解决中文文本分类问题；Zhang 等人[48]提出一种基于图的多样例学习方法用于各种视频领域研究；Carlson 等人[49]将耦合 SSL 用于从网页提取类别和关系的信息；Guillaumin 等人[50]将多模态 SSL 用于图像分类；He[51]将半监督子空间学习用于图像检索；Balcan 等人[52]用基于图的 SSL 方法进行低质量摄像头图像中的身份识别；Wang 等人[53]提出半监督散列方法用于处理大规模图像检索问题.

2.2　半监督学习依赖的假设

SSL 的成立依赖于模型假设，当模型假设正确时，无类标签的样例能够帮助改进学习性能[10]. SSL 依赖的假设有以下 3 个：

(1) 平滑假设(Smoothness Assumption). 位于稠密数据区域的两个距离很近的样例的类标签相似，也就是说，当两个样例被稠密数据区域中的边连接时，它们在很大的概率下有相同的类标签；相反地，当两个样例被稀疏数据区域分开时，它们的类标签趋于不同.

(2) 聚类假设(Cluster Assumption)[1,54]. 当两个样例位于同一聚类簇时，它们在很大的概率下有相同的类标签. 这个假设的等价定义为低密度分离假设(Low Sensity Separation Assumption)，即分类决策边界应该穿过稀疏数据区域，而避免将稠密数据区域的样例分到决策边界两侧.

(3) 流形假设(Manifold Assumption)[4,55]. 将高维数据嵌入到低维流形中，当两个样例位于低维流形中的一个小局部邻域内时，它们具有相似的类标签.

许多实验研究表明当 SSL 不满足这些假设或模型假设不正确时，无类标签的样例不仅不能对学习性能起到改进作用，反而会恶化学习性能，导致 SSL 的性能下降. 但是还有一些实验表明，在一些特殊的情况下即使模型假设正确，无类标签的样例也有可能损害学习性能[55]. 例如，Shahshahani 和 Landgrebe[10]通过实验证明了如何利用无类标签的样例帮助减轻休斯现象(Hughes Phenomenon)(休斯现象指在样例数量一定的前提条件下，分类精度随着特征维数的增加先增后降的现象)，但是同时实验中也出现了无类标签的样例降低学习性能的情况. Baluja[56]用朴素贝叶斯分类器和树扩展朴素贝叶斯(Tree Augmented Naïve Bayesian，TAN)分类器得到很好的分类结果，但是其中也存在无类标签的样例降低学习性能的情况. Balcan 和 Blum[57]提出容许函数使分类器能够很好的服从无类标签的样例的分布，但是这种方法同样会损害学习性能.

2.3　半监督学习的分类

SSL 按照统计学习理论的角度包括直推(Transductive) SSL[58]和归纳(Inductive) SSL 两类模式. 直推 SSL 只处理样本空间内给定的训练数据，利用训练数据中有类标签的样本和无类标签的样例进行训练，预测训练数据中无类标签的样例的类标签；归纳 SSL 处理整个样本空间中所有给定和未知的样例，同时利用训练数据中有类标签的样本和无类标签的样例，以及未知的测试样例一起进行训练，不仅预测训练数据中无类标签的样例的类标签，更主要的是预测未知的测试样例的类标签.

从不同的学习场景看，SSL 可分为 4 大类：

(1) 半监督分类(Semi-Supervised Classification)[11,59]. 在无类标签的样例的帮助下训练有类标签的样本，获得比只用有类标签的样本训练得到的分类器性能更优的分类器，弥补有类标签的样本不足的缺陷，其中类标签 y_i 取有限离散值 $y_i \in \{c_1, c_2, \cdots, c_C\}, c_j \in \boldsymbol{N}$.

(2) 半监督回归(Semi-Supervised Regression)[60-61]. 在无输出的输入的帮助下训练有输出的输入，获得比只用有输出的输入训练得到的回归器性能更好的回归器，其中输出 y_i 取连续值 $y_i \in \boldsymbol{R}$.

(3) 半监督聚类(Semi-Supervised Clustering)[62-63]. 在有类标签的样本的信息帮助下获得比只用无类标签的样例得到的结果更好的簇，提高聚类方法的精度.

(4) 半监督降维(Semi-Supervised Dimensionality Reduction)[64]. 在有类标签的样本的信息帮助下找到高维输入数据的低维结构，同时保持原始高维数据和成对约束(Pair-Wise Constraints)的结构不变，即在高维空间中满足正约束(Must-Link Constraints)的样例在低维空间中相距很近，在高维空间中满足负约束(Cannot-Link Constraints)的样例在低维空间中距离很远.

为便于更加清晰地介绍各种 SSL 方法，这里按照图 2 对各种 SSL 方法进行归类.

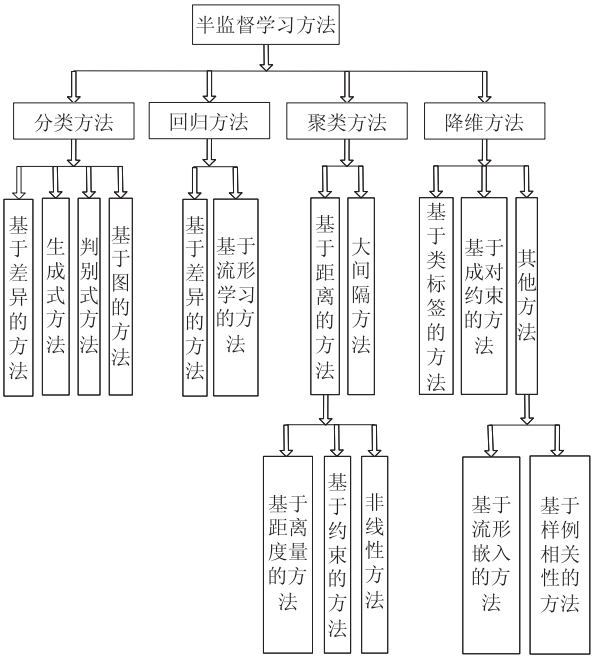


图 2 半监督学习方法结构

3 半监督分类方法

半监督分类问题是 SSL 中最常见的问题，其中有类标签的样本数量相比聚类问题多一些，引入大量的无类标签的样例 $\boldsymbol{U}=\{\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$ 和 $\boldsymbol{T}=\{\boldsymbol{x}_{test_1},\cdots,\boldsymbol{x}_{test_t}\}$ 弥补有类标签的样本 $\boldsymbol{L}=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$ 不足的缺陷，改进监督分类方法的性能，训练得到分类性能更优的分类器，从而预测无类标签的样例的类标签. 其中样例 $\boldsymbol{x}_i\in\boldsymbol{R}^m$，类标签 $y_i\in\{c_1,c_2,\cdots,c_C\}$，$i=1,\cdots,l,\cdots,l+u,\cdots,l+u+t$，训练样例数量为 $n_{train}=l+u$，测试样例数量为 $n_{test}=t$. 主要的半监督分类方法有基于差异的方法(Disagreement-Based Methods)、生成式方法(Generative Methods)、判别式方法(Discriminative Methods)和基于图的方法(Graph-Based Methods)等，下面分别对这几种方法进行描述与分析.

3.1 基于差异的方法

ML 中的数据有时可以用多种方式表示其特征. 例如，在网页分类问题中，网页可以用每页出现的词描述，也可以用超链接描述；癌症诊断可以用 CT、超声波或 MRI 等多种医学图像技术确定患者是否患有癌症. 基于这些朴素的思想，产生了基于差异的方法.

1998 年，Blum 和 Mitchell[11] 提出了协同训练方法. 如图 3 所示，协同训练方法的基本训练过程为：在有类标签的样本的两个不同视图(View)上分别训练，得到两个不同的学习机，然后用这两个学习机预测无类标签的样例的类标签，每个学习机选择标记结果置信度最高的样例和它们的类标签加入另一个学习机的有类标签的样本集中. 这个过程反复迭代进行，直到满足停止条件. 这个方法需要满足两个假设条件：(1) 视图充分冗余(Sufficient and Redundant)假设，即给定足够数量的有类标签的样本，基于每个视图都能通过训练得到性能很好的学习机；(2) 条件独立假设，即每个视图的类标签都条件独立于另一视图给定的类标签.

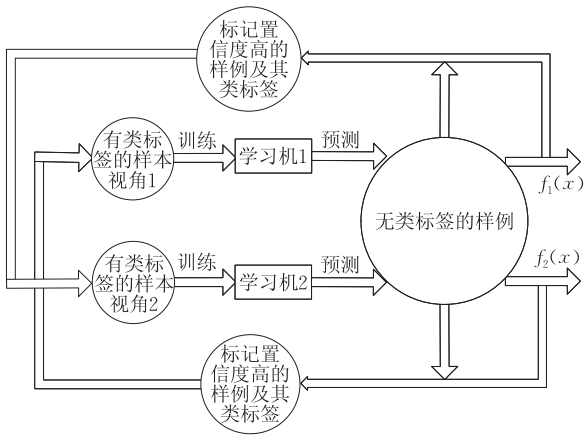


图 3 协同训练方法示意

许多研究人员通过理论分析和实验证明了基于差异的方法的有效性. Dasgupta 等人[65] 从理论上说明，当训练数据满足视图充分冗余假设时，基于差异的方法通过使基于不同视图的学习机在无类标签的样例上的一致性达到最大化，得到相同的分类预测结果，可以降低误分类率. Zhou 等人[66] 证明当训练数据满足视图充分冗余假设时，即使只给定一个有类标签的样本，也能有效地进行 SSL. Wang 和 Zhou[67] 进行了理论证明和实验验证，理论结果显示出，基于差异的方法并不是必须具备多个视图，为单视图类型的方法提供了理论支持.

尽管基于差异的方法已经广泛应用于许多实际领域，如统计语法分析、名词短语识别等，但是在大多数实际问题中，训练数据往往不满足视图充分冗余假设. 因此，研究人员开始致力于研究基于放松的视图充分冗余假设或不需要满足视图充分冗余假设的基于差异的方法. Nigam 和 Ghani[68] 在不具有充分冗余视图的问题上对基于差异的方法的性能进行了研究，通过实验证明，将训练数据随机划分到两个视图后，基于差异的方法的误分类率明显降低. 2000 年，Goldman 和 Zhou[69] 提出基于差异的改进方法，这个方法不需要训练数据满足视图充分冗余假设，而

是用两个不同的 SL 方法，将样本空间分到一组等价类中，通过交叉校验来确定如何对无类标签的样例进行标记. 2002 年，Abney[70] 提出一种使无类标签的样例的一致性最大化的贪婪算法，在命名实体分类基于差异的训练实验中产生好的学习效果. 2003 年，Clark 等人[55] 提出间接寻找无类标签的样例的最大一致性的朴素基于差异的训练过程. 2004 年，Zhou 和 Goldman[29] 通过使用多个不同类型的学习机对之前提出的基于差异的训练改进方法进行了扩展，在一定程度上放宽了标准协同训练方法的假设条件，但是这个方法要求两个学习机所采用的学习方法能够将样本空间划分为等价类集合，而且训练过程耗时很大. 为了解决这个问题，2005 年，Zhou 和 Li[71] 提出了三训方法(Tri-Training)，用三个学习机分别进行训练，按投票选举的方式间接得到标记置信度，如果两个学习机对同一个无类标签的样例的预测结果相同，则认为该样例具有较高的标记置信度，将其与它的类标签加入到第三个学习机的训练数据集中. 他们在 UCI 数据集和网页分类问题上进行实验，证明能够有效地利用无类标签的样例提高学习机性能. 三训方法利用三个学习机来选择标记置信度，不仅有效地降低了时间耗费，而且能够利用集成学习提高学习机的泛化能力. 但是当初始学习机性能较差时，在训练过程中将会引入噪声，导致预测精度下降. 为此，2007 年，Li 和 Zhou[43] 对三训方法进行了扩展，提出可以更好发挥集成学习作用的 Co-Forest 方法，并将这个方法用于乳腺癌诊断中的微钙化检测，通过实验证明这个方法能够有效提高预测精度.

基于差异的方法由于性能优越而得到了广泛的应用，由此出现了许多变形[72]. Nigam 和 Ghani[68] 提出协同 EM 方法，只用有类标签的样本初始化第一视图学习机，然后用这个学习机以概率方式标记所有无类标签的样例，第二视图学习机训练所有数据，将得到的新的样本提供给第一视图学习机进行再训练. 这个过程反复迭代进行，直到学习机的预测结果收敛. Steedman 等人[73] 提出了一种基于差异训练的统计句法分析方法，用两个功能完整的不同统计句法分析机进行基于差异的训练，通过实验证明，基于差异的训练方法能够显著提高句法分析机的性能. Hwa 等人[34] 将主动学习与 SSL 相结合，提出一种基于差异训练的主动半监督句法分析方法，在学习过程中，一个学习机挑选并标记自己最确定的样本给另一个学习机，而另一个学习机则挑选自己最不确定的样本请用户标记后再提交给该学习机用于模型更新. 他们的研究结果表明，该方法可以减少大约一半的人工标记量. Zhou 等人[44] 将基于差异的训练引入图像检索，提出了基于差异训练的主动半监督相关反馈方法. Wang 和 Zhou[74] 将基于差异的方法和基于图的方法结合. Yan 等人[75] 提出一种概率 SSL 模型，用多个分类器进行学习，并通过实验证明了该方法的优越性能.

3.2　生成式方法

生成式方法假定样例和类标签由某个或有一定结构关系的某组概率分布生成，已知类先验分布 $p(y)$ 和类条件分布 $p(\boldsymbol{x}|y)$，重复取样 $y\sim p(y)$ 和 $\boldsymbol{x}\sim p(\boldsymbol{x}|y)$，从这些分布中生成有类标签的样本 $\boldsymbol{L}$ 和无类标签的样例 $\boldsymbol{U}$. 根据概率论公理得到后验分布 $p(y|\boldsymbol{x})$，找到使 $p(y|\boldsymbol{x})$ 最大的类标签对 $\boldsymbol{x}$ 进行标记[76-77].

生成样例的模型有高斯模型、贝叶斯网络、S 型信度网(Sigmoidal Belief Networks)、GMM、多项混合模型(Multinomial Mixture Model，MMM)、隐马尔可夫模型(Hidden Markov Model，HMM)和隐马尔可夫随机场模型(Hidden Markov Random Field，HMRF)等.

(1) 高斯模型[10] 中的样例服从高斯分布

$$p(\boldsymbol{x}|y)=N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$$
$$=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad (1)$$

式(1)中 $\boldsymbol{\mu}$ 是均值，$\boldsymbol{\Sigma}$ 是协方差阵.

(2) 贝叶斯网络[78] 中的样例的概率分布如图 4 所示.

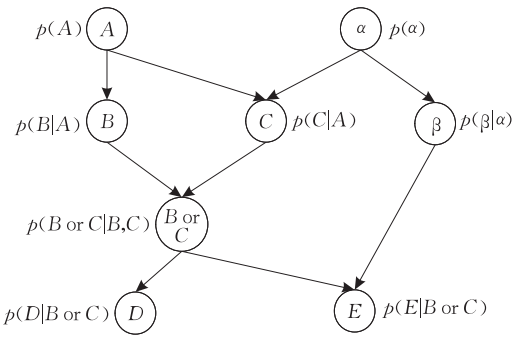


图 4　贝叶斯网络示意

(3) S 型信度网[79] 中的样例服从概率分布

$$p(\boldsymbol{x}_i|pa(\boldsymbol{x}_i))=\frac{\exp\left(\left(\sum_j \boldsymbol{J}_{ij}\boldsymbol{x}_j+h_i\right)\boldsymbol{x}_i\right)}{1+\exp\left(\sum_j \boldsymbol{J}_{ij}\boldsymbol{x}_j+h_i\right)} \quad (2)$$

式(2)中 $pa(\boldsymbol{x}_i)\subseteq\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{i-1}\}$ 表示 $\boldsymbol{x}_i$ 的父节

点，$\boldsymbol{J}_{ij}$ 和 h_i 是网络中的权值和偏差.

(4) GMM[80] 是多个高斯分布的混合分布模型，假定样例由多个模型加权混合生成 $\sum_i \pi_i p_i(\boldsymbol{x}|y)$，$\sum_i \pi_i = 1$，每个模型的分布服从式(1)的高斯分布.

(5) MMM 是多个多模态分布的混合分布模型，假定样例由多个模型加权混合生成 $\sum_i \pi_i p_i(\boldsymbol{x}|\boldsymbol{\mu})$，$\sum_i \pi_i = 1$，每个模型的分布服从多模态分布

$$p(\boldsymbol{x}=(x_{\cdot 1},\cdots,x_{\cdot d})|\boldsymbol{\mu}) = \frac{(\sum_{i=1}^{D} x_{\cdot i})!}{x_{\cdot 1}!\cdots x_{\cdot D}!}\prod_{d=1}^{D}\mu_d^{x_{\cdot d}} \quad (3)$$

式(3)中 $\boldsymbol{\mu}$ 是多个模态共同选择的概率向量，D 是模态数.

(6) HMM[81] 用于建立样例序列的模型，指定状态间的转移概率矩阵按一定周期从一个状态转移到另一状态来形成序列，序列中每个样例由隐状态生成，其中状态条件分布可以是高斯混合分布或多模态混合分布. 当前状态只依赖前一状态，并且输出只依赖当前状态.

(7) HMRF[82] 的每个模型都与之前的模型无关. 定义两个随机场：隐随机场 $\boldsymbol{X}_H$ 和可观测的随机场 $\boldsymbol{X}$. 根据 MRF 的局部特性，当给定 $\boldsymbol{X}_H$ 和它的领域 $\boldsymbol{X}_N$，$(\boldsymbol{X}_H,\boldsymbol{X})$ 的联合概率分布为 $p(\boldsymbol{x},\boldsymbol{x}_H|\boldsymbol{x}_N) = p(\boldsymbol{x}|\boldsymbol{x}_H)p(\boldsymbol{x}_H|\boldsymbol{x}_N)$. $\boldsymbol{X}$ 的边缘条件概率依赖于参数 $\theta=(\mu,\boldsymbol{\Sigma})$ 和 $\boldsymbol{X}_H$ 的领域分布 $\boldsymbol{X}_N$

$$p(\boldsymbol{x}|\boldsymbol{x}_N,\theta) = \sum_{l\in L} p(\boldsymbol{x},l|\boldsymbol{x}_N,\theta) = \sum_{l\in L} p(\boldsymbol{x};\theta_l)p(l|\boldsymbol{x}_N) \quad (4)$$

式(4)中 $l\in\boldsymbol{L}$ 为 $\boldsymbol{X}_H$ 的取值空间，$p(\boldsymbol{x};\theta_l)$ 为 $\boldsymbol{X}$ 的条件概率分布.

常用的生成式方法是朴素贝叶斯分类器，假设样例的各属性条件独立，对样例 $\boldsymbol{x}_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$，$i=1,\cdots,n$ 进行分类的过程实际就是利用朴素贝叶斯计算 $\boldsymbol{x}_i$ 的类标签 $y_i\in\{c_1,c_2,\cdots,c_C\}$ 的后验概率，然后将 $\boldsymbol{x}_i$ 标记为具有最高后验概率的类标签 y_i. 其目标函数为

$$y_i = \arg\max_{y_i=c_1}^{c_C} p(y_i|\boldsymbol{x}_i) = \arg\max p(\boldsymbol{x}_i|y_i)p(y_i) \quad (5)$$

也可以用逻辑斯蒂回归进行训练，预测样例的类标签

$$p(y|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})} \quad (6)$$

式(6)中 $\boldsymbol{\theta}$ 是调节的参数向量.

常用极大似然估计(Maximum Likelihood Estimation，MLE)或最大后验估计(Maximum A Posteriori，MAP)求解这个问题[83-84]. 混合模型或更复杂的生成式模型的目标函数非凸并且难于优化，无法用 MLE 或 MAP 解析地求解，这种情况下通常用直接梯度下降法或 EM 算法等迭代算法求解得到局部极大值.

3.3 判别式方法

判别式方法利用最大间隔算法同时训练有类标签的样本和无类标签的样例学习决策边界，如图 5 所示，使其通过低密度数据区域，并且使学习得到的分类超平面到最近的样例的距离间隔最大[85-86].

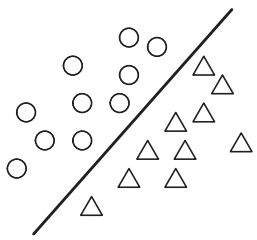


图 5 判别式方法示意

判别式方法包括 LDA、广义判别分析法(Generalized Discriminant Analysis，GDA)、半监督支持向量机(Semi-Supervised Support Vector Machine，S^3VM)、熵正则化法和 KNN 法等.

LDA 也叫做费希尔线性判别法(Fisher Linear Discriminative Analysis，FDA)，最初由 Fisher[87] 于 1936 年提出，其基本思想是将样例投影到合适维数的低维空间中，使投影后的样例在新的子空间中有最大的类间距离和最小的类内距离，即样例按照类别能被分成许多簇. Baudat 和 Anouar[88] 将 LDA 发展到多类问题，提出 GDA. 通过一个非线性映射，将样例映射到高维特征空间，在这个特征空间中运用 FDA 进行训练. LDA 和 GDA 用于 SSL 时，有一部分样例的类标签的值是不知道的，目标是要求解类标签的值，由于 $y_i\in\{c_1,c_2,\cdots,c_C\}$，因此这是一个混合整数规划问题.

TSVM 最初由 Vapnik 和 Sterin[12] 提出，用于估计类标签的线性预测函数 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$，由于它也可以用来估计未知的测试样例的类标签，实际得到的是整个样例空间上的决策边界，不是严格的直推方法，而是归纳半监督方法，因此被称为 S^3VMs. TSVM 的目标函数为

$$\min_{\boldsymbol{w},b,\boldsymbol{Y}_u} \frac{1}{2}\|\boldsymbol{w}\|^2 + C_1\sum_{i=1}^{l} V(y_i,f(\boldsymbol{x}_i)) + C_2\sum_{i=l+1}^{l+u} V(\hat{y}_i,f(\boldsymbol{x}_i)) \quad (7)$$

式(7)中 $V(y,f(\boldsymbol{x}))$ 是损失函数，两个正则化参数 C_1 和 C_2，是有类标签和无类标签的样例上损失的加

权值，分别用于权衡有类标签和无类标签的数据上的复杂度和实际误差. 式(7)非凸并且难以找到全局最优解，可以用半定规划(Semi-Definite Programming，SDP)、分支界定法(Branch and Bround，BB)、确定性模拟退火算法(Deterministic Annealing，DA)和同伦连续法等方法求解.

熵正则化法[1,89-90]采用香农条件熵来度量类之间的重叠程度，其目标函数为

$$\max_{\theta}\frac{1}{2}\|\boldsymbol{w}\|^2+C_1\sum_{i=1}^{l}\ln p(y_i\mid\boldsymbol{x}_i,\theta)+\lambda\sum_{i=l+1}^{n}\sum_{y_i=c_1}^{c_C}p(y_i\mid\boldsymbol{x}_i,\theta)\ln p(y_i\mid\boldsymbol{x}_i,\theta)\tag{8}$$

KNN 法[91]在所有样例中找到与测试样例距离最近的 k 个近邻样例，其中各类所占的数量为 k_j，$j=c_1,\cdots,c_C$，用决策规则 $\arg\max_j k_j, j=c_1,\cdots,c_C$ 选择类标签对样例进行标记.

3.4　基于图的方法

基于图的方法的实质是标签传播(Label Propagation)，基于流形假设，根据样例之间的几何结构构造图(Graph)，用图的结点(Vertice)表示样例，利用图上的邻接关系将类标签从有类标签的样本向无类标签的样例传播.

如图 6 所示，基于图的方法的基本训练过程为：(1)选择合适的距离函数计算样例间的距离. 可供选取的距离函数有欧氏距离、曼哈顿距离、切比雪夫距离、明氏距离、马氏距离和归一化欧氏距离等；(2)根据计算得到的距离选择合适的连接方式，将样例用边(Edge)连接，构造连接图. 构造的连接图分为稠密图(Dense Graph)和稀疏图(Sparse Graph)，稠密图的典型代表是全连接图，如图 7(a)所示，任意两个结点之间都有边连接；稀疏图如图 7(b)所示，按照某种准则将距离最近的某几个结点连接，包括 KNN 图、ε 近邻(ε-Nearest Neighbor，εNN)图、正切权图和指数权图等；(3)用核函数(Kernel)给图的连接边赋权值(Weight)，用权反映这个边所连接的两个结点之间的相似程度，当两个结点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 距离很近时，连接这两个结点的边的权 w_{ij} 就很大，这两个样例有相同的类标签的概率就很大；反之，当两个结点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 距离很远时，连接这两个结点的边的权 w_{ij} 就很小，这两个样例有相同类标签的概率就很小. 常用的核函数有线性核、多项式核、高斯核、径向基核、双曲正切核、神经网络核、费希尔核和样条核等；(4)根据学习目标确定优化问题并求解. 半监督分类问题的目标是找到使目标函数最小的类标签的预测函数 $f(\boldsymbol{x})$，这个问题可以看作是一个由损失函数和正则化函数组成的复合目标函数的正则化风险最小化问题[92-94]，因此基于图的方法解决半监督分类问题的目标函数一般表示为

$$\min_{f(\boldsymbol{x})}V(y,f(\boldsymbol{x}))+\lambda\Omega(f)\tag{9}$$

式(9)中损失函数 $V(y,f(\boldsymbol{x}))$ 用来惩罚样例的预测类标签不等于给定类标签的情况，正则化函数 $\Omega(f)$ 用来保证预测函数的平滑性，使近邻点的预测类标签相同. 根据具体的学习任务可以选择不同的损失函数和正则化函数，如损失函数可以选取平方误差函数、绝对值函数、对数函数、指数函数和铰链损失函数等. 一般将损失函数和正则化函数限制在再复制核希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)中，用核学习算法求解学习机[95].

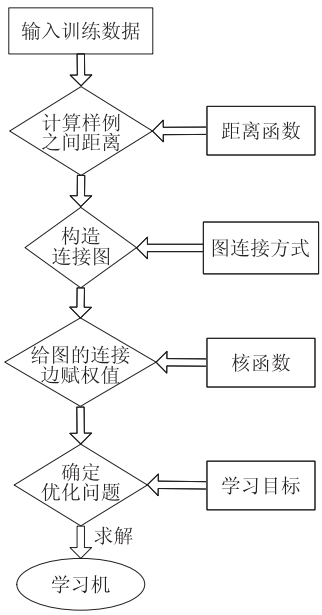
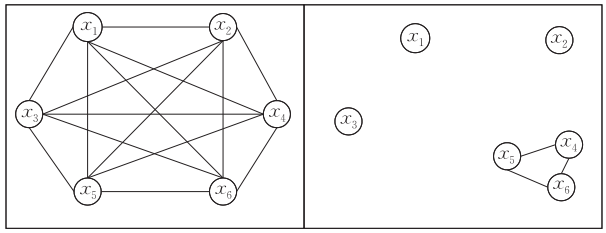


图 6　基于图的方法训练过程示意



(a) 全连接图　　(b) 稀疏图

图 7　连接图示意

2001 年，Blum 和 Chawla[17]提出第一个基于图的 SSL 方法——最小割法(Mincut)，将已标记为正的样例看作源结点，已标记为负的样例看作目标结点，找到一组边，使移除这些边之后，源结点和目标结点之间没有连接，即图被分割成两个独立的云团，将

这组边称为图割(Cut).当图被分割后,将连接源结点的结点类标签标记为正,将连接目标结点的结点类标签标记为负.这个方法选择带无限权的平方损失作为损失函数 $V(y,f(\boldsymbol{x}))=\infty\cdot\sum_{i=1}^{l}(y_i-f(\boldsymbol{x}_i))^2$,用割的大小表示正则化函数 $\Omega(f)=\boldsymbol{f}^{\mathrm{T}}L\boldsymbol{f}$,并且限制在任意有类标签的结点 $\boldsymbol{x}_i$ 上,预测类标签等于给定类标签 $f(\boldsymbol{x}_i)=y_i$.随后,Blum 等人[96]通过人工引入边权的随机噪声,指出基本最小割法的一些缺点,并对最小割法进行改进.求解最小割法的目标函数实际上就是求解函数 $f(\boldsymbol{x})\in\{+1,-1\}$,如果忽略 $f(\boldsymbol{x})$ 上的整数约束,那么这个问题就变为凸二次优化问题,此时有近似形式的解.但是与此同时,在没有约束的情况下,无法得到唯一的确定解.为了解决这个问题,Zhu 等人[97]提出比例割法(Ratio Cut),在正则化函数中引入分割后云团的势来使分割后的云团所包含结点的数量平衡,可以得到唯一的解.但是在许多实际情况中,边的权存在较大差异,此时这种比例割法不一定能得到较好的解.为此,Zhou 等人[98]提出归一化割法(Normalized Cut),损失函数同时惩罚有类标签和无类标签的样例上预测类标签不同于给定类标签的情况,同时为了充分考虑到分割后云团中权的分布的平衡特性,在正则化函数中引入分割后云团的大小.2003 年,Zhu 等人[18]提出调和函数法(Harmonic Function),也称为高斯随机场法(Gauss Random Field),在图上建立预测函数的分布模型,从而将离散预测函数扩展为连续预测函数.与最小割法一样,这个方法也选择带无限权的平方损失作为损失函数,用割的大小表示正则化函数,区别只在于预测函数取连续值,而不是直接等于类标签值,充分考虑了样例的分类概率,解决了最小割法不能解决的问题.2004 年,Belkin 和 Niyogi[99]提出拉普拉斯正则化法(Laplacian Regularization),假定训练数据分布在流形(Manifold)上,用离散频谱和近邻图的特征函数将学习问题嵌入到希尔伯特空间中.Zhou 等人[100]用归一化图拉普拉斯算子阵作为正则化因子,提出了一种迭代的标签传播法.上面介绍的方法都是转导 SSL 方法,无法直接处理训练数据之外的数据,必须重新构造图进行重新计算,并且这些方法都限制有类标签的样本的预测类标签必须等于给定类标签,无法解决训练数据中有噪声的问题.2005 年,Chen 和 Wang[101]通过用图拉普拉斯算子组合产生式混合模型和判别式正则化项来克服非归纳和适用性的限制.2006 年,Belkin 等人[19]将流形学习的思想用于 SSL 场景,提出一种归纳 SSL 方法——流形正则化法(Manifold Regularization).这个方法并不限制有类标签的样本的预测类标签必须等于给定类标签,损失函数为 $V(y,f(\boldsymbol{x}))=\|\boldsymbol{f}-\boldsymbol{Y}_l\|^2$,并且有两个正则化函数,其中一个 $\Omega_1(f)=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{f}$ 用来控制预测函数的复杂性,用图拉普拉斯算子阵使 f 平滑变化,另一个 $\Omega_2(f)=\|\boldsymbol{f}\|^2$ 的作用是保持样本分布的内在流形结构,防止 f 在测试数据上发生急剧改变,波动太大.2008 年,Goldberg 等人[102]提出在线流形正则化结构,提高了流形正则化在大规模数据和实时问题中的适用性.Yan 和 Wang[103]基于 l_1 图提出一种新的 SSL 结构,l_1 图的想法来源于每个数据可以通过训练数据的稀疏线性叠加进行重构,通过解决稀疏表示上的 l_1 优化问题得到稀疏重构系数,用得到的系数推断出有向 l_1 图的权值,以参数自由的方式同时得到 l_1 图中的近邻结构和权值,并进行人脸识别和图像分类实验,实现结果显示出该方法的性能比传统的图方法更加优越.Liu 等人[104]提出一种处理大型数据集的 SSL 方法,用图正则化执行学习过程.Dhillon 等人[105]提出一种用于半监督结构化输出学习的新方法,用无类标签的样例上的放松标记来解决类标签空间的组合特性,并进一步用领域约束来指导学习.Breve 等人[106]提出一种新的基于图的半监督分类模型,用粒子的组合随机贪婪行走结合竞争和合作机制,将类标签传播到整个网络.Zhang 等人[48]提出快速基于图半监督多样例学习算法,用于搜索并训练小规模专家标记视频和大规模未标记视频得到模型.

近年来,也出现了许多将基于图的方法与其他方法相结合的 SSL 方法.Sindhwani 等人[107]将流形正则化嵌入到定义在整个输入空间的 SSL 核学习结构中,产生新的 RKHS.Tang 等人[108]提出基于敏感型结构的基于图的 SSL 方法.He 等人[109]提出一种基于图的生成式 SSL 方法,利用图的传播来估计类条件概率,用线性回归估计类先验信息.Johnson 和 Zhang[35]将基于频谱分解的无监督核与基于图的 SSL 方法结合,提高了预测性能.Mallapragada 等人[36]提出一种 SSL 的改进框架——SemiBoost,利用无类标签的样例提高已有的 SL 方法的分类准确性.Zha 等人[30]提出用于多类标签 SSL 场景的新的基于图的学习结构.Zhang 和 Wang[110]提出能自动构造最优图的线性近邻传播法,接着提出使这个方法能用于大型数据集的多层次方法.

影响 SSL 性能的除了 SSL 方法的性能之外，还有图本身的性能，图的构造的好坏甚至对学习性能起决定性作用. 因此，有关图的学习问题也受到一些研究人员的关注与研究. Carreira-Perpinan 和 Zemel[111] 提出构造用于学习的鲁棒图；Wang 和 Zhang[112] 用局部线性嵌入的思想得到图的连接边的权；Hein 和 Maier[113] 尝试移除噪声数据来得到更好的图；Shin 等人[37] 提出解决反向边问题的方法，即当图的连接边从无类标签的样例指向有类标签的样本时，调整这个连接边的权来减少不确定信息的传播，从而提高学习性能.

基于图的方法性能优越，具有坚实的数学理论基础、计算速度快及准确性高等优点，已被用于解决一些实际问题. 例如，Goldberg 和 Zhu[26] 利用基于图的方法解决了情绪分级问题，Chen 等人[45] 将标签传播法用于关系抽取，Camps-Valls 等人[46] 提出基于图的混合核分类方法，并将其应用于解决超光谱图像问题.

4　半监督回归方法

与半监督分类方法一样，半监督回归方法引入大量的无输出的输入 $\boldsymbol{U}=\{\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$ 和 $\boldsymbol{T}=\{\boldsymbol{x}_{test_1},\cdots,\boldsymbol{x}_{test_t}\}$ 弥补有输出的输入 $\boldsymbol{L}=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$ 的不足，改进监督学习方法的性能，训练得到性能更优的回归器，从而预测输入的输出. 其中输入 $\boldsymbol{x}_i\in\boldsymbol{R}^m$，输出 $y_i\in\boldsymbol{R}$，$i=1,\cdots,l,\cdots,l+u,\cdots,l+u+t$，训练输入数量为 $n_{train}=l+u$，测试输入数量为 $n_{test}=t$. 虽然在 SL 中，回归问题与分类问题近乎同等重要，但是对 SSL 方法的研究主要集中在半监督分类问题上，而对半监督回归问题的研究比较有限. 产生这个现象的一个原因是半监督分类中的聚类假设对回归问题不一定成立，因此不能直接将大多数半监督分类方法用于回归问题. 但值得庆幸的是，流形假设在回归问题中仍然成立，因此，利用特征空间中的局部平滑性的基于流形学习的半监督回归方法是可行的. 主要的半监督回归方法有基于差异的方法和基于流形学习的方法等，下面分别对这几种方法进行描述与分析.

4.1　基于差异的方法

Zhou 和 Li[114] 最早用协同训练方法进行半监督回归，提出基于协同训练的半监督回归方法——COREG(Co-Training Regressors)，用两个不同阶明氏距离的 KNN 回归器作为学习机，每个学习机挑选无输出的输入并进行回归预测，然后加入对方的训练集中供对方学习，最后的回归预测结果是两个学习机的预测平均值. 这个方法不要求充分冗余视图，用不同的距离度量或近邻数寻找两个回归机之间的差异，而不需要两个视图的输入，适用于没有自然属性分割的回归问题.

Brefeld 等人[60] 提出一种半监督最小二乘回归方法——coRLSR (co-Regularised Least Squares Regression)，将协同训练用于希尔伯特空间中的归一化风险最小化问题，并提出线性计算无输出的输入数量的半参数近似法.

Ma 和 Wang[115] 提出基于 SVM 协同训练的半监督回归模型，使用两个 SVM 回归模型进行协同训练，适用于解决缺少大量有输出的输入的情况，缓解了只使用单一回归模型造成的错误累加，提高了回归模型的泛化能力，同时由于 SVM 的优势，可用于解决小样本、非线性回归问题.

4.2　基于流形学习的方法

Wang 等人[116] 提出半监督核回归法 (Semi-Supervised Kernel Regression，SSKR)，利用所有观察到的有输出的和无输出的输入，用加权因子调节无输出的输入的影响，并通过实验说明这个方法比传统的核回归和基于图的半监督回归方法更有效. Verbeek 等人[117] 将高斯场框架用于高维数据上的半监督回归. 提出基于熵最小化和极大似然模型选择法的主动学习策略，并将广义 LLE 用于高斯场框架. Pozdnoukhov 等人[118] 将 Belkin 和 Niyogi[99] 提出的方法用于回归问题，通过实验得到这个方法仅适用于线性 ε 不敏感损失函数的回归问题. Yang 等人[119] 提出拉普拉斯正则化框架，导出基于一类广义损失函数的拉普拉斯半监督回归，能够利用数据所在流形的内在几何结构进行回归估计，并给出几种损失函数的拉普拉斯半监督回归方法. Navaratnam 等人[120] 说明如何利用来自边际分布的无输出的输入改进拟合，并用高斯过程隐变量模型学习共隐低维流形特征和参数空间的映射. Ji 等人[121] 提出用于半监督回归问题的算法，将从有类标签和无类标签的样例得到的积分算子的最初几个特征函数看做基函数，利用简单的线性回归过程学习预测函数.

基于流形学习的半监督回归方法利用数据所在流形的内在几何结构进行回归，但是由于流形学习的复杂性，这种方法中的参数较多，并且没有一种指导性的选择参数的方法.

5 半监督聚类方法

半监督聚类问题与分类和回归问题不同,在大量的无类标签的样例 $\boldsymbol{U}=\{\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$ 中引入少量的有类标签的样本 $\boldsymbol{L}=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$,用有类标签的样本包含的监督信息指导算法,将样例 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ 划分到 c 个簇 $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_c$ 中,提高聚类的性能. 其中样例 $\boldsymbol{x}_i\in\boldsymbol{R}^m$,类标签 $y_i\in\{c_1,c_2,\cdots,c_C\}$,$i=1,\cdots,l,\cdots,l+u$,训练样例数量为 $n=l+u$,簇 $\boldsymbol{C}_k$ 的中心为 $\boldsymbol{m}_k(k=1,\cdots,c)$. SSL 中可利用的监督信息除了类标签之外,还有成对约束. 成对约束包括正约束和负约束,正约束指两个样例属于同一类,负约束指两个样例不属于同一类[122]. 将正约束集记为 $\boldsymbol{ML}$,负约束集记为 $\boldsymbol{CL}$. 主要的半监督聚类方法有基于距离的方法和大间隔方法等,下面分别对这些方法进行描述与分析.

5.1 基于距离的方法

基于距离的方法通过学习样例之间的距离,将距离很近的样例划分到同一簇,将距离很远的样例划分到不同簇. 根据具体的学习方式,基于距离的方法可以分为基于距离度量的方法、基于约束的方法及非线性方法.

5.1.1 基于距离度量的方法

基于距离度量的方法通过训练得到某种自适应距离度量,使其满足类标签或成对约束,然后用学习到的距离度量执行聚类. 常用的距离度量有马氏距离[32]、改进的欧氏距离[20]和 K-L 离差(Kullbac-Leibler Divergence)等.

一种常用的基于距离度量的聚类方法是谱聚类(Spectral Clustering)[123],用加权图的结点表示样例,连接结点的边的权表示两个样例之间的相似度,通过将图的结点分割到不同的部分得到簇. 2002 年,Klein 等人提出第一个半监督距离度量学习聚类方法,根据被正约束影响的相似图中的最短路径学习一种距离度量[20]. 他们通过研究认为正约束在样例上具有二值传递关系,根据这种传递关系可以将正负约束进行传播以反映样例的空间分布信息. 这个方法首先通过求最短路径施加正约束,得到度量矩阵,再利用完全链接层次聚类算法间接施加负约束. Xing 等人[32]定义一种距离度量 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)=d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}=\sqrt{(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{x}_j)}$. 若 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相似,则 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{ML}$;若 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 不相似,则 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{CL}$. 目标函数为凸优化问题

$$\min_{\boldsymbol{A}}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{ML}}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2$$
$$\text{s.t.}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{CL}}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}\geq 1,\ \boldsymbol{A}\geq 0 \tag{10}$$

Ng 等人[124]提出一种谱聚类算法,首先计算邻接矩阵 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$,若 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{ML}$,则 $A_{ij}=A_{ji}=1$;若 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\boldsymbol{CL}$,则 $A_{ij}=A_{ji}=0$,且 $A_{ii}=0$,然后构造矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$,其中 $\boldsymbol{D}$ 为对角阵,对角线元素为 $D_{ij}=\sum_{k=1}^{n}A_{ik}$,接着计算 $\boldsymbol{L}$ 的 k 个最大的特征值所对应的特征向量 $\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_k$,构造 $\boldsymbol{S}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_k]\in\boldsymbol{R}^{n\times k}$. 对 $\boldsymbol{S}$ 的每一行进行单位化处理,得到矩阵 $\boldsymbol{M}$,$M_{ij}=\dfrac{S_{ij}}{\left(\sum_{j=1}^{k}S_{ij}^2\right)^{\frac{1}{2}}}$,然后将 $\boldsymbol{M}$ 的每一行看作 $\boldsymbol{R}^k$ 空间中的一个点,使用 K 均值法或其他方法进行聚类. 若 $\boldsymbol{M}$ 的第 i 行分配到 c_j 类,则将 $\boldsymbol{x}_i$ 也分配到 c_j 类. Kamvar 等人[125]提出的谱聚类算法与 Ng 等人[125]提出的谱聚类算法基本相同,区别只在于他们对 $\boldsymbol{L}$ 的构造方式进行了改进,$\boldsymbol{L}=\dfrac{\boldsymbol{A}+d_{\max}\boldsymbol{I}-\boldsymbol{D}}{d_{\max}}$,$d_{\max}$ 为对角线元素的最大值. Basu 等人[62]提出一种 HMRF 法,将距离度量与约束条件相结合来解决图像分割问题,充分利用了相邻模型之间的相关信息,克服了均值场算法对初始化条件要求非常苛刻的缺点. Ji 和 Xu[126]提出半监督归一化割谱聚类法(Semi-Supervised Spectral Clustering with Normalized Cuts,SS-SNC),利用监督信息用谱方法改变有成对约束信息的聚类距离度量. Zhang 和 Li[127]提出基于密度的约束扩展方法(Density-Based Constraint Expansion Method,DCE),将样例的结构信息引入聚类,在约束条件较少,不足以反映样例分布特点时,可以得到更好的聚类效果,扩展后的约束集可用于各种半监督聚类算法. Wu 等人[31]提出密度敏感的半监督聚类法(Density-Sensitive Semi-Supervised Clustering,DS-SC),引入一种密度敏感的距离度量,能较好地反映聚类假设,并充分利用样例中复杂的内在结构信息,同时与基于图的 SSL 方法相结合,使算法在聚类性能上有了显著的提高. 随后,Wu 等人[128]又提出一种改进的密度敏感的半监督聚类法,得到一种改进的密度敏感的距离度量,可以有效地增大位于不同稠密区域的样例的距离,并缩小位于同一稠密区域内的样例的距离. Luo 和

Wang[129]提出双相似性度量半监督聚类法(Double Similarity Measure Semi-Supervised Clustering, DMSC)，结合主空间和辅助空间来共同影响聚类过程，引入两个近邻度量函数，一个基于辅助空间，采取 K-L 离差，另一个基于主空间. 这个方法不易陷入局部最优，受初始点影响小，可以提高聚类的有效性，但是必须找到合适的 k 值和下降函数. Bijral 等人[130]以基于密度的距离估计为基础，提出一种用图上的最短路径进行计算的简单有效的方法，该方法适用于稠密的全连接图情况，能有效减少运行时间.

5.1.2　基于约束的方法

在许多实际问题中，成对约束信息比类标签信息更普遍，如在语音识别、GPS 导航和图像检索等问题中往往不知道样例的类标签，而知道两个样例是否属于同一类. 基于约束的方法通过将约束条件加入目标函数或修改目标函数使其满足约束条件来进行训练，得到更合适的数据划分.

2000 年 Wagstaff 和 Cardie[122]提出成对约束之后，许多研究人员对利用成对约束的聚类方法进行了研究. 一种基本的半监督聚类方法是 K 均值法[131]，其目标函数为

$$\min_{\boldsymbol{m}_j}\sum_{i=1}^{c}\sum_{j=1}^{n_k}\|\boldsymbol{x}_j-\boldsymbol{m}_i\|^2 \tag{11}$$

式(11)中 n_i 为第 i 个簇的样例数量，$\boldsymbol{m}_i=\dfrac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j$.

Wagstaff 等人[35]将成对约束引入 K 均值聚类法中，提出约束 K 均值法(Constrained K-Means Clustering，COP-K-Means)，调整簇的数量使其满足成对约束，并通过大量说明这个方法可以减少迭代次数和聚类的总运行时间，在给定较好的初始解的情况下能够提高聚类的准确性. Basu 等人[132]在 K 均值法的基础上，引入由少量已标记的样本组成的 seed 集，采用 EM 算法进行优化，提出两种半监督 K 均值法——Seeded-K-Means 和 Constrained-K-Means. Klein 等人[20]通过使两两近似的样例变形将成对约束引入完全连接聚类中. Shental 等人[80]在 EM 算法中考虑成对约束，能够提高混合模型簇和类标签的相似度. Xing 等人[32]将梯度下降法和迭代映射组合作为凸优化问题，利用成对约束学习用于 K 均值的马氏距离，引入学习度量的思想进行聚类，并通过实验用数据说明用成对约束的马氏距离度量能提高 K 均值聚类的准确性. Bar-Hillel 等人[133]以及 Chang 和 Yeung[134]也进行了类似的实验. Basu 等人[62]和 Lange 等人[135]将成对约束信息加入 K 均值聚类法中，在违反概率簇分配的约束上增加了惩罚项. Davidson 和 Ravi[136]提出一种 K 均值式算法，使约束向量量化误差最小，但是这种方法在每次迭代中并不一定满足所有约束. Cohn 等人[137]和 Jain 等人[138]提出允许迭代地提供一些数据上的反馈的聚类方法. Bilenko 等人[139]将基于约束的方法和距离度量学习相结合，为每个簇学习一个单独的距离度量. Gao 等人[140]综合考虑有类标签的样本的背景信息与无类标签的样例的特征，将问题表示为有约束的优化问题，并提出两种学习方法解决硬聚类和模糊聚类问题. Tang 等人[141]提出了一种改进的高维数据半监督聚类的方法，利用约束指导特征映射，而不是用距离度量. Lu 等人[142]和 Nelson 等人[143]提出半监督聚类概率方法，通过贝叶斯先验信息将成对约束与聚类算法合并. Chen 等人[144]提出用于文本聚类的半监督非负矩阵分解框架(Semi-Supervised Nonnegative Matrix Factorization，SS-NMF)，通过迭代算法执行文本间相似矩阵的对称三次因式分解，来推断出文本簇. SS-NMF 可以看做是现有的半监督聚类方法的一般框架. Yin 等人[145]提出一种基于成对约束的判别式半监督聚类分析法(Discriminative Semi-Supervised Clustering Analysis with Pairwise Constraints，DSCA)，首先利用正负约束产生投影矩阵，在投影空间中对样例聚类生成簇标号，然后利用 LDA 选择子空间，接着用基于成对约束的 K 均值法对子空间中的样例聚类. 这个方法有效地利用了监督信息进行聚类，降低了基于约束的 SSL 聚类算法的计算复杂度. Xia[146]将 Tuy 的凹割平面方法用于半监督聚类，并给出了半监督聚类局部最优解的一些特性. Lu 和 Ip[147]提出一种详细有效的用成对约束进行谱聚类的约束传播方法，将约束传播问题分解为一组独立的约束传播子问题，用基于 KNN 的 SSL 方法在平方时间内解决这些子问题，并在真实数据集上进行实验，结果说明了该方法的优越性.

5.1.3　非线性方法

非线性方法通过核函数间接将样例映射到特征空间中，在原始空间的非线性边界的帮助下完成聚类[148].

Kulis 等人[149]将 Basu 等人[62]提出的方法扩展为基于核的半监督聚类，提出半监督核 K 均值法(Semi-Supervised Kernel K-Means，SS-KK)，不是增加违反成对约束的惩罚项，用有奖励和惩罚约束

的加权核 K 均值法来变换聚类距离度量，执行向量形式或图形式的样例的半监督聚类. Yan 等人[150]提出一种半监督聚类自适应核学习法(Adaptive Kernel Learning Method for Semi-Supervised Clustering，ASSKKM)，将 Basu[62]提出的方法中的目标函数核化. Chang 和 Yeung[134]提出一种用正约束寻找局部线性度量的方法. 随后，Yeung 和 Chang[151]指出之前提出的方法的目标函数有许多局部最优解，并且在训练过程中并不能很好的保持拓扑结构，提出两种基于核的度量学习方法. Tsang 等人[152]提出核相关组分分析方法，利用合适的核函数将 Bar-Hillel[133]的方法泛化到非线性问题. Chen 等人[153]提出非线性自适应距离度量学习，首先用核函数将样例映射到高维空间，然后应用线性映射找到低维流形，最后在映射得到的低维空间中完成聚类. Chang 等人[154]提出一种度量自适应方法，迭代地调整样例的位置，使相似的点距离越来越近，相异的点距离越来越远. 但是这个方法缺乏显式变换映射，因此不能直接处理变换空间中新加入的样例. Xiang 等人[155]提出用追踪比优化问题更合适作为的目标函数，并提出一个很好的启发式搜索方法用于解决这个问题.

5.2 大间隔方法

最近，将大间隔方法用于聚类的研究成为 SSL 研究热点之一，已提出了 MMC[155-156]、IterSVR[157]、GMMC[158]、CPMMC[159]、CPM3C[160]和 MKC[161]等算法，并且与当前主流聚类算法如 K 均值聚类和归一化割谱聚类法相比有较大的优势. 大间隔方法用于聚类时，求解原问题 Γ_2 为

$$\min_{w,\xi,b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2+C\xi_i \tag{12a}$$

$$\text{s.t.} \ \ y_i\cdot(\langle \boldsymbol{x}_i,\boldsymbol{w}\rangle+b)\geq 1-\xi_i \tag{12b}$$

$$\xi_i\geq 0 \tag{12c}$$

$$-l\leq\sum_{i=1}^{n} y_i\leq l \tag{12d}$$

$$y_i\in\{c_1,c_2,\cdots,c_C\},\ i=1,\cdots,n \tag{12e}$$

式(12)中 $C>0$，$l\geq 0$ 为待定的参数，式(12d)为类平衡约束，避免出现平凡解. Γ_2 中一部分样例 $\boldsymbol{x}_i$ 的类标签 y_i 的值是不知道的(一般是全不知道)，目标是不但求解出 $\boldsymbol{w}$ 和 b，还要求解类标签 y_i 的值，由于式(12e)，因此这是一个整数规划问题. 以两类为例，由于约束 $y_i\in\{\pm 1\}\Leftrightarrow y_i^2-1=0$ 是非凸的，整个问题为非凸整数规划问题. Xu 提出的 MMC 算法先求出聚类大间隔方法对偶问题

$$\max_{\lambda} \ 2\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{e}-\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{K}\circ\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}})\boldsymbol{\lambda}$$

$$\text{s.t.} \ \ 0\leq\boldsymbol{\lambda}\leq C\boldsymbol{e}$$

$$\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{y}=0$$

$$-l\leq\sum_{i=1}^{n} y_i\leq l$$

$$y_i\in\{\pm 1\},\ i=1,\cdots,n \tag{13}$$

式(13)中 $\boldsymbol{K}\in\boldsymbol{R}^{n\times n}$ 为输入为 $[\boldsymbol{K}]_{i,j}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 的核矩阵，$\boldsymbol{y}=(y_1,\cdots,y_n)^{\mathrm{T}}$，$\boldsymbol{e}=(1,\cdots,1)^{\mathrm{T}}$，$\boldsymbol{A}\circ\boldsymbol{B}$ 表示矩阵元素相乘运算. 再次引入等式约束对偶变量 $\bar{b}$ 和不等式约束对偶变量 $\boldsymbol{\mu}$，$\boldsymbol{\upsilon}$，得

$$\min_{\boldsymbol{y},\delta,\boldsymbol{\mu},\boldsymbol{\upsilon},\bar{b}} \ \delta$$

$$\text{s.t.} \ \begin{bmatrix} (\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}})\circ\boldsymbol{K} & \boldsymbol{e}+\boldsymbol{\mu}-\boldsymbol{\upsilon}-\bar{b}\boldsymbol{y} \\ (\boldsymbol{e}+\boldsymbol{\mu}-\boldsymbol{\upsilon}-\bar{b}\boldsymbol{y})^{\mathrm{T}} & \delta-2C\boldsymbol{\upsilon}^{\mathrm{T}}\boldsymbol{e} \end{bmatrix}\geq 0$$

$$\boldsymbol{\mu}\geq 0,\ \boldsymbol{\upsilon}\geq 0$$

$$-l\leq\sum_{i=1}^{n} y_i\leq l$$

$$y_i\in\{\pm 1\},\ i=1,\cdots,n \tag{14}$$

Xu 提出用 $n\times n$ 实值正定矩阵 $\boldsymbol{M}\geq 0$ 代替 $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\in\{\pm 1\}^{n\times n}$，此时平衡约束变为 $-l\leq\sum_{i=1}^{n} y_i\leq l\Leftrightarrow -l\boldsymbol{e}\leq\boldsymbol{M}\boldsymbol{e}\leq l\boldsymbol{e}$，为避免正定约束中 $\bar{b}$ 和 $\boldsymbol{y}$ 出现共线性，放松使得 $\bar{b}=0$，等于假定分类超平面通过原点. 经过以上放松，问题变为以下半定规划问题

$$\min_{\boldsymbol{M},\delta,\boldsymbol{\mu},\boldsymbol{\upsilon},\bar{b}} \ \delta$$

$$\text{s.t.} \ \begin{bmatrix} \boldsymbol{M}\circ\boldsymbol{K} & \boldsymbol{e}+\boldsymbol{\mu}-\boldsymbol{\upsilon} \\ (\boldsymbol{e}+\boldsymbol{\mu}-\boldsymbol{\upsilon})^{\mathrm{T}} & \boldsymbol{\delta}-2C\boldsymbol{\upsilon}^{\mathrm{T}}\boldsymbol{e} \end{bmatrix}\geq 0$$

$$\boldsymbol{\mu}\geq 0,\ \boldsymbol{\upsilon}\geq 0$$

$$\boldsymbol{M}\geq 0,\ \mathrm{diag}(\boldsymbol{M})=\boldsymbol{e}$$

$$-l\boldsymbol{e}\leq\boldsymbol{M}\boldsymbol{e}\leq l\boldsymbol{e} \tag{15}$$

式(15)的半定规划问题算法复杂性仍然很高，不适合实际应用，Zhang 等人[162]提出把铰链损失换为拉普拉斯损失或平方损失，直接求解非凸优化问题的迭代 SVR 算法，为了解决 MMC 算法只能求解中等规模问题，Li 等人[163]使用多类标签组合核学习方法，提出 LG-MMC(Label-Generating MMC)算法. 为解决聚类精度不稳定问题，Hu 等人[164]提出引入成对约束，用约束凹凸过程(Constrained Concave-Convex Procedure，CCCP)迭代求解一系列二次规划问题.

针对 $\boldsymbol{M}\in\boldsymbol{R}^{n\times n}$ 中要求解的参数数量是样例数量的二次方，无法处理大规模数据，分类超平面通过原点，不适合聚类类不平衡数据等问题，赵兵与张长水提出割平面算法，把原问题的 n 个约束条件依次加入目标函数求近似解，从而逐次逼近原问题的

解[160]. Valizadegan[158]提出放松非凸问题为凸问题时并不增加求解变量，求解的问题与样例个数成线性关系，把核矩阵逆替换为归一化图拉普拉斯算子，从而具有无监督核学习能力，能够同时自动确定适当的核矩阵和簇成员关系的算法. Xu[157]和 Zhao 等人[161]分别给出了大间隔方法聚类的多类版本并扩展到多核情形.

另外，Gieseke 等人[165]提出在核空间中求解正则化最小二乘损失函数

$$\min_{f\in \boldsymbol{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \|\boldsymbol{f}\|_{\boldsymbol{H}}^2 \qquad (16)$$

由表示理论知 $\|\boldsymbol{f}\|_{\boldsymbol{H}}^2 = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{c}$，$\boldsymbol{K}\in \boldsymbol{R}^{n\times n}$ 为输入为 $[\boldsymbol{K}]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 的核矩阵，Gieseke 等人提出用进化算法求解以下问题

$$\min_{\boldsymbol{c}\in \boldsymbol{R}^n} J(\boldsymbol{y}, \boldsymbol{c}) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{c})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{c}) + \lambda \boldsymbol{c}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{c} \qquad (17)$$

Zhang 等人[166-167]提出多样例最大间隔聚类法(Maximum Margin Multiple Instance Clustering, M3IC)，使用割平面方法和 CCCP 组合求解最优化问题.

6　半监督降维方法

在许多实际问题中，如数字图像、金融时间序列、基因表达微序列等，常常会遇到高维数据，直接处理这些高维数据容易出现维数灾难(Curse of Dimensionality)问题，这就需要对数据进行降维. 降维通常被看做改进分类、回归、聚类等任务的工具，通过对数据进行降维处理，可以改进随后的训练过程的学习性能. 半监督降维的目的是在大量的无类标签的样例 $\boldsymbol{U} = \{\boldsymbol{x}_{l+1}, \cdots, \boldsymbol{x}_{l+u}\}$ 中引入少量的有类标签的样本 $\boldsymbol{L} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_l, y_l)\}$，利用监督信息找到高维数据 $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ 的低维结构表示 $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$，$\boldsymbol{z}_i \in \boldsymbol{R}^d$，$d < m$，$i = 1, \cdots, n$，$n = l + u$，与此同时保持数据的内在固有信息(Intrinsic Information)，即保持原始数据及成对约束信息 $\boldsymbol{ML}$ 和 $\boldsymbol{CL}$ 的结构，也就是说，$\boldsymbol{ML}$ 中的样例最终应该距离很近，$\boldsymbol{CL}$ 中的样例最终应该互相远离. 其中样例 $\boldsymbol{x}_i \in \boldsymbol{R}^m$，样例的低维表示 $\boldsymbol{z}_i \in \boldsymbol{R}^d$，$d < m$，类标签 $y_i \in \{c_1, c_2, \cdots, c_C\}$，$i = 1, \cdots, l, \cdots, l+u$，训练样例数量为 $n = l + u$. 当降维方法为线性时，训练过程为学习一个映射矩阵 $\boldsymbol{W} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_d\}$，$\boldsymbol{w}_j \in \boldsymbol{R}^m$，$j = 1, \cdots, d$，使得 $\boldsymbol{Z} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$. 当降维方法为非线性时，训练过程不需要学习这个映射矩阵 $\boldsymbol{W}$，而是直接从原始数据中学习得到数据的低维表示 $\boldsymbol{Z}$. SSL 中利用的监督信息既可以是样例的类标签，也可以是成对约束信息，还可以是其他形式的监督信息. 主要的半监督降维方法有基于类标签的方法、基于成对约束的方法及其他方法等，下面分别对这几种方法进行描述与分析.

6.1　基于类标签的方法

当前应用广泛的降维方法之一是 LDA，寻找一个映射矩阵，使降维后的同类数据之间的距离尽量减小，不同类数据距离尽量增大. Baudat 和 Anouar[168]用核技巧将 LDA 扩展到非线性形式，提出广义判别分析法(Generalized Discriminant Analysis, GDA). Yan 等人[169]和 Sugiyama[170]将 FDA 扩展为间隔费希尔判别分析法(Margin Fisher Discriminative Analysis, MFA)和局部费希尔判别分析法(Local Fisher Discriminative Analysis, LFDA). Costa 和 Hero[171]为了提取与分类任务相关的低维特征，将具有流形结构的样例的类标签引入低维数据嵌入结构中调整拉普拉斯法，提出非线性降维方法——分类约束降维法(Classification Constrained Dimensionality Reduction, CCDR)，将每类中所有结点的中心点作为新的结点加入近邻图中，并用权为 1 的边连接同一类中的结点和它们的中心点. Yu 等人[33]在概率 PCA 模型的基础上加入了类标签信息，提出半监督概率 PCA 模型(Semi-Supervised Probabilistic Principal Component Analysis, S^2 PPCA)[93]，将类标签信息引入映射过程中，能很好地利用所有的信息定义映射，并得到解决这个问题的有效的 EM 学习算法，而且从理论上分析了这个方法的性能. 这个方法能用于处理多输出问题，不仅表现出优越的性能，而且具有较好的可扩展性. Chen 等人[172]将 LDA 重写成最小平方的形式，通过加入拉普拉斯正则化项将该模型转化为最小平方正则化问题. Cai 等人[173]在传统的 LDA 中引入流形正则化项，提出半监督判别分析法(Semi-Supervised Discriminative Analysis, SDA)，在最大化类间离散度的同时可以保持数据的局部结构信息. 与 LDA 一样，也可以将 SDA 的目标函数转化为一个广义特征分解问题. Zhang 和 Yeung[174]提出一种新的半监督降维方法——半监督判别分析法(Semi-Supervised Discriminant Analysis, SSDA)，也是使用正则化项来保持数据的流形结构，但与 SDA 不同的是这个方法使用一种基于路径的鲁棒相似度量来构造近邻图，并用得到的相似性使不同类间的可分离性最大化，并提出用于半监督非线性降维的核化方法. 这个方法能够利用

数据的全局结构,并且在定义近邻关系中对噪声具有鲁棒性. Song 等人[175]提出半监督降维方法框架,在原始 LDA 中加入一个正则化项,这个正则化项以有类标签的样本和无类标签的样例提供的先验信息为基础,可以用图拉普拉斯构造,并且这个方法可以用核技巧进行核化. 这个方法能够发现每类的子流形结构,并将判别子流形嵌入到低维全局坐标系中,能够很好地解决半监督归纳问题,并且 PCA、LDA 等及其核化方法可以看做这个统一框架下的特殊情况.

Zhang 和 Yeung[176]提出一种新的 SSDA——约束凹凸过程半监督判别分析法(Semi-Supervised Discriminant Analysis Algorithm Constrained Concave-Convex Procedure,$SSDA_{CCCP}$),利用无类标签的样例最大化 LDA 的优化准则,并用约束凹凸过程解决优化问题. 这个方法克服了当有类标签的样本数量有限时,LDA 受到严重限制的情况. 随后提出 $SSDA_{CCCP}$的变形 M-$SSDA_{CCCP}$,依赖流形假设来利用无类标签的样例,综合了 TSVM 和基于图的半监督学习方法的特点.

Sugiyama 等人[177]将局部费希尔判别分析法与 PCA 相结合,提出半监督局部费希尔判别分析法(Semi-Supervised Local Fisher Discriminant Analysis,SELF),在保留局部费希尔线性判别分析法优点的同时,可以保持无类标签的样例的全局结构,基于特征分解能够得到全局优化的解析解.

Chatpatanasiri 和 Kijsirikul[178]从流形学习的角度提出广义的半监督降维框架,在该框架下,可以很容易地把传统的费希尔判别分析法扩展为半监督的形式,即使在每类的训练数据形成复杂非线性流形上的单独簇的情况下,基于这个框架的方法也能够找到很好的低维子空间. 最近出现的基于谱分解的半监督框架可以看做是这个框架的特殊情况. 还提出一个新的非线性化框架——核主成分分析(Kernel Principal Component Analysis,KPCA)技巧框架,并将其用于半监督场景.

6.2 基于成对约束的方法

Tang 和 Zhong[179]将成对约束引入到半监督降维过程中,目标函数加上低维空间中满足负约束的点之间的距离,并减去所有满足正约束的点之间的距离和,然后使这个目标函数最大. 这个方法能同时利用正负约束信息,但是没有考虑大量的无类标签的样例,并且构造的所有约束都占同等重要的地位,这可能会产生不好的结果.

Hoi 等人[180]提出利用成对约束进行度量学习的判别成分分析法(Discriminative Component Analysis,DCA),像 LDA 一样寻找一个映射矩阵,使降维后的满足正约束的数据之间的距离尽量减小,满足负约束的数据尽量远离,并提出核化的 DCA 法.

Bar-Hillel 等人[133]提出可以有效解决降维问题的约束费希尔线性判别法(Constrained Fisher Linear Discrimination,CFLD),利用成对约束信息对数据进行预处理,随后产生改进的分类和聚类,并说明需要多少成对约束信息修改数据表示才能得到好的学习结果,从而解决学习度量问题. CFLD 是相关组分分析法(Relevant Component Analysis,RCA)的一个中间步骤. RCA 搜索并删除数据在全局内不需要的变量,将用于数据表示的特征空间进行改变,用全局线性变换给相关维分配大的权,给不相关维分配小的权,是一种简单有效的学习马氏距离的方法. 但是 CFLD 只能解决正约束问题.

Zhang 等人[64]提出半监督降维法(Semi-Supervised Dimensionality Reduction,SSDR),这个方法并不像 CFLD 那样利用约束信息构造散布矩阵,而是使数据满足约束来直接指导降维过程,同时能够像 PCA 一样保持数据的内部结构信息,在一些特定的拉普拉斯矩阵的本征问题上有近似解,可以从更直观有效的角度同时利用无类标签的样例和正负约束来指导降维过程. 但是这个方法只能保持全局协方差结构,而没有保持数据的局部结构.

Cevikalp 等人[181]在局部保持映射法中引入约束信息,提出约束局部保持映射法(Constrained Locality Preserving Projections,CLPP),首先构造数据的加权近邻图,然后利用约束信息修改近邻关系和加权阵,增大满足正约束的数据之间的权值,减小满足负约束的数据之间的权值,同时修改与满足约束的结点直接相连的结点的权值,对约束信息进行传播,最后通过解最小特征值得到损失函数对应的最优映射矩阵. 不同于 SSDR,CLPP 在降维过程中保持数据的局部结构信息. 但是,没有找到确定构造近邻图所需优化参数的可靠方法,同时通过实验发现,参数的一个小扰动就会产生截然不同的训练结果.

Wei 和 Peng[182]提出基于近邻保持的半监督降维法(Neighbourhood Preserving based Semi-Supervised Dimensionality Reduction,NPSSDR). 与 CLPP 不同,NPSSDR 不需要构造数据的邻接矩

阵，而是通过添加正则化项的方法来实现，组合给定数据集的流形结构和近邻保持，然后最小化重构误差. NPSSDR 在利用约束信息指导降维过程的同时，还能保持正负约束与嵌入低维子空间中的数据的局部结构，有近似形式的解，因此可以很容易的用于解决未知测试数据的情况.

Peng 和 Zhang[183]在传统的典型相关分析法中引入成对约束信息，提出半监督典型相关分析法(Semi-Supervised Canonical Correlation Analysis，Semi-CCA)，同时利用无类标签的样例和成对约束信息指导降维过程，并验证了两者的相对重要性. Baghshah 和 Shouraki[184]将正负约束引入拓扑结构中，提出解决非线性变换的基于核的距离学习方法，将 NPSSDR 用于度量学习，并用二分搜索法来优化求解过程.

大多数降维方法都是由目标函数驱动的，可能只有一部分满足或者甚至不满足性能需求. 因此，Davidson[185]提出通过线性映射的图驱动约束降维法(Graph-Driven Constrained Dimension Reduction via Linear Projection，GCDR-LP)，将编码成图领域的专业知识引入降维过程中，使其成为更普遍的特征值问题. 这个方法引入了约束图，可以更加灵活地利用占不同重要性的约束建立模型，并强调了局部几何结构的重要性.

6.3　其他方法

还有许多基于其他形式监督信息的降维方法.

6.3.1　基于流形嵌入的方法

Ham 等人[186]给出部分样例的嵌入结果作为监督信息，用核方法解释了流形上的降维方法，如 ISOMAP、LLE、图拉普拉斯特征映射(Graph Laplacian Eigenmap，GLE)等，这些方法都是利用局部近邻信息构造流形并将其全局映射到低维空间，并说明这些方法都可以被描述为特定结构的格莱姆矩阵上的核 PAC 法. Yang 等人[187]将特定数据的流形坐标形式的先验信息用于降维过程，将 LLE、ISOMAP、局部切空间排列法(Local Tangent Space Alignment，LTSA)等基本的非线性降维方法扩展到半监督的形式，并说明利用哪些先验信息能更好地提高解的性能. 但是，获取数据的流形坐标比获得数据的成对约束要困难得多，因此基于流形嵌入的方法并不常用.

6.3.2　基于样例相关性的方法

Memisevic 和 Hinton[188]提出的多关系嵌入法(Multiple Relational Embedding，MRE)可以综合利用多种相似性关系学习数据的低维嵌入，通过分别给基本隐空间的维数再加权选择不同的相似性表示，并用一些简单的例子证明这个方法的有效性. Yu 和 Tian[189]提出语义子空间映射法(Semantic Subspace Projection，SSP)，在嵌入到高维空间中的非线性图像子空间上建立图像模型，充分利用语义和图像之间的相似或相异信息，通过增量学习有效地结合相关反馈，并通过理论分析证明 LDA 可以表示为这个方法的特例. Liu 等人[190]通过询问样例相关或不相关来对其进行标记，提出相关聚合映射法(Relevance Aggregation Projections，RAP)，用半监督的方式学习有效的子空间映射. 这个方法根据反馈给定样例之间的相关性或不相关性，生成一个子空间，在这个子空间中，相关的样例被聚合到一个点，不相关的样例被大的间隔分开[115].

7　半监督学习理论分析

大量研究证实 SSL 能够利用无类标签的样例提高学习算法预测的准确性和预测算法的速度. 因此，许多研究人员对 SSL 进行了理论分析.

Castelli 和 Cover[76]说明在无类标签的样例数量无限的情况下，可识别的混合模型的误差率以指数收敛到贝叶斯风险. Ratsaby 和 Venkatesh[191]得到服从两个 GMM 分布的 SSL 的 PAC 结构的学习率. Lafferty 和 Wasserman[192]从极小极大理论提出半监督回归方法的理论分析，并说明这些方法在合适的假设下可以产生更好的性能. Ben-David 等人[55]在生成式场景和诊断场景中对 SSL 模型和标准的监督 PAC 模型进行了比较，得到的结论是只有在满足类标签分布的强假设的条件下，SSL 才能够提供合适的样本大小来保证其性能优于 SL 的性能，并提出当前 SSL 方法中普遍存在的一个问题，目前没有方法来实际验证样本-类标签之间的数据结构关系是否成立. Wang 和 Zhou 在文献[74]中给出了充分必要性定理，为协同训练方法的成功提供了充分必要条件，该定理深刻揭示出，只要学习机之间存在显著差异并且学习机不太差时，就可以通过协同训练提高学习的性能，若希望用协同训练得到理想的结果，则必须将每个无类标签样例在某个联合假设空间中与有类标签样本相连. 该结果被 Mitchell 等人认为是 SSL 领域的重要进展，并在 ICML2010 的报告上进行介绍. Urner 等人[193]研究如何用无类标签的样例帮助构造更快速的分类器，

提出一种 SSL 算法架构,利用一组预定的快速分类器中的无类标签的样例来学习分类器,并分析在何种条件下无类标签的样例能够有效提高这种方法的性能. Darnstadt 和 Simon[194] 分析了 SSL 的样本复杂度,说明的确存在巧妙的 PAC 学习机,利用有限数量的有类标签样本在不用关于基本领域分布的任何先验信息的情况下完成学习目标,从信息论的角度说明了完整的领域分布先验信息并不能有效减少完成学习目标所需要的有类标签样本的数量. Dillon 等人[195] 基于随机组合似然的拓展形式,定量分析了生成式 SSL 的渐近准确性,通过提供与不同标记方法相关的度量值的可替代架构,补充分析了分布自由的数据训练性能,解决了标记多少数据及以何种方式标记数据这些基本问题.

与此同时,也有许多研究表明由于引入了无类标签的样例而造成学习算法性能下降,由此,许多研究人员对 SSL 方法进行改进.

Balcan 和 Blum[57] 提出同时利用有类标签和无类标签的样例的 PAC 模型,引入相容函数使满足假设的分类机能够很好的服从无类标签的样例的分布,通过得到的样本复杂度得出结论,在训练误差为零且相容性很高的条件下,只用很少的有类标签的样本就能得到好的假设. Nigam 等人[28] 在文本分类任务中,用具有固定结构和大量特征的朴素贝叶斯分类器讨论了无类标签的样例降低分类性能的情况,提出一些技术来抑制性能的降低,他们提出分类性能的降低可能是由于特征空间中的自然簇和真实类标签之间的失配. 与其他 SSL 方法类似,研究发现利用无类标签的样例也会降低 S^3VMs 的学习性能. 为了解决这个问题,Li 和 Zhou[86] 提出了可靠 S^3VMs(Safe Semi-Supervised Support Vector Machines,S^4VM)方法,不根据目标值选择最优的低密度分离机,而是考虑所有低密度分离机,特别针对直推场景,优化最坏情况下的无类标签的样例的类标签分配来构造 S^4VMs. Li 和 Zhou[196] 提出相比于利用所有无类标签的样例而造成 S^3VMs 性能下降,应当选择更能起到帮助作用的无类标签样例,而避免使用高风险的无类标签样例,因此提出了 S3VM-us(Semi-Supervised Support Vector Machine with Unlabeled Instances Selection)方法,用层次聚类来选择无类标签的样例,并用实验结果说明 S3VM-us 方法比现有的 S^3VMs 在引起学习性能下降方面具有很大改进.

有许多研究人员从理论上分析给出了 SSL 算法的误差界.

Leskes[27] 给出了协同训练的广义误差界,并提出当协同训练的不同学习机在给定相同的训练数据集上得到的结果一致,并且训练结果的误差较低时,广义误差界更紧. El-Yaniv 和 Pechyony[197] 在更广的学习场景中,在没有假定来自未知分布的样本独立同分布的情况下,给出一致稳定学习算法的边界. Kaariainen[198] 讨论不用类标签条件分布先验假设的 SSL 方法的广义误差界. Derbeko 等人[199] 明确地给出了直推学习的误差界,得到了一个在直推场景下构造误差界的方法,并将这种方法用于压缩和聚类场景得到误差界. 假定 p 是依赖全部样例的类上的先验分布,给定 $\delta\in(0,1)$,训练误差 $R(\boldsymbol{X}_l)=\sum_{i=1}^{l}V(y_i,f(\boldsymbol{x}_i))$,直推风险 $R(\boldsymbol{X}_u)=\sum_{i=l+1}^{l+u}V(\hat{y}_i,f(\boldsymbol{x}_i))$. 以至少 $1-\delta$ 的概率在全部样例上选择训练样本,得到直推场景的广义误差界为

$$R(\boldsymbol{X}_u)\leqslant R(\boldsymbol{X}_l)+\sqrt{\left(\frac{2R(\boldsymbol{X}_l)(l+u)}{u}\right)\frac{\log\frac{1}{p}+\ln\frac{l}{\delta}+7\log(l+u+1)}{l-1}}+\frac{2\left(\log\frac{1}{p}+\ln\frac{l}{\delta}+7\log(l+u+1)\right)}{l-1} \tag{18}$$

当 $R(\boldsymbol{X}_l)\to 0$ 时,式(18)中的平方根消失,得到更快的学习率. 可以根据式(18)使用训练样例 $\boldsymbol{X}_{l+u}$ 来选择先验分布 p,不过该方法在实际应用中还没有得到广泛应用. 对于来自固定数据集 $\boldsymbol{X}_{l+u}$ 的随机选择的大小为 l 的子样例,式(18)以至少 $1-\delta$ 的概率成立.

利用多重先验分布 $p_1,\cdots,p_k$,扩展到 PAC 贝叶斯结构,以至少 $1-\delta$ 的概率在来自全部样例的随机选择的大小为 l 的子样例上选择训练样本,得到实际直推场景的紧致误差界为

$$R(\boldsymbol{X}_u)\leqslant R(\boldsymbol{X}_l)+\sqrt{\left(\frac{2R(\boldsymbol{X}_l)(l+u)}{u}\right)\frac{\min\limits_{1\leqslant i\leqslant k}\log\frac{1}{p_i}+\ln\frac{kl}{\delta}+7\log(l+u+1)}{l-1}}+\frac{2\left(\min\limits_{1\leqslant i\leqslant k}\log\frac{1}{p_i}+\ln\frac{kl}{\delta}+7\log(l+u+1)\right)}{l-1} \tag{19}$$

相比用一个先验分布 p,根据式(19)可以用 k 个先验分布 $p_1,\cdots,p_k$,并从 k 个相应的 PAC 贝叶斯边界中选择最优的那个.

一个确定性算法对每个二分法最多只能确定一个假设 $h\in\boldsymbol{H}$. 用确定性学习算法基于大小为 s 的小

压缩集产生假设，以至少 $1-\delta$ 的概率对所有 $h\in \boldsymbol{H}$ 成立，得到压缩算法的紧致误差界为

$$R(\boldsymbol{X}_u)\leq R(\boldsymbol{X}_l)+\sqrt{\left(\frac{2R(\boldsymbol{X}_l)(l+u)}{u}\right)\frac{s\log\frac{2e(l+u)}{s}+\ln\frac{l^2}{\delta}+7\log(l+u+1)}{l-1}}+\frac{2\left(s\log\frac{2e(l+u)}{s}+\ln\frac{l^2}{\delta}+7\log(l+u+1)\right)}{l-1} \tag{20}$$

训练分类器时很容易计算得出式(20)的误差界，并且当压缩集非常小时，就可以得到紧致误差界.

用聚类算法将训练样例 $\boldsymbol{X}_{l+u}$ 聚类到 $2,3,\cdots,c$ 个聚类簇中，其中 $c\leq l$，产生将 $\boldsymbol{X}_{l+u}$ 划分 $\tau=2,3,\cdots,c$ 个聚类簇中的分割集合，以至少 $1-\delta$ 的概率对所有 τ 成立，得到聚类算法的误差界为

$$R(\boldsymbol{X}_u)\leq R(\boldsymbol{X}_l)+\sqrt{\left(\frac{2R(\boldsymbol{X}_l)(l+u)}{u}\right)\frac{\tau+\ln\frac{lc}{\delta}+7\log(l+u+1)}{l-1}}+\frac{2\left(\tau+\ln\frac{lc}{\delta}+7\log(l+u+1)\right)}{l-1} \tag{21}$$

当聚类算法用少量的聚类簇得到训练样例的数据结构时，用式(21)可以得到紧致误差界.

同时，有许多研究人员从理论上分析给出了 SSL 的样本复杂度.

Gentile 和 Helmbold[200] 研究在噪声场景中，固定分布的 d 个间隔的并集的样本复杂度下界为 $\Omega\left(\frac{2d\log(1/\Delta)}{\Delta(1-2\eta)^2}\right)$，其中 Δ 是学习算法以高概率保证到目标的距离，η 是错误类标签出现的概率. Ben-David 等人[55] 在生成式场景和诊断场景中，比较 SSL 模型和监督 PAC 模型的样本复杂度. 置信度 $\delta>0$，准确度 $\varepsilon>0$，在生成式场景中，SL 的样本复杂度上界是 $\frac{\ln(1/\delta)}{\varepsilon}$，SSL 的样本复杂度上界是 $\frac{\ln(1/\delta)}{2\varepsilon}+O\left(\frac{1}{\varepsilon}\right)$，下界是 $\frac{\ln(1/\delta)}{2.01\varepsilon}-O\left(\frac{1}{\varepsilon}\right)$. 忽略低阶项，对绝对连续的无类标签数据分布，可以得到在生成式场景中 SL 的样本复杂度至多是 SSL 的 2 倍. 在诊断场景中，Ben-David 等人明确地构造噪声场景，并利用信息论方法，得到 SSL 在(0,1)均匀分布上学习阈值的样本复杂度为 $\Theta\left(\frac{\ln(1/\delta)}{\varepsilon^2}\right)$，学习至多 d 个间隔的并集的样本复杂度为 $\Theta\left(\frac{2d+\ln(1/\delta)}{\varepsilon^2}\right)$. 该结果说明在诊断场景中，SSL 在样本复杂度方面的改进不超过 SL 的常数倍.

8　未来研究方向

经过大量研究人员的长期努力，SSL 领域的研究已取得了一定发展，提出了不少 SSL 方法，同时已将 SSL 应用于许多实际领域. 但目前这个领域的研究仍存在许多有待进一步解决的问题，我们认为未来的研究方向包括以下一些内容.

8.1　理论分析

目前对 SSL 的理论分析还不够深入. 在类标签错误或成对约束不正确时学习方法的性能如何改变，选择不同的正约束和负约束的比例会对降维的性能造成什么影响，除了通常采用的分类精度和运算速度之外，还有没有其他更合适的评价指标，对学习性能起到改进作用的是准确的最优化求解算法，还是使用的学习模型中的数据表示和学习方法，最优解对学习结果的影响有多大，未来还需要进一步探讨这些问题.

8.2　抗干扰性与可靠性

当前大部分 SSL 利用的数据是无噪声干扰的数据，而且依赖的基本假设没有充分考虑噪声干扰下无类标签数据分布的不确定性以及复杂性，但是在实际应用中通常难以得到无噪声数据. 未来需要研究如何根据实际问题选择合适的 SSL 方法，更好地利用无类标签的样例帮助提高学习的准确性和快速性，并减小大量无类标签数据引起的计算复杂性，可以考虑引入鲁棒统计理论解决该抗噪声干扰问题. 此外，大量实验研究证明当模型假设正确时，无类标签的样例能够帮助改进学习性能；而在错误的模型假设上，SSL 不仅不会对学习性能起到改进作用，甚至会产生错误，恶化学习性能. 如何验证做出的模型假设是否正确，选择哪种 SSL 方法能够更合适地帮助提高学习性能，除了已有的假设之外，还可以在无类标签的样例上进行哪些假设，新的假设是否会产生新的算法，SSL 能否有效用于大型的无类标签的数据，这些问题还有待未来研究. 此外，导致 SSL 性能下降的原因除了模型假设不符合实际情况外，还有学习过程中标记无类标签的样例累积的噪声，是否还有其他原因使无类标签的样例造成学习性能的下降，也是未来需要进一步研究的问题.

8.3　训练样例与参数的选取

通常训练数据是随机选取的，即有类标签的样例和无类标签的样例独立同分布，但是在实际应用

中,无类标签的样例可能来自与有类标签的样例分布不同或未知的场景,并且有可能带有噪声.未来的研究需要找到一个好的方法将 SSL 和主动学习相结合,选取有利于学习模型的训练样例,并确定 SSL 能够有效进行所需要的有类标签的样本数量的下界.此外,许多研究人员将 SL 和 UL 算法扩展用于 SSL,但是许多这些算法是根据先验信息得到训练数据集的参数,并利用这些参数改进算法在 SSL 中的性能.目前都是人工选取一种 SSL 方法,并设定学习参数,保证 SSL 的性能优于 SL 和 UL,但是当选取的 SSL 方法与学习任务不匹配或者参数的设定不合适时,会造成 SSL 的性能比 SL 或 UL 更差.如何自动根据学习任务选取合适的 SSL 方法并准确得到参数是未来 SSL 需要深入研究的内容,可以考虑用全贝叶斯学习理论解决.

8.4 优化求解

从各种 SSL 算法的实现过程可以看出,SSL 问题大多为非凸、非平滑问题,或整数规划和组合优化问题,存在多个局部最优解,例如求解 SSL 产生式方法目标函数的 EM 算法只能得到局部极大值.目前主要采用各种放松方法把目标函数近似转化为凸或连续最优化问题,不易得到全局最优解,算法的时空复杂性很高,问题的求解依赖于最优化理论的突破,未来需要研究新的算法求解全局最优解.

8.5 研究拓展

SSL 从产生以来,主要用于实验室中处理人工合成数据,未来的研究一方面需要讨论 SSL 可以显著提高哪些学习任务的性能,拓展 SSL 在现实领域的实际应用,另一方面需要制定出一个统一的令人信服的 SSL 方法的使用规程.此外,目前有许多的半监督分类方法,而对半监督回归问题的研究比较有限.未来有待继续研究半监督分类和半监督回归之间的关系,并提出其他半监督回归方法.

9 结束语

SSL 作为 ML 中的一个重要问题,能够同时利用大量的无类标签的样例和有限的有类标签的样本一起训练,近几十年来得到了越来越多的关注,许多 SSL 理论和方法得以发展.本文详细概述了 SSL 的相关基本概念和研究发展历程,分别从分类、回归、聚类及降维这四个方面全面总结了 SSL 的理论和方法,综述了 SSL 的理论分析、误差界和样本复杂度,并在最后从理论分析、抗干扰性与可靠性、训练样例与参数的选取、优化求解和研究拓展五个方面指出了 SSL 未来的研究方向.随着 SSL 理论与方法研究的深入,SSL 将被更加广泛地应用在各个领域.

参 考 文 献

[1] Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning. Cambridge, USA: MIT Press, 2006

[2] Zhu X. Semi-supervised learning literature survey. Department of Computer Science, University of Wisconsin-Madison, Wisconsin: Technical Report 1530, 2006

[3] Zhou Z H, Li M. Semi-supervised learning by disagreement. Knowledge and Information Systems, 2010, 24(3): 415-439

[4] Wang W, Zhou Z H. Analyzing co-training style algorithms //Proceedings of the 18th European Conference on Machine Learning (ECML'07). Warsaw, Poland, 2007: 454-465

[5] Yarowsky D. Unsupervesed word sense disambiguation rivaling supervised methods//Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, USA, 1995: 189-196

[6] Scudder H J. Probability of error of some adaptive pattern-recognition machines. IEEE Transductions on Information Theory, 1965, 11(3): 363-371

[7] Fralick S. Learning to recognize patterns without a teacher. IEEE Transductions on Information Theory, 1967, 13(1): 57-64

[8] Agrawala A K. Learning with a probabilistic teacher. IEEE Transductions on Information Theory, 1970, 16(4): 373-379

[9] Merz C J, St Clair D C, Bond W E. Semi-supervised adaptive resonance theory//Proceedings of the 1992 International Joint Conference on Neural Networks. Baltimore, USA, 1992: 851-856

[10] Shahshahani B M, Landgrebe D A. 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

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

[12] Vapnik V, Sterin A. On structural risk minimization or overall risk in a problem of pattern recognition. Automation and Remote Control, 1997, 10(3): 1495-1503

[13] Joachims T. Transductive inference for text classification using support vector machines//Proceedings of the 16th International Conference on Machine Learning. San Francisco, USA, 1999: 200-209

[14] De Bie T, Cristianini N. Convex methods for transduction// Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA, 2003: 73-80

[15] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1977, 39(1): 1-38

[16] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labeled and unlabelled data//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 1996: 571-577

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

[18] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic function//Proceedings of the 20th International Conference on Machine Learning. Washington, USA, 2003: 912-919

[19] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006, 7: 2399-2434

[20] Klein D, Kamvar S D, Manning C D. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 307-314

[21] Castelli V, Cover T M. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Transactions on Information Theory, 1996, 42(6): 2102-2117

[22] Sinha K, Belkin M. The value of labeled and unlabeled examples when the model is imperfect//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2007: 1361-1368

[23] Balcan M F, Blum A. A discriminative model for semi-supervised learning. Journal of the Association for Computing Machinery, 2010, 57(3): 19

[24] Singh A, Nowak R D, Zhu X. Unlabeled data: Now it helps, now it doesn't//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2008, 21: 1513-1520

[25] Balcan M F, Blum A, Yang K. Co-training and expansion: Towards bridging theory and practice//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA, 2004, 17: 89-96

[26] Goldberg A B, Zhu X. Seeing stars when there aren't many stars: Gragh-based semi-supervised learning for sentiment categorization//Proceedings of the 1st Workshop on Graph Based Methods for Natural Language Processing. New York, USA, 2006: 45-52

[27] Ratsaby J, Venkatesh S S. Learning from a mixture of labeled and unlabeled examples with parametric side information//Proceedings of the 8th Annual Conference on Computational Learning Theory. Santa Cruz, USA, 1995: 412-417

[28] Nigam K, Mccallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2-3): 103-134

[29] Zhou Y, Goldman S. Democratic co-learning//Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, USA, 2004: 594-602

[30] Zha Z J, Mei T, Wang J, et al. Graph-based semi-supervised learning with multiple labels. Journal of Visual Communication and Image Representation, 2009, 20(2): 97-103

[31] Wu Y, Yuan P. Application of density-sensitive distance measure in special image clustering. Computer Engineering, 2009, 35(6): 210-212

[32] Xing E P, Ng A Y, Jordan M I, et al. Distance metric learning, with application to clustering with side-information//Proceedings of the 16th Annual Conference on Neural Information Processing Systems. Cambridge, UK, 2003: 521-528

[33] Yu S, Yu K, Tresp V, et al. Supervised probabilistic principal component analysis//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2006: 464-473

[34] Hwa R, Osborne M, Sarkar A, Steedman M. Corrected co-training for statistical parsers//Proceedings of the ICML Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining at the 20th International Conference of Machine Learning. Washington, USA, 2003: 95-102

[35] Johnson R, Zhang T. Graph-based semi-supervised learning and spectral kernel design. IEEE Transactions on Information Theory, 2008, 54(1): 275-288

[36] Mallapragada P K, Jin R, Jain A K, Liu Y. SemiBoost: Boosting for semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2000-2014

[37] Shin H H, Hill N J, Raetsch G. Graph based semi-supervised learning with sharper edges//Proceedings of the 17th European Conference on Machine Learning. Berlin, Germany, 2006: 401-412

[38] Shang F, Jiao L C, Liu Y, et al. Semi-supervised learning with nuclear norm regularization. Pattern Recognition, 2013, 46(8): 2323-2336

[39] Wang J, Jebara T, Chang S F. Semi-supervised learning using greedy max-cut. The Journal of Machine Learning Research, 2013, 14(1): 771-800

[40] Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping//Proceedings of the 16th National Conference on Artificial Intelligence. Orlando, USA, 1999: 474-479

[41] Collins M, Singer Y. Unsupervised models for named entity classification//Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, USA, 1999: 100-110

[42] Yu Z T, Su L, Li L, et al. Question classification based on co-training style semi-supervised learning. Pattern Recognition Letters, 2010, 31(13): 1975-1980

[43] Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2007, 37(6): 1088-1098

[44] Zhou Z H, Chen K J, Dai H B. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems, 2006, 24(2): 219-244

[45] Chen J, Ji D, Tan C L, et al. Relation extraction using label propagation based semi-supervised//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA, 2006: 129-136

[46] Camps-Valls G, Bandos Marsheva T, Zhou D. Semi-supervised graph-based hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(10): 3044-3054

[47] Cheng S, Shi Y, Qin Q. Particle swarm optimization based semi-supervised learning on chinese text categorization// Proceedings of the 2012 IEEE Congress on Evolutionary Computation. Brisbane, Australia, 2012: 1-8

[48] Zhang T, Xu C, Zhu C, et al. A generic framework for video annotation via semi-supervised learning. IEEE Transactions on Multimedia, 2012, 14(4): 1206-1219

[49] Carlson A, Betteridge J, Wang R C, et al. Coupled semi-supervised learning for information extraction//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 101-110

[50] Guillaumin M, Verbeek J, Schmid C. Multimodal semi-supervised learning for image classification///Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 902-909

[51] He X. Incremental semi-supervised subspace learning for image retrieval//Proceedings of the 12th Annual ACM International Conference on Multimedia. New York, USA, 2004: 2-8

[52] Balcan M F, Blum A, Choi P P, et al. Person identification in webcam images: An application of semi-supervised learning //Proceedings of the International Conference on Machine Learning Workshop on Learning from Partially Classified Training Data. Las Vegas, USA, 2005: 1-9

[53] Wang J, Kumar S, Chang S F. Semi-supervised hashing for scalable image retrieval///Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3424-3431

[54] Chapelle O, Zien A. Semi-supervised learning by low density separation//Proceedings of the 10th Information Workshop on Artificial Intelligence and Statistics. Savannah Hotel, Barbados, 2005: 57-64

[55] Ben-David S, Lu T, Pai D. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning//Proceedings of the 21st Annual Conference on Learning Theory. Helsinki, Finland, 2008: 33-44

[56] Baluja S. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 1998: 854-860

[57] Balcan M F, Blum A. An augmented PAC model for semi-supervised learning//Chapelle O, Scholkopf B, Zien A eds. Semi-Supervised Learning. Cambridge, USA: MIT Press, 2006: 61-89

[58] Vapnik V. Transductive inference and semi-supervised learning //Chapelle O, Scholkopf B, Zien A eds. Semi-Supervised Learning. Cambridge, USA: MIT Press, 2006: 453-472

[59] Zhang T, Ando R K. Analysis of spectral kernel design based semi-supervised learning//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2006, 18: 1601-1608

[60] Brefeld U, Gartner T, Scheffer T, et al. Efficient co-regularised least squares regression//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 137-144

[61] Zhou Z H, Li M. Semi-supervised regression with co-training //Proceedings of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland, 2005: 908-916

[62] Basu S, Bilenko M, Mooney R J. A probabilistic framework for semi-supervised clustering//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 59-68

[63] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA, 2001: 577-584

[64] Zhang D, Zhou Z H, Chen S. Semi-supervised dimensionality reduction//Proceedings of the 7th SIAM International Conference on Data Mining. Minneapolis, USA, 2007: 629-634

[65] Dasgupta S, Littman M L, McAllester D. PAC generalization bounds for co-training//Dietterich T G, Becker S, Ghahramani Z eds. Advances in Neural Information Processing System 14. Cambridge, USA, 2002: 375-382

[66] Zhou Z H, Zhan D C, Yang Q. Semi-supervised learning with very few labeled training examples//Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2007: 675-680

[67] Wang W, Zhou Z H. Co-training with insufficient views// Proceedings of the Asian Conference on Machine Learning. Canberra, Australia, 2013: 467-482

[68] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training//Proceedings of the 9th ACM International Conference on Information and Knowledge Management. McLean, USA, 2000: 86-93

[69] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data//Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA, 2000: 327-334

[70] Abney S. Bootstrapping//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 360-367

[71] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[72] Wu J, Xiao Z B, Wang H S, et al. Learning with both unlabeled data and query logs for image search. Computers & Electrical Engineering, 2014, 40(3): 964-973

[73] Steedman M, Osborne M, Sarkar A, et al. Bootstrapping statistical parsers from small data sets//Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics - Volume 1. Budapest, Hungary, 2003: 331-338

[74] Wang W, Zhou Z H. A new analysis of co-training//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 1135-1142

[75] Yan Y, Rosales R, Fung G, Dy J. Modeling multiple annotator expertise in the semi-supervised learning scenario//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA, 2012: 674-682

[76] Castelli V, Cover T M. On the exponential value of labeled samples. Pattern Recognition Letters, 1995, 16(1): 105-111

[77] Cozman F G, Cohen I. Unlabeled data can degrade classification performance of generative classifiers//Proceedings of the 15th International Florida Artificial Intelligence Society Conference. Pensacola, USA, 2002: 327-331

[78] Jensen F V. An Introduction to Bayesian Networks. London: UCL Press, 1996

[79] Saul L K, Jaakkola T, Jordan M I. Mean field theory for sigmoid belief networks. Journal of Artificial Intelligence Research, 1996, 4(1): 61-76

[80] Shental N, Bar-Hillel A, Hertz T, et al. Computing Gaussian mixture models with EM using equivalence constraints//Proceedings of the Advances in Neural Information Processing Systems 16. Cambridge, USA, 2004, 16(8): 465-472

[81] Rabiner L R. A tutorial on hidden markov models and selected applications in speech recogntion. Proceedings of the IEEE, 1989, 77(2): 257-286

[82] Yedidia J S, Freeman W T, Weiss Y. Generalized belief propagation//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 2000: 689-695

[83] Duda R O, Hart P E. Pattern Classification and Scene Analysis. New York, USA: Wiley, 1973

[84] Meir R. Empirical risk minimization versus maximum-likelihood estimation: A case study. Neural Computation, 1995, 7(1): 144-157

[85] Chapelle O, Sindhwani V, Keerthi S S. Optimazation techniques for semi-supervised support vector machines. The Journal of Machine Learning Research, 2008, 9(6): 203-233

[86] Li Y F, Zhou Z H. Towards making unlabeled data never hurt//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 1081-1088

[87] Fisher R A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7(2): 179-188

[88] Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. Neural Computation, 2000, 12(10): 2385-2404

[89] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization///Proceedings of the Advances in Neural Information Processing Systems 17. Cambridge, USA, 2004: 529-536

[90] Zhu X, Lafferty J, Rosenfeld R. Semi-Supervised Learning with Graphs[Ph.D. dissertation]. Carnegie Mellon University, Pennsylvania, USA, 2005

[91] Cover T M, Hart P. The nearest neighbor decision rule. IEEE Transactions on Information Theory, 1967, 13(1): 21-27

[92] Corduneanu A, Jaakkola T S. Distributed information regularization on graphs//Proceedings of the Advances in Neural Information Processing Systems 17. Cambridge, USA, 2004: 297-304

[93] Smola A, Kondor R. Kernels and regularization on graphs//Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop. Washington, USA, 2003: 144-158

[94] Szlam A D, Maggioni M, Coifman R R. Regularization on graphs with function-adapted diffusion processes. The Journal of Machine Learning Research, 2008, 9(6): 1711-1739

[95] Huang T M, Kecman V, Kopriva I. Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-Supervised, and Unsupervised Learning. New York, USA: Springer, 2006

[96] Blum A, Lafferty J, Rwebangira M R, et al. Semi-supervised learning using randomized mincuts//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 13

[97] Zhu X, Lafferty J D. Harmonic mixtures: Combining mixture models and graph-based methods for inductive and scalable semi-supervised learning//Proceedings of the 22nd International Conference of Machine Learning. Bonn, Germany, 2005: 1052-1059

[98] Zhou D, Scholkopf B. Learning from labeled and unlabeled data using random walks//Proceedings of the 26th Pattern Recognition Symposium. Tubingen, Germany, 2004: 237-244

[99] Belkin M, Niyogi P. Semi-supervised learning on riemannian manifolds. Machine Learning, 2004, 56(1-3): 209-239

[100] Zhou D, Scholkopf B. A regularization framework for learning from graph data//Proceedings of the Workshop on Statistical Relational Learning at 21st International Conference on Machine Learning. Vancouver, Canada, 2004: 132-137

[101] Chen K, Wang S. Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 129-143

[102] Goldberg A B, Li M, Zhu X. Online manifold regularization: A new learning setting and empirical study//Proceedings of the 18th European Conference on Principles of Data Mining and Knowledge Discovery. Antwerp, Belgium, 2008: 393-407

[103] Yan S, Wang H. Semi-supervised learning by sparse representation//Proceedings of the 6th VLDB Workshop on Secure Data Management. Lyon, France, 2009: 792-801

[104] Liu W, He J, Chang S F. Large graph construction for scalable semi-supervised learning//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 679-686

[105] Dhillon P S, Keerthi S S, Bellare K, et al. Deterministic annealing for semi-supervised structured output learning//Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. La Palma, Canary Islands, 2012: 299-307

[106] Breve F, Zhao L, Quiles M, et al. Particle competition and cooperation in networks for semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(9): 1686-1698

[107] Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: From transductive to semi-supervised learning//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 824-831

[108] Tang J, Hua X S, Qi G J, et al. Structure-sensitive manifold ranking for video concept detection//Proceedings of the 15th International Conference on Multimedia. Augsburg, Germany, 2007: 852-861

[109] He J, Carbonell J G, Liu Y. Graph-based semi-supervised learning as a generative model//Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 2492-2497

[110] Zhang C, Wang F. Graph-based semi-supervised learning. Artificial Life and Robotics, 2009, 14(4): 445-448

[111] Carreira-Perpinan M A, Zemel R S. Proximity graphs for clustering and manifold learning//Proceedings of the Advances in Neural Information Processing Systems 17. Cambridge, USA, 2004, 17: 225-232

[112] Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55-67

[113] Hein M, Maier M. Manifold denoising//Proceedings of the Advances in Neural Information Processing Systems 19. Cambridge, USA, 2006, 19: 561-568

[114] Zhou Z H, Li M. Semi-supervised regression with co-training style algorithm. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(11): 1479-1493

[115] Ma L, Wang X. Semi-supervised regression based on support vector machine co-training. Computer Engineering and Applications, 2011, 47(3): 177-180

[116] Wang M, Hua X S, Song Y, et al. Semi-supervised kernel regression//Proceedings of the 6th IEEE International Conference on Data Mining. Hong Kong, China, 2006: 1130-1135

[117] Verbeek J J, Vlassis N. Gaussian fields for semi-supervised regression and correspondence learning. Pattern Recognition, 2006, 39(10): 1864-1875

[118] Pozdnoukhov A, Bengio S. Semi-supervised kernel methods for regression estimation//Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France, 2006: 577-580

[119] Yang J, Wang J, Zhong N. Laplacian semi-supervised regression on a manifold. Journal of Computer Research and Development, 2007, 44(7): 1121-1127

[120] Navaratnam R, Fitzgibbon A W, Cipolla R. The joint manifold model for semi-supervised multi-valued regression //Proceedings of the IEEE 11th International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-8

[121] Ji M, Yang T, Lin B, et al. A simple algorithm for semi-supervised learning with improved generalization error bound//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, UK, 2012: 1223-1230

[122] Wagstaff K, Cardie C. Clustering with instance-level constraints//Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA, 2000: 1103-1110

[123] Ding C, He X. Linearized cluster assignment via spectral ordering//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 30-37

[124] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2001: 849-856

[125] Kamvar K, Sepandar S, Klein K, et al. Spectral learning//Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003: 561-566

[126] Ji X, Xu W. Document clustering with prior knowledge//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA, 2006: 405-412

[127] Zhang L, Li M. Density-based constraint expansion method for semi-supervised clustering. China Computer Engineering, 2008, 34(10): 13-15

[128] Wu Y, Yuan P, Yu N. An improved density-sensitive semi-supervised clustering algorithm//Proceedings of the 5th International Conference on Visual Information Engineering. Xi'an, China, 2008: 106-110

[129] Luo X, Wang S. A novel semi-supervised clustering method based on double similarity measure. China Computer Applications and Software, 2008, 25(4): 219-250

[130] Bijral A S, Batliff N, Srebro N. Semi-supervised learning with density based distances. Machine Learning, 2012: 43-50

[131] Nesterov Y. Introductory Lectures on Convex Optimization. Boston, USA: Kluwer Academic Publishers, 2004

[132] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 27-34

[133] Bar-Hillel A, Hertz T, Shental N, et al. Learning a Mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2005, 6(6): 937-965

[134] Chang H, Yeung D Y. Locally linear metric adaptation with application to semi-supervised clustering and image retrieval. Pattern Recognition, 2006, 39(7): 1253-1264

[135] Lange T, Law M H C, Jain A K, Buhmann J M. Learning with constrained and unlabelled data//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA, 2005: 731-738

[136] Davidson I, Ravi S S. Clustering with constraints: Feasibility issues and the k-means algorithm//Proceedings of the 2005 SIAM International Conference on Data Mining. Newport Beach, USA, 2005: 201-211

[137] Cohn D, Caruana R, McCallum A. Semi-supervised clustering with user feedback. Constrained Clustering: Advances in Algorithms, Theory, and Applications, 2003, 4(1): 17-32

[138] Jain A K, Mallapragada P K, Law M. Bayesian feedback in data clustering//Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China, 2006: 374-378

[139] Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 11

[140] Gao J, Tan P N, Cheng H. Semi-supervised clustering with partial background information//Proceedings of the 6th SIAM International Conference on Data Mmining. Bethesda, USA, 2006: 20-22

[141] Tang W, Xiong H, Zhong S, et al. Enhancing semi-supervised clustering: A feature projection perspective//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 707-716

[142] Lu Z, Leen T K. Semi-supervised learning with penalized probabilistic clustering///Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2004: 849-856

[143] Nelson B, Cohen I. Revisiting probabilistic models for clustering with pair-wise constraints//Proceedings of the 24th International Conrerence on Machine Learning. Corvallis, USA, 2007: 673-680

[144] Chen Y, Rege M, Dong M, et al. Incorporating user provided constraints into document clustering//Proceedings of the 7th IEEE International Conference on Data Mining. Washington, USA, 2007: 103-112

[145] Yin X S, Hu E L, Chen S C. Discriminative semi-supervised clustering analysis with pairwise constraints. Journal of Software, 2008, 19(11): 2791-2802

[146] Xia Y. A global optimization method for semi-supervised clustering. Data Mining and Knowledge Discovery, 2009, 18(2): 214-256

[147] Lu Z, Ip H H S. Constrained spectral clustering via exhaustive and efficient constraint propagation//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece, 2010: 1-14

[148] Yeung D Y, Chang H. A kernel approach for semi-supervised metric learning. IEEE Transactions on Neural Networks, 2007, 18(1): 141-149

[149] Kulis B, Basu S, Dhillon I, et al. Semi-supervised graph clustering: A kernel approach. Machine Learning, 2009, 74(1): 1-22

[150] Yan B, Domeniconi C. An adaptive kernel method for semi-supervised clustering//Proceedings of the 17th European Conference on Machine Learning. Berlin, Germany, 2006: 521-532

[151] Soleymani B M, Afsari F, Bagheri S S, et al. Scalable semi-supervised clustering by spectral kernel learning. Pattern Recognition Letters, 2014, 45: 161-171

[152] Tsang I W, Cheung P M, Kwok J T. Kernel relevant component analysis for distance metric learning//Proceedings of the 2005 IEEE International Joint Conference on Neural Networks. Montreal, Canada, 2005: 954-959

[153] Chen J H, Zhao Z, Ye J P, et al. Nonlinear adaptive distance metric learning for clustering//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 123-132

[154] Chang H, Yeung D Y, Cheung W K. Relaxational metric adaptation and its application to semi-supervised clustering and content-based image retrieval. Pattern Recognition, 2006, 39(10): 1905-1917

[155] Xiang S, Nie F, Zhang C. Learning a Mahalanobis distance metric for data clustering and classification. Pattern Recognition, 2008, 41(12): 3600-3612

[156] Xu L, Neufeld J, Larson B, et al. Maximum margin clustering //Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2004: 1537-1544

[157] Xu L, Schuurmans D. Unsupervised and semisupervised multi-class support vector machines//Proceedings of the 20th International Conference on Artificial Intelligence. Pittsburgh, USA, 2005: 904-910

[158] Valizadegan H, Jin R. Generalized maximum margin clustering and unsupervised kernel learning//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2007: 1417

[159] Zhang K, Tsang I W, Kwok J T. Maximum margin clustering made practical. IEEE Transactions on Neural Networks, 2009, 20(4): 583-596

[160] Zhao B, Wang F, Zhang C. Efficient maximum margin clustering via cutting plane algorithm//Proceedings of the SIAM International Conference on Data Mining. Atlanta, USA, 2008: 751-762

[161] Zhao B, Wang F, Zhang C. Efficient multiclass maximum margin clustering//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 1248-1255

[162] Harris T. Credit scoring using the clustered support vector machine. Expert Systems with Applications, 2015, 42(2): 741-750

[163] Li Y F, Tsang I W, Kwok J T, et al. Tighter and convex maximum margin clustering//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater Beach, USA, 2009: 344-351

[164] Hu Y, Wang J, Yu N, et al. Maximum margin clustering with pairwise constraints//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 253-262

[165] Gieseke F, Pahikkala T, Kramer O. Fast evolutionary maximum margin clustering//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 361-368

[166] Zhang D, Wang F, Si L, Li T. M3IC: Maximum margin multiple instance clustering//Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, USA, 2009: 1339-1344

[167] Zhang D, Wang F, Si L, et al. Maximum Margin multiple instance clustering with applications to image and text clustering. IEEE Transactions on Neural Networks, 2011, 22(5): 739-751

[168] Iosifidis A, Tefas A, Pitas I. Kernel reference discriminant analysis. Pattern Recognition Letters, 2014, 49: 85-91

[169] Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40-51

[170] Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. The Journal of Machine Learning Research, 2007, 8(5): 1027-1061

[171] Costa J A, Hero A O. Classification constrained dimensionality reduction//Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, USA, 2005: 1077-1080

[172] Chen J, Ye J, Li Q. Integrating global and local structures: A least squares framework for dimensionality reduction//Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[173] Cai D, He X, Han J. Semi-supervised discriminant analysis //Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-7

[174] Zhang Y, Yeung D Y. Simi-supervised discriminant analysis using robust path-based similarity//Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[175] Song Y, Nie F, Zhang C, et al. A unified framework for semi-supervised dimensionality reduction. Pattern Recognition, 2008, 41(9): 2789-2799

[176] Zhang Y, Yeung D Y. Semi-supervised discriminant analysis via CCCP//Proceedings of the 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Antwerp, Belgium, 2008: 644-659

[177] Sugiyama M, Ide T, Nakajima S, et al. Semi-supervised local fisher discriminant analysis for dimensionality reduction. Machine Learning, 2010, 78(1-2): 35-61

[178] Chatpatanasiri R, Kijsirikul B. A unified semi-supervised dimensionality reduction framework for manifold learning. Neurocomputing, 2010, 73(10): 1631-1640

[179] Tang W, Zhong S. Pairwise constraints-guided dimensionality reduction//Proceedings of the SDM Workshop on Feature Selection for Data Mining. Bethesda, USA, 2006: 59-66

[180] Hoi S, Liu W, Lyu M R, et al. Learning distance metrics with contextual constraints for image retrieval//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 2072-2078

[181] Cevikalp H, Verbeek J, Jurie F, et al. Semi-Supervised dimensionality reduction using pairwise equivalence constraints //Proceedings of the 3rd International Conference on Computer Vision Theory and Applications. Funchal, Portuguesa, 2008: 489-496

[182] Wei J, Peng H. Neighbourhood preserving based semi-supervised dimensionality reduction. Electronics Letters, 2008, 44(20): 1190-1192

[183] Peng Y, Zhang D Q. Semi-Supervised canonical correlation analysis algorithm. Journal of Software, 2008, 19(11): 2822-2832

[184] Baghshah M S, Shouraki S B. Semi-Supervised metric learning using pairwise constraints//Proceedings of the 21st International Joint Conference on Artificial Intelligence. San Francisco, USA, 2009: 1217-1222

[185] Davidson I. Knowledge driven dimension reduction for clustering//Proceedings of the 21st International Joint Conference on Artificial Intelligence. San Francisco, USA, 2009: 1034-1039

[186] Ham J, Lee D D, Mika S, et al. A kernel view of the dimensionality reduction of manifolds//Proceedings of the 21st International Conferences on Machine Learning. Banff, Canada, 2004: 47-54

[187] Yang X, Fu H, Zha H, et al. Semi-supervised nonlinear dimensionality reduction//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 1065-1072

[188] Memisevic R, Hinton G E. Multiple relational embedding// Proceedings of the Advances in Neural Information Processing Systems 17. Cambridge, USA, 2004: 913-920

[189] Yu J, Tian Q. Learning image manifolds by semantic subspace projection//Proceedings of the 14th Annual ACM International Conference on Multimedia. Santa Barbara, USA, 2006: 297-306

[190] Liu W, Jiang W, Chang S F. Relevance aggregation projections for image retrieval//Proceedings of the 7th ACM International Conference on Image and Video Retrieval. New York, USA, 2008: 119-126

[191] Fox-Roberts P, Rosten E. Unbiased generative semi-supervised learning. The Journal of Machine Learning Research, 2014, 15(1): 367-443

[192] Lafferty J D, Wasserman L A. Statistical analysis of semi-supervised regression//Proceedings of the Advances in Neural Information Processing System. Vancouver, Canada,

2007：801-808

[193] Urner R, Ben-David S, Shalev-Shwartz S. Access to unlabeled data can speed up prediction time//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011：641-648

[194] Darnstadt M, Simon H U. Smart PAC-learners. Theoretical Computer Science, 2011, 412(19)：1756-1766

[195] Dillon J V, Balasubramanian K, Lebanon G. Asymptotic analysis of generative semi-supervised learning//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010：295-302

[196] Li Y F, Zhou Z H. Improving semi-supervised support vector machines through unlabeled instances selection//Proceedings of the 25th AAAI Conference on Artificial Intelligence. Chicago, Illinois, USA, 2010：386-391

[197] El-Yaniv R, Pechyony D. Transductive rademacher complexity and its applications//Proceedings of the 20th Annual Conference on Learning Theory. San Diego, USA, 2007：157-171

[198] Kaariainen M. Generalization error bounds using unlabeled data//Proceedings of the 18th Annual Conference on Learning Theory. Bertinoro, Italy, 2005：127-142

[199] Derbeko P, El-Yaniv R, Meir R. Explicit learning curves for transduction and application to clustering and compression algorithms. Journal of Articial Intelligence Research, 2004, 22：117-142

[200] Gentile C, Helmbold D P. Improved lower bounds for learning from noisy examples and information-theoretic approach//Proceedings of the 11th Annual Conference on Learning Theory. Madison, USA, 1998：104-115

LIU Jian-Wei, born in 1966, Ph.D., associate professor. His main research interests include intelligent information processing, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing.

LIU Yuan, born in 1989, M. S. candidate. Her main research interests include machine learning and digital image processing.

LUO Xiong-Lin, born in 1963, Ph. D., professor. His main research interests include intelligent control, and analysis, prediction, controlling of complicated nonlinear system.

Background

Semi-supervised learning is a hot topic in machine learning with interesting theoretical properties and practical applications in recent years and has received significant attention, and many researchers dedicate to the research in this field. Several literatures summarizing semi-supervised learning methods exist at present. Chapelle and his colleagues reviewed research and development on semi-supervised learning of the early stage. Zhu summarized semi-supervised learning methods detailedly. Zhou and Li summarized semi-supervised learning methods based on disagreement. Zhou reviewed co-training in detail. Researchers have come up with a great number of semi-supervised learning methods, including generative methods, discriminative methods, graph-based methods, disagreement- based methods, measure-based methods, constraint-based methods. Semi-supervised learning methods have been applied to solve practical problems, such as image retrieval, natural language processing, text categorization, and so on.

In this paper we reviewed semi-supervised learning methods systematically. In contrast to existing work in semi-supervised learning, we expounded new research and development on semi-supervised learning methods. We summarized some basic concepts about semi-supervised learning, and detailed semi-supervised learning methods from four aspects, namely classification, regression, clustering, and dimension reduction. Then we studied theoretical analysis on semi-supervised learning, and gave error bounds and sample complexity. We discussed the future research on semi-supervised learning at last.

This work is supported by the National Basic Research Program (973 Program) of China (2012CB720500), the National Natural Science Foundation of China (21006127), Basic Scientific Research Foundation of China University of Petroleum (JCXK-2011-07).

The work can be viewed as reviewed semi-supervised learning methods systematically, which include semi-supervised classification methods, semi-supervised regression methods, semi-supervised clustering methods, and semi-supervised dimension reduction methods. The main aim of this paper is to expound new research and development on semi-supervised learning methods. We also summarized some basic concepts about semi-supervised learning, studied theoretical analysis on semi-supervised learning, and discussed the future research on semi-supervised learning.