半监督聚类综述

秦 悦[1] 丁世飞[1,2]
(中国矿业大学计算机科学与技术学院 江苏 徐州 221116)[1]
(中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)[2]

摘 要 半监督聚类是结合半监督学习与聚类分析而提出的新的学习方法,其在机器学习中得到了广泛的重视和应用。传统无监督聚类算法在划分数据时并不需要任何数据属性,但在实际应用中,存在少量带有独立类标签或成对约束的监督信息的数据样本,学者们致力于将这些为数不多的监督信息运用于聚类,以得到更优的聚类结果,从而提出了半监督聚类。文中主要介绍了半监督聚类的理论基础和算法思想,并对半监督聚类的最新研究进展进行了综述。首先,对半监督学习的研究现状和分类进行了概述,并将生成式半监督学习、半监督SVM、基于图的半监督学习和协同训练这4种分类方法进行了对比;其次,针对半监督学习的聚类进行了详细的描述,并对4种典型半监督聚类算法(Cop-Kmeans算法、LCop-Kmeans算法、Seeded-Kmeans算法和SC-Kmeans算法)的算法思想进行了分析和总结,同时对这4种算法的优缺点进行了评价;然后,按照基于约束的半监督聚类和基于距离的半监督聚类两种情况,分别对半监督聚类的研究现状进行了阐述;最后,探讨了半监督聚类在生物信息学、图像分割以及计算机其他领域内的应用以及未来的研究方向。文中旨在使初学者能够快速了解半监督聚类的进展,理解典型的算法思想,并在之后的实际应用中能起到一定的指导作用。

关键词 半监督学习,聚类,成对约束,标签,半监督聚类,机器学习
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.002

Survey of Semi-supervised Clustering

QIN Yue[1] DING Shi-fei[1,2]
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Semi-supervised clustering is a new learning method combining semi-supervised learning and clustering analysis, and it has been used widely in machine learning. The traditional unsupervised clustering algorithms do not need any data attributes when dividing data, but in practical applications, there are a small number of data samples for supervised information with independent class labels or paired constraints, so scholars are committed to applying these few supervised information into clustering to obtain better clustering results, thus proposing semi-supervised clustering. This paper mainly introduced the theoretical basis and algorithm ideas of semi-supervised clustering, and summarized the latest progress of semi-supervised clustering. Firstly, the current situation and classification of semi-supervised learning were reviewed, and the generative semi-supervised learning, semi-supervised SVM, semi-supervised learning based on graph and collaborative training were compared. Secondly, the clustering of semi-supervised learning was described in detail, four typical semi-supervised clustering algorithms (Cop-Kemans algorithm, LCop-Kmeans algorithm, Seeded-Kmeans algorithm and SC-Kmeans algorithm) were analyzed and summarized, and their advantages and disadvantages were evaluated. Then, according to the two situations of semi-supervised clustering based on constraints and the semi-supervised clustering based on distance, the research status of semi-supervised clustering was expounded respectively. Finally, the applications of semi-supervised clustering in bioinformatics, image segmentation and other fields of computer and the future research directions were discussed. This paper aims to enable beginners to quickly know about the progress of semi-supervised clustering and understand the typical algorithm ideas, and it can play a guiding role in actual applications afterwards.

到稿日期:2018-09-13 返修日期:2018-11-26 本文受国家自然科学基金(61672522,61379101),国家重点基础研究计划(2013CB329502)资助。
秦 悦(1994-),女,硕士生,主要研究方向为人工智能、机器学习、聚类,E-mail:qinyue23333@163.com;丁世飞(1963-),男,博士,教授,CCF理事,主要研究方向为模式识别与人工智能、机器学习与数据挖掘、粗糙集与软计算、粒度计算、感知与认知计算,E-mail:dingsf@cumt.edu.cn(通信作者)。

Keywords Semi-supervised learning, Clustering, Pairwise constraints, Label, Semi-supervised clustering, Machine learning

1 引言

当今社会处于大数据时代,数据量越来越多且越来越繁杂。在面对海量数据时,如何挖掘数据中有用的信息,并利用有标签的数据来优化算法,有着重要的研究意义。机器学习中两种最基本的学习方法是:监督学习和无监督学习。传统的聚类是一种无监督分析方法,属于探索性分析,在对数据进行划分的过程中不依赖任何背景知识和相应的假设,单纯地按照相似性规则自然划分,让每一类中的数据的关系尽可能相似而类与类之间数据的关系尽可能不相似。

在现实生活的应用中,只有小部分数据带有标签信息,这些数据不足以支撑监督学习,若继续使用传统无监督学习的聚类分析方法,会让这些找出的标签白白浪费。为了合理运用数据对象中或多或少的背景知识,半监督学习应运而生。

半监督学习作为监督学习与无监督学习相结合的一种学习方法,可以同时利用已标签数据和未标签数据。按照学习任务的不同,半监督学习可以分成半监督分类和半监督聚类。

与 K-means 算法[1]、EM 算法[2]等传统的聚类算法不同,半监督聚类是一种新型的研究算法,其主要思想是将聚类和半监督学习相结合,通过海量数据中少量的标签数据和先验知识来提高聚类性能,从而得到性能更优的结果。半监督聚类方法可以分为:基于约束的半监督聚类算法、基于距离的半监督聚类算法、基于约束和距离相结合的半监督聚类算法。

半监督聚类在数据挖掘领域内具有很好的应用前景,近年来的相关研究成果越来越多。聚类和半监督学习方向的综述文献已有不少,但半监督聚类作为一个新的研究方向,国内的综述型文献极少,且发表的年份也较为久远。例如,李昆仑等[3]对半监督聚类的相关进展进行了阐述,着重讨论了半监督聚类的分类,并提出了基于约束的半监督模糊 C-means 聚类算法;但其对半监督聚类的发展现状以及相关典型算法的描述并不是特别全面。熊建斌等[4]也对半监督聚类算法的研究现状进行了综述,且分析了几种典型半监督聚类算法的优点与局限性;但其也只是提及了几种典型算法,并未对算法进行更为详细的描述。

本文将对半监督聚类进行综述,旨在向读者详细介绍这种新型算法,使相关专业人员能清楚地了解半监督聚类的相关知识、典型算法以及研究现状。第 2 节介绍了半监督学习的研究现状和分类;第 3 节将对半监督聚类的概念和分类进行详细阐述,并分别介绍半监督聚类的相关典型算法;最后总结全文,并对未来研究方向进行讨论。

2 半监督学习

半监督学习早在二十世纪八九十年代就成为了机器学习的研究热点,它能够将监督学习与无监督学习相结合,同时利用已标签数据和未标签数据共同进行训练学习。文献[5]在基本概念、发展历程、分类以及相关算法方面对半监督学习进行了详细的叙述。

2.1 半监督学习的研究现状

半监督学习是由自学习发展而来的,在之后的发展中逐渐出现了自学习、直推学习、生成式模型等学习方法。Sceudder[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监督学习的方法。该类方法主要使用了迭代的思想,不断地重复监督学习,将得出的最优化标记结果运用到下一轮中,并与类标签一并加入到样本集中,继续迭代,反复监督。此方法的优势在于简单易懂、便于操作,但是很容易产生误差错误被迭代而导致恶性循环的弊端。"Semi-supervised"由 Merz 等[9]首次提出,并且可以将半监督用于分类。Shahshahani 等[10]证明了用无标签样本能减缓小样本下的"Hughes"现象,此思想的提出使得无标签样本和半监督学习受到广泛关注。Kingma 等[11]提出了深度生成模型的半监督学习。Klein 等[12]首次提出了用于聚类的半监督距离度量学习方法。Cheng 等[13]提出了一种基于半监督分类器的粒子群优化算法,用于解决中文文本的分类问题。Wang 等[14]提出了半监督散列方法,用于处理大规模图像的检索问题。

2.2 半监督学习的分类

按照学习场景的不同,半监督学习划分为 4 类:半监督降维[15]、半监督回归[16]、半监督聚类[17]、半监督分类[18]。而半监督学习主要还是按照半监督分类,可以分为:基于生成式的半监督学习、基于支持向量机的半监督学习、基于图的半监督学习和协同训练[19]。

2.2.1 基于生成式半监督学习

生成式学习方法[20](Generative Learning Algorithm,GLA)的主要思想是:假设样本和类标签是某种概率分布,按照生成式模型[21],找出让后验概率分布最高的标签,并对其进行标记。通俗来说,生成式方法就是把测试用例分别放进各个模型中,最后比较其结果,选择最优的作为类标签。

2.2.2 基于支持向量机的半监督学习

转导支持向量机 TSVM 由 Vapnik 和 Sterin[22]提出,是对半监督问题的支持向量机(Support Vector Machine,SVM)扩展算法。当给定某测试集并对其进行分类时,将 SVM 和转导推理相结合得出转导支持向量机 TSVM,其可以充分利用未标记的数据,有着比传统的 SVM 更为突出的性能表现[23]。TSVM 可以利用转导学习的核心思想,采用局部搜索的策略来进行迭代求解,同时用已标记样本和未标记样本两种数据来确定最优分类边界,找到最大边缘,使超平面最大化地去分离原来已有的数据,这样不仅可以最小化误差,还可以满足原未标记样本数据的分类。

2.2.3 基于图的半监督学习

基于图的半监督学习在近些年得到了快速发展,主要思想为:构建样本图(其中顶点代表样本),定义出边的关系(边的权值代表两点之间的相似程度),用有效的能量函数作为评

价标准。最小分割算法和流行算法是两种典型的基于图的半监督学习算法。最小分割算法由 Blum 和 Chawla[24] 首次提出,源结点是正标记样本实例,目标结点是负标记样本实例,找到一组删除后可以将源结点和目标结点分割开的边,这组边就为图割,图也被分割成两个独立的部分。Belkin 等[25] 归纳总结了流形半监督学习方法,并提出了流行正规化法。

2.2.4　协同训练

最初的协同训练方法由 Blum 和 Mitchell[26] 提出。Goldman[27] 提出了不需要充分冗余视图的改进算法,该算法需要对两个分类器中的未标记数据进行可信度估计,并根据可信度进行分类。

表 1 详细对比分析了上述 4 种半监督学习方法。

表 1　4 种半监督学习分类方法的对比

Table 1　Comparison of four classification methods of semi-supervised learning

半监督学习分类方法	算法介绍	优点	缺点
生成式半监督学习	把未标记样本的每个类别看作一组缺失参数,然后对生成式模型的参数进行最大似然估计	简单直观,当有标记样本极少时能够取得更好的性能	需要充分的先验知识;当模型假设与数据分布不一致时,使用大量未标记数据容易降低模型的泛化能力
半监督 SVM	利用有标记数据训练 SVM 并估计未标记数据的标记。迭代式使得间隔最大化,并更新预测模型	比传统 SVM 得到的结果更优,能应对高维数据	损失函数非凸,会陷入局部最小点,从而影响泛化能力
图的半监督学习	利用有标记和未标记数据构建图,按照邻接关系将标记从有标记的数据点向未标记数据点传播。传播方式分为显式标记传播和隐式标记传播	更加直观,并具有很好的解释性和良好的学习性能	需要良好的数学基础;难以对大规模未标记数据进行半监督学习
协同训练	假设数据集有两个视图,在给定训练数据集的情况下训练每个分类器,对大量的未标注数据进行标注,找出以较高置信度标注的未标注数据,把这些数据加入到训练集中不断进行迭代,直到未标注数据都被标注过	与生成式模型及 EM 算法相比,特征集足够大时,可以将特征集随机地划分成两个视图,在此基础上进行协同训练得到的效果最佳	如果初始分类器较弱,未标记数据可能被错误标记,导致引入噪声

3　半监督聚类

在现实生活中,如何利用未标记数据进行聚类一直是人们研究的焦点。半监督学习的全面发展为半监督聚类奠定了基础。在聚类和半监督学习的基础上,半监督聚类利用少量标记信息进行数据预处理,改进了传统聚类,得到了更好的结果。

3.1　半监督聚类的分类

一般情况下,半监督聚类方法分为如下 3 类。

(1)基于约束的半监督聚类 (Constraint-based Semi-supervised Clustering,CBSSC)算法。这类算法的思想特点是在传统聚类的基础上加入约束限制信息来使聚类效果达到最佳。

(2)基于距离的半监督聚类(Distance-based Semi-supervised Clustering,DBSSC)算法。这类算法的特点是在对数据进行预处理的过程中,对样本之间的相似性度量进行变换,从而得到一个新的测量函数,使得相关联的正约束样本之间更加相近而负样本则更加相反。

(3)基于约束和距离相结合的半监督聚类 (Constraint and Distance based Semi-supervised Clustering,CDBSSC)算法。这类算法是将前两种方法相结合而得到的一种新的算法,可以获得更好的聚类效果。

3.2　半监督聚类算法

3.2.1　COP-Kmeans 算法

Wagstaff 等[28] 将成对约束的思想运用到传统 K-means 算法中,提出了 Cop-Kmeans 算法。Cop-Kmeans 算法的基本聚类思想与 K-means 相同,只是在数据分配过程中要求数据对象必须满足 Must-link(ML)约束和 Cannot-Link(CL)约束条件(ML 代表被选中的两个点一定是属于同一类,而 CL 代表被选中的两个点一定不是同一类的元素),并且约束具有对称性和传递性[29]。

对称性:

$(x_i,x_j)\in ML\Rightarrow(x_j,x_i)\in ML$

$(x_i,x_j)\in CL\Rightarrow(x_j,x_i)\in CL$

传递性:

$(x_i,x_j)\in ML\&(x_j,x_k)\in ML\Rightarrow(x_i,x_k)\in ML$

$(x_i,x_j)\in ML\&(x_j,x_k)\in CL\Rightarrow(x_i,x_k)\in CL$

对称性和传递性在成对约束中至关重要,这种特性导致样本在强制分配约束关系时,只有在 CL 约束情况下才会造成约束违反,在其他情况下并不会有样本分配失败的情况。

3.2.2　改进的 LCop-Kmeans 算法

基于广度优先搜索(Breadth First Search,BFS)的 COP-Kmeans 算法,也可以命名为 LCop-Kmeans(Linked Cop-Kmeans)算法[30]。此算法主要解决了 CL 约束造成违反的问题,主要思想是:数据在迭代过程中将有约束的数据和无约束的数据区分开来,对于无约束的数据,在进行分类时可以直接将其分配到最近的聚类中,之后确定其他有约束数据之间的关系。当 ML 约束存在时,按照传递性进行分配,以保证约束不违反,并且得到分配结果。当 CL 约束存在时,采用 BFS 搜索算法,将与某顶点有关联的所有 CL 约束的数据都按顺序插入一个先进先出的 List 结构中,然后再访问与该顶点相邻的所有顶点。此方法不仅可以防止数据被重复访问,还可以保证 CL 约束不违反。

3.2.3 Seeded-Kmeans 算法

与之前介绍的成对约束算法不同，当先验知识是独立的类标签，而不再是成对的约束信息时，Basu 等[31]基于 Seeds 集改进的思想，提出 Seeded-Kmeans 算法。该算法的主要目的是将标记样本引入 K-means 中，其中标记样本可以是少量的，将其作为 Seeds 集，同时采用最大期望算法将样本划分成 K 个簇，初始化的集群中心是每个簇类的均值。

3.2.4 改进的 SC-Kmeans 算法

SC-Kmeans 算法[32]是 Seeded-Kmeans 算法的一种改进，针对如何充分利用半监督聚类算法中有效的先验知识而提出。该算法可以同时利用成对约束和独立类标签这两种监督信息，并将主动学习算法引入到 SC-Kmeans 算法中，提高了 SC-Kmeans 算法的聚类精度。在这个算法中，需要扩充成对约束，得到新的 ML 和 CL 约束集，即 New-ML 和 New-CL 约束。根据 ML 的传递性和对称性，得出新的等价类 TML-Set，其中 $TML\text{-}Set=\bigcup_{i=1}^{s} T_i$。

3.2.5 算法优缺点分析

(1)Cop-Kmeans 算法的优缺点

Cop-Kmeans 算法要求样本数据在划分过程中满足约束关系，因此针对有约束信息的样本来说，这种算法的效率极高。但是，约束信息的质量会直接影响聚类结果的好坏，因此只有获得更优的约束信息，才能提高聚类效果。

(2)LCop-Kmeans 算法的优缺点

LCop-kmeans 算法解决了之前所提到的约束违反的问题，可以让成对约束的聚类结果更加安全可靠。但是，该算法在数据样本稳定性上的表现并不突出，可能会将样本分配到不合适的簇中，从而导致最后的聚类结果不准确。

(3)Seeded-Kmeans 算法的优缺点

在数据样本的监督信息是独立的类标签的情况下，Seeded-Kmeans 算法可以对数据进行聚类，优化了聚类的结果。但是，当遇到先验知识同时包括类标签成对约束时，如果只拿出其中的某一方面来指导聚类，得出的结果则并不是很理想。

(4)SC-Kmeans 算法的优缺点

SC-Kmeans 算法可以同时利用类标签和成对约束这两种监督信息进行聚类，尤其是加入了主动学习来主动标记样本，能获得质量更高的监督信息，得到更好的聚类结果。但是，在处理大规模数据方面，此算法的效率并不是很高。

3.3 半监督聚类的研究现状

根据半监督聚类的分类，可以从以下两个方面来介绍此类研究的现状。

(1)基于约束的半监督聚类的研究现状

在基于约束的半监督聚类中，最典型、最基础的算法是 Seeded-Kmeans 算法和 Cop-Kmeans 算法。这两种算法分别从监督信息的不同角度出发，在半监督聚类的发展中发挥着至关重要的作用。Li 等[33]提出了半监督层次聚类算法，在层次聚类中也可以使用成对约束。朱煜等[34]在 LCop-Kmeans 算法的基础上提出了改进的基于广度优先搜索的 Cop-Kmeans 算法，该算法提高了数据稳定性，使得聚类结果更加准确。文献[35]证明半监督聚类算法可以在顶点低层和组件高层两个部分随机游走。低层随机游走时得到的约束顶点可以对其他顶点的约束能力进行估测，然后在高层随机游走中传播约束，最终将其归结在一个簇里。文献[36]提出了一种基于 AP 算法的半监督聚类方法，间接地使聚类效果得到优化。Yang 等[37]利用 MapReduce 对 Cop-Kmeans 算法进行改进，让大规模的数据集可以并行运算。李晁铭等[38]提出了基于成对约束的交叉熵半监督聚类算法，这种方法利用样本的交叉熵来表达成对约束信息，能够在成对约束信息较少的情况下获得较高的聚类精度和较优的结果。柴变芳等[39]提出了主动学习先验的半监督 K-means 聚类算法来补充监督信息。该方法将主动学习加入到约束半监督聚类中，可以主动选择每个聚类中最有用的无标签数据样本。相比传统的半监督聚类算法，这种方法的效率更高，并且得到的结果更佳。除此之外，半监督聚类还可以与马尔可夫模型相结合。例如，Basu 等[40]提出了一种基于隐马尔可夫随机场的成对约束的半监督聚类；Jia 等[41]提出了一种基于 HMRF 模型的监督近似谱聚类算法。

(2)基于距离的半监督聚类的研究现状

在基于距离的半监督聚类中，Alok 等[42]将特征选择和半监督聚类相结合，解决了无监督分类框架下的特征选择问题。相比基于多目标的自动聚类技术，单个客观聚类技术和传统 K-means 聚类能够从具有点对称聚类的数据集中检测出适当的特征组合和划分。Wei 等[43]提出了基于成对约束与度量的半监督聚类集合，在成对约束标记的数据情况下，分别使用基于约束的半监督聚类和基于度量的半监督聚类来生成不同的基础聚类分区，然后通过积分得到目标聚类。这种方法能够提高聚类精度。

4 应用

半监督聚类已经逐渐成为机器学习等领域的研究热点，能够解决实际问题，涉及道路检测、图像、分类、信息检索、语音识别、生物信息学等多个领域。下面分不同领域来介绍半监督聚类在近些年的应用情况。

(1)在生物信息学领域内，如在癌症诊断方面，Saha 等[44]将半监督聚类应用于基因表达数据集的划分中，解决了癌症组织的分类问题；尤其针对 3 种开源基准癌症数据集进行了评估，最终在合理的时间范围内得到了分类方案。文献[45]将半监督聚类集成应用在了生物分子模式挖掘中，通过聚类可以将庞大的癌症数据量进行分类，但是在癌症的数据集中，还可能会出现已知的癌症诊断信息，包括专家意见指导、相关癌症病情等。若想癌症诊断得更为精准，就需要运用这些背景知识，由此提出了两种半监督聚类集成框架。此框架可以有效去除数据中的噪声和无关数据，从而得到精准的潜在关系，为治疗癌症提供帮助。文献[46]则是将半监督聚类运用到生物信息学中，对提取到的心电图的特征进行聚类，由此得到的结果简单直接，训练时间也更短。

(2)在图像处理领域,半监督聚类得到了广泛的应用。图像分割结果的好坏将影响后续的工作和分析。图像分割的本质是像素的聚类问题,即按照像素之间的相似性进行划分。安强强等[47]针对图像分割不清晰等问题进行了研究,通过半监督 K-means 算法进行图像分割,对少数的独立像素点进行标记,并运用 Seeds 集进行聚类。而李巧兰[48]则是在成对约束的监督信息的基础上加入 Seeds 集来进行图像分割,这样可以避免算法陷入局部最优,并且提升了图像分割的效率。在面对含噪图像时,文献[49]提出了鲁棒半监督聚类分割算法,即在数据预处理阶段先进行引导滤波预处理,之后再进行半监督约束,这种做法可以使含噪情况下的图像分割效果不受影响。在处理不清晰的卫星图片时,Alok 等[50]将卫星图像的像素划分为均匀区域,并按照半监督聚类对卫星图像进行分割。

(3)在计算机其他领域,Fiore 等[51]提出使用受限的玻尔兹曼机进行网络的异常检测,梁辰等[52]提出新的半监督入侵检测方法。在其他相关领域,彭太乐等[53]在微视频方向上也结合了半监督聚类;Zhong 等[54]在文档聚类上也有涉及半监督聚类;程雪梅等[55]提出基于半监督聚类方法的测试用例选择技术,针对软件回归测试中修改后的大量用例集,将回归测试用例选择技术与半监督聚类相结合,从而得出判别型半监督 K-means 聚类算法(Discriminative Semi-supervised K-means clustering Method,DSKM)。DSKM 依旧是运用了背景知识中的成对约束信息和标签数据,通过半监督聚类得到了更加优化的测试结果,提高了回归测试的效率。

5　未来展望

未来,半监督聚类算法可以从以下角度进行进一步的研究。

(1)基于不同的聚类思想,可以形成不同的半监督聚类算法,如半监督层次聚类、半监督密度聚类和半监督谱聚类等。

(2)可以尝试从不同的角度改进半监督聚类,如减少数据噪声、提高精度和充分利用先验知识等方法。

(3)与多个领域进行融合。在不同领域内运用半监督聚类算法的思想,加入不同领域的知识,可以得到更加优化的效果。

(4)现在的数据量是十分庞大的,在遇上大规模数据,同时还是高维数据时,如何高效率地解决大规模高维数据,是半监督聚类算法需要面对的。

(5)无论是半监督聚类,还是传统聚类,甚至是其他方向的算法,若想保证输出的最终结果最优,对数据做预处理是相当重要的一个环节。

(6)半监督聚类的评价准则十分重要,未来可以探索合理的公式以及相似性评价准则来判断算法的优劣性。

(7)在成对约束的半监督聚类中,如何避免约束违反,从而达到更加安全、可靠的结果,也是研究的重中之重。

结束语　本文从半监督学习和半监督聚类两个角度,介绍了半监督学习及其分类的优劣对比,更好地阐释了半监督聚类,对半监督聚类的典型算法进行了描述,使初学者从理论上学习半监督聚类,为之后的实践做准备。未来可以在高维数据的处理和约束违反的方向上,对半监督聚类进行改进。

参 考 文 献

[1] HARTIGAN J A,WONG M A. Algorithm AS 136:A k-means clustering algorithm[J]. Applied Statistics,1979,28(1):100-108.

[2] MADDAH M,CRIMSON W E L,WARFIELD S K. Statistical modeling and EM clustering of white matter fiber tracts[C]// 3rd IEEE International Symposium on Biomedical Imaging:Nano to Macro. New York:IEEE Press,2006:53-56.

[3] LI K L,CAO Z,CAO L P,et al. Some Developments on Semi-Supervised Clustering [J]. Pattern Recognition and Artificial Intelligence,2009,22(5):735-742. (in Chinese)
李昆仑,曹铮,曹丽苹,等. 半监督聚类的若干新进展[J]. 模式识别与人工智能,2009,22(5):735-742.

[4] XIONG J B,LI Z K,LIU Y J. Research on the Present Situation of Semi-Supervised Clustering Algorithm[J]. Modern Computer,2009(12):61-64,77. (in Chinese)
熊建斌,李振坤,刘怡俊. 半监督聚类算法研究现状[J]. 现代计算机(专业版),2009(12):61-64,77.

[5] LIU J W,LIU Y,LUO X L. Semi-Supervised Learning Methods [J]. Chinese Journal of Computers,2015,38(8):1592-1617. (in Chinese)
刘建伟,刘媛,罗雄麟. 半监督学习方法[J]. 计算机学报,2015,38(8):1592-1617.

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

[7] FRALICK S. Learning to recognize patterns without a teacher [J]. IEEE Transactions on Information Theory,2003,13(1):57-64.

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

[9] MERZ C J,CLAIR D C,BOND W E. Semi-supervised adaptive resonance theory (SMART2)[C]// International Joint Conference on Neural Networks. Baltimore:IEEE Press,1992:851-856.

[10] SHAHSHAHANI B M,LANDGREBE D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience & Remote Sensing,1994,32(5):1087-1095.

[11] KINGMA D P,REZENDE D J,MOHAME D S. Semi-Supervised Learning with Deep Generative Models[J]. Advances in Neural Information Processing Systems,2014,4:3581-3589.

[12] 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[C]// Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,2002:307-314.

[13] CHENG S,SHI Y,QIN Q. Particle swarm optimization based semi-supervised learning on Chinese text categorization[C]//

IEEE Congress on Evolutionary Computation. New York:IEEE Press,2012:1-8.

[14] WANG J,KUMAR S,CHANG S F. Semi-supervised hashing for scalable image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition. San Francisco: DBLP, 2010: 3424-3431.

[15] CHEN S G,ZHANG D Q. Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods [J]. Journal of software,2011,22(1):28-43. (in Chinese)
陈诗国,张道强. 半监督降维方法的实验比较[J]. 软件学报,2011,22(1):28-43.

[16] ZHOU Z H,LI M. Semi-supervised regression with co-training [C]// International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 2005: 908-913.

[17] MEHRKANOON S,ALZATE C,MALL R,et al. Multiclass Semi-supervised Learning Based Upon Kernel Spectral Clustering[J]. IEEE Transactions on Neural Networks & Learning Systems,2015,26(4):720.

[18] CALLUT J,FRANCOISSE K,SAERENS M,et al. Semi-supervised Classification from Discriminative Random Walks[M]// ECML PKDD 2008. Berlin:Springer,2008:162-177.

[19] 周志华. Machine learning[M]. 北京:清华大学出版社,2016.

[20] COZMAN F G,COHEN I. Unlabeled Data Can Degrade Classification Performance of Generative Classifiers[C]// Fifteenth International Florida Artificial Intelligence Society Conference. California:AAAI Press,2009:327-331.

[21] CASTELLI V,COVER T M. On the exponential value of labeled samples[M]. Elsevier Science Inc,1995.

[22] VAPNIK V,STERIN A. On structural risk minimization or overall risk in a problem of pattern recognition[J]. Automation & Remote Control,1977,10(10):1495-1503.

[23] WANG X,YU H. How to Break MD5 and Other Hash Functions[M]// Advances in Cryptology-EUROCRYPT 2005. DBLP,2005:19-35.

[24] BLUM A,CHAWLA S. Learning from Labeled and Unlabeled Data using Graph Min-cuts[C]// Eighteenth International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,2001:19-26.

[25] BELKIN M,NIYOGE P,SINDHWANI V. Manifold Regularization:A Geometric Framework for Learning from Labeled and Unlabeled Examples[J]. Journal of Machine Learning Research, 2006,7(1):2399-2434.

[26] BLUM A,MITCHELL T. Combining labeled and unlabeled data with co-training[C]// Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT98). Wisconsin,MI,1998:92-100.

[27] COLDMAN S,ZHOU Y. Enhancing supervised learning with unlabeled data[C]// Proceedings of the 17th International Conference on Machine Learning(ICML'00). San Francisco: CA, 2000:327-334.

[28] WAGSTAFF K,CARDIE C. Clustering with instance-level constraints[C]// Proceedings of 17th International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,2000:1097-1103.

[29] WAGSTAFF K,CARDIE C,ROGERS S,et al. Constrained K-means Clustering with Background Knowledge[C]// Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc,2001:577-584.

[30] YANG Y,TAN W,LI T. Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems[J]. Knowledge Based Systems,2012,32:101-115.

[31] BASU S,BANERJEE A,MOONEY R. Semi-Supervised Clustering by Seeding[C]// Proceedings of 19th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc,2002:19-26.

[32] CHEN Z Y,WANG H J,HU M,et al. An active semi-supervised clustering algorithm based on seeds set and pairwise constraints[J]. Journal of Jilin University(Science Edition),2017, 55(3):664-672. (in Chinese)
陈志雨,王慧君,胡明,等. 一种基于 Seeds 集和成对约束的主动半监督聚类算法[J]. 吉林大学学报(理学版),2017,55(3):664-672.

[33] ZHENG L,LI T. Semi-supervised Hierarchical Clustering[C]// International Conference on Data Mining. 2011.

[34] ZHU Y,QIAN J H,JI Z B. An improved COP-Kmeans algorithm based on BFS[EB/OL]. Beijing: China science and technology paper online [2015-07-09]. http://www.paper.edu.cn/releasepaper/content/201507-93. (in Chinese)
朱煜,钱景辉,季正波. 改进的基于广度优先搜索的 COP-Kmeans 算法[EB/OL]. 北京:中国科技论文在线 [2015-07-09]. http://www.paper.edu.cn/releasepaper/content/201507-93.

[35] HE P,XU X H,LU L,et al. Semi-Supervised Clustering via Two-Level Random Walk [J]. Journal of Software, 2014, 25(5):997-1013. (in Chinese)
何萍,徐晓华,陆林,等. 双层随机游走半监督聚类[J]. 软件学报,2014,25(5):997-1013.

[36] TANG Q,LIAO Z G. A Semi-Supervised Clustering Method Based on Affinity Propagation Algorithm[J]. Electronic Information Warfare Technology,2017,32(1):8-12. (in Chinese)
汤琼,廖泽广. 一种基于 AP 算法的半监督聚类方法[J]. 电子信息对抗技术,2017,32(1):8-12.

[37] YANG Y,RUTAYISIRE T,LIN C,et al. An Improved Cop-Kmeans Clustering for Solving Constraint Violation Based on MapReduce Framework [J]. Fundamental Information, 2013, 126(4):301-318.

[38] LI C M,XU S B,HAO Z F. Cross-Entropy semi-supervised clustering based on pairwise constraints[J]. Pattern Recognition and Artificial Intelligence,2017,30(7):598-608. (in Chinese)
李晁铭,徐圣兵,郝志峰. 基于成对约束的交叉熵半监督聚类算法[J]. 模式识别与人工智能,2017,30(7):598-608.

[39] CHAI B F,LV F,LI W B,et al. Semi-supervised Kmeans Clustering Algorithm based on Active Learning Priors [J/OL].

[2018-11-25]. http://kns. cnki. net/kcms/detail/51. 1307. TP. 20180719. 1030. 038. html. (in Chinese)
柴变芳,吕峰,李文斌,等. 基于主动学习先验的半监督 Kmeans 聚类算法[J/OL]. [2018-11-25]. http://kns. cnki. net/kcms/detail/51. 1307. TP. 20180719. 1030. 038. html.

[40] BASU S, BILENKO M, MOONEY R J. A probabilistic framework for semi-supervised clustering[C]// Proceedings of the Tenth ACM 0SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-04). New York: MIT Press, 2008:59-68.

[41] DING S, JIA H, DU M, et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model[J]. Information Sciences, 2018, 429:215-228.

[42] ALOK A K, SAHA S, EKBAL A. Feature Selection and Semi-supervised Clustering Using Multi-objective Optimization[J]. Springer Plus, 2014, 3(1):465.

[43] WEI S, LI Z, ZHANG C. Combined constraint-based with metric-based in semi-supervised clustering ensemble[J]. International Journal of Machine Learning & Cybernetics, 2018, 9(7): 1085-1100.

[44] SAHA S, KAUSHIK K, ALOK A K, et al. Multi-objective semi-supervised clustering of tissue samples for cancer diagnosis[J]. Soft Computing, 2016, 20(9):3381-3392.

[45] CHEN H S. Semi-supervised clustering ensemble for bio-molecular pattern mining[D]. Guangzhou: South China University of Technology, 2016. (in Chinese)
陈弘晟. 半监督聚类集成在生物分子模式挖掘中的应用[D]. 广州:华南理工大学,2016.

[46] OROZCO-DUQUE A, BUSTAMANTE J, CASTELLANOS-DOMINGUEZ G. Semi-supervised clustering of fractionated electrograms for electroanatomical atrial mapping[J]. Biomedical Engineering Online, 2016, 15(1):44.

[47] AN Q Q, ZHANG F, LI Z X, et al. Research on image segmentation based on machine learning[J]. Automation & Instrumentation, 2018(6):29-31. (in Chinese)
安强强,张峰,李赵兴,等. 基于机器学习的图像分割研究[J]. 自动化与仪器仪表,2018(6):29-31.

[48] LI Q L. Semi-supervised clustering based on constraints for images segmentation[D]. Xi'an: XiDian University, 2014. (in Chinese)
李巧兰. 基于约束的半监督聚类的图像分割算法研究[D]. 西安:西安电子科技大学,2014.

[49] LI Y W. Research on robust segmentation algorithm based on semi-supervised fuzzy clustering[D]. Xi'an: Xi'an University of Posts & Telecommunications, 2018. (in Chinese)
李亚文. 鲁棒半监督模糊聚类分割算法研究[D]. 西安:西安邮电大学,2018.

[50] ALOK A K, SAHA S, EKBAL A. Multi-objective semi-supervised clustering for automatic pixel classification from remote sensing imagery[J]. Soft Computing, 2016, 20(12):4733-4751.

[51] FIORE U, PALMIERI F, CASTIGLIONE A, et al. Network anomaly detection with the restricted Boltzmann machine[J]. Neuro Computing, 2013, 122:13-23.

[52] LIANG C, LI C H. Novel Intrusion Detection Method Based on Semi-supervised Clustering[J]. Computer Science, 2016, 43(5): 87-90. (in Chinese)
梁辰,李成海. 一种新的半监督入侵检测方法[J]. 计算机科学,2016,43(5):87-90.

[53] PENG T L, ZHANG W J, LAN J L, et al. Micro video annotation method based on semi-supervised clustering[J]. Application Research of Computers, 2016, 33(3):948-952. (in Chinese)
彭太乐,张文俊,蓝建梁,等. 基于半监督聚类的微视频标注方法[J]. 计算机应用研究,2016,33(3):948-952.

[54] ZHONG S. Semi-supervised model-based document clustering: A comparative study[J]. Machine Learning, 2006, 65(1):3-29.

[55] CHENG X M, YANG Q H, ZHAI Y P, et al. Test Case Selection Technique Base on Semi-supervised Clustering Method[J]. Computer Science, 2018, 45(1):249-254. (in Chinese)
程雪梅,杨秋辉,翟宇鹏,等. 基于半监督聚类方法的测试用例选择技术[J]. 计算机科学,2018,45(1):249-254.