软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):981−990 [doi: 10.13328/j.cnki.jos.005919]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于 *k* 个标记样本的弱监督学习框架*

付 治[1,2], 王红军[1,2], 李天瑞[1,2], 滕 飞[1,2], 张 继[1,2]

[1](西南交通大学 信息科学与技术学院,四川 成都 611756)

[2](综合交通大数据应用技术国家工程实验室(西南交通大学),四川 成都 611756)

通讯作者: 王红军, E-mail: wanghongjun@swjtu.edu.cn

摘 要: 聚类是机器学习领域中的一个研究热点,弱监督学习是半监督学习中一个重要的研究方向,有广泛的应用场景.在对聚类与弱监督学习的研究中,提出了一种基于 *k* 个标记样本的弱监督学习框架.该框架首先用聚类及聚类置信度实现了标记样本的扩展.其次,对受限玻尔兹曼机的能量函数进行改进,提出了基于 *k* 个标记样本的受限玻尔兹曼机学习模型.最后,完成了对该模型的推理并设计相关算法.为了完成对该框架和模型的检验,选择公开的数据集进行对比实验,实验结果表明,基于 *k* 个标记样本的弱监督学习框架实验效果较好.

关键词: 机器学习;弱监督学习;聚类

中图法分类号: TP181

中文引用格式: 付治,王红军,李天瑞,滕飞,张继.基于 *k* 个标记样本的弱监督学习框架.软件学报,2020,31(4):981−990. http://www.jos.org.cn/1000-9825/5919.htm

英文引用格式: Fu Z, Wang HJ, Li TR, Teng F, Zhang J. Weakly supervised learning framework based on *k* labeled samples. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):981−990 (in Chinese). http://www.jos.org.cn/1000-9825/5919.htm

Weakly Supervised Learning Framework Based on *k* Labeled Samples

FU Zhi[1,2], WANG Hong-Jun[1,2], LI Tian-Rui[1,2], TENG Fei[1,2], ZHANG Ji[1,2]

[1](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

[2](National Engineering Laboratory of Integrated Transportation Big Data Application Technology (Southwest Jiaotong University), Chengdu 611756, China)

Abstract: Clustering is an active research topic in the field of machine learning. Weakly supervised learning is an important research direction in semi-supervised learning, which has wide range of application scenarios. In the research of clustering and weakly supervised learning, it is proposed that a framework of weakly supervised learning is based on *k* labeled samples. Firstly, the framework expands labeled samples by clustering and clustering confidence level. Secondly, the energy function of the restricted Boltzmann machine is improved, and a learning model of the restricted Boltzmann machine based on *k* labeled samples is proposed. Finally, the model of ratiocination and algorithm are proposed. In order to test the framework and the model, a series of public data sets are chosen for comparative experiments. The experimental results show that the proposed weakly supervised learning framework based on *k* labeled samples is more effective.

Key words: machine learning; weakly supervised learning; clustering model

在机器学习中,无监督学习与有监督学习有很大的区别.判断的关键在于区分给定的训练样本是否包含了样本的类别信息.其中,无监督学习的目标是:发现样本数据潜在的结构或者规律[1,2].它不包含任何先验知识,学

* 基金项目: 四川省国际科技创新合作重点项目(2019YFH0097)

Foundation item: Key Program for Int'l S&T Cooperation of Sichuan Province of China (2019YFH0097)

本文由"非经典条件下的机器学习方法"专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-03-10; 修改时间: 2019-07-11; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-17 14:06:32, http://kns.cnki.net/kcms/detail/11.2560.TP.20200117.1405.002.html

习过程也没有指导者,完全靠自身进行学习、归纳、总结;而在有监督学习中,又包含了强监督学习(即监督学习)、半监督学习[3−6],它们之间区别的依据在于给定学习样本中含有样本的类别信息所占的比重.然而,在很多任务中,由于数据标注过程的高昂代价,很难获得强监督信息.因此,弱监督学习一直是研究者未来研究的方向.本文在非经典条件下机器学习的研究中,提出了基于 k 个标记样本的弱监督学习框架.

弱监督学习是机器学习领域的重要研究热点[5−7],其目的是同时利用少量有标签的样本与无标签的样本训练学习模型,使学习得到的模型或者模型参数更准确.当前已提出的弱监督学习主要与分类和聚类两个方面相结合[5,6].弱监督学习分类是分类中比较理想的情况.Scudder[8]提出的自训练方法是将无标签的样本用于监督学习的方法.Vapnik 和 Sterin[9]提出了转导支持向量机,用于估计类标签的线性预测函数.Blum 和 Mitchell[10]提出了一种协同训练的方法,基于不同的视图训练两个完全不同的学习机,提高了训练样本数据的置信度.Zhou 和 Goldman[11]提出了一种协同训练改进算法,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.Chapelle 等人[12]提出直推学习的概念和生成式模型,其中,假设生成数据的概率密度函数为多项式分布,并基于该假设同时利用有类标签的样本和无类标签的样例进行学习,以估计该模型中的参数.这些弱监督学习主要是与分类相结合,希望用“较少”的带标签的数据来训练分类模型,但这个“较少”并没有被量化.

弱监督聚类问题也是弱监督学习中常见的一类问题.Klein 等人提出弱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[13].王玲等人提出一种基于密度敏感的弱监督谱聚类算法,使用两类先验信息在指导聚类搜索的过程中起到相辅相成的作用,相对于仅利用成对限制信息的聚类算法,在聚类性能上有了显著的提高[14].高滢等人提出了一种弱监督 k 均值多关系数据聚类算法[15].王娜等人提出了一种基于监督信息特性的主动学习策略[16].Wang 等人提出一种成对约束邻域投影的方法进行弱监督聚类学习[17].Xiong 等人研究了弱监督聚类中的成对约束的主动学习问题[18].Huang 等人扩展了一种基于流形正则化的弱监督任务和无监督任务的极限学习机,从而极大地扩展了极限学习机的适用性[19].Portela 等人将弱监督聚类应用到不同解剖结构或组织类型的磁共振脑图像分割上[20].Lad 和 Parikh 使用基于图像本身属性的推理作为弱监督图像聚类约束的获取方式[21].Anand 等人提出了一个只使用成对约束来指导聚类过程的弱监督核均值漂移聚类框架[22].Lai 等人引入了一个新的交互式弱监督聚类模型,其中的先验信息通过图像之间的成对约束进行整合[23].Ding 等人提出了一种有效的基于成对约束的弱监督谱聚类算法,其中,数据点的相似度矩阵通过成对约束来调整和优化[24].Huang 等人对成对约束投影进行协同扩展,提出了协同成对约束投影的弱监督协同聚类方法[25].Yi 等人提出了一个高效的动态弱监督聚类框架,将聚类问题转化为一个可行的凸集上的搜索问题[26].Triguero 等人[27]尝试使用基于多目标优化的弱监督分类技术来解决基因表达数据聚类问题.Mehrkanoon 等人介绍了一种在线弱监督学习算法,作为正则化核心谱聚类方法[28].Honda 等人提出了一种由多项式混合概念引入的用于对共生信息进行局部监督的模糊协同聚类的新框架[29].Nie 等人提出一种新的框架对标准的谱学习模型进行重构,可以进行多视图的聚类或弱监督学习[30].Wang 等人提出了约束传播的弱监督非负矩阵分解方法[31].Saha 等人提出了一种利用多目标优化概念的弱监督聚类技术,并将该技术应用于强度空间中脑核磁共振图像的自动分割[32].Yu 等人提出了一种使用弱监督聚类进行分类的方法来改善文本分类[33].Fang 等人提出了一种鲁棒弱监督子空间聚类方法以解决低秩表示的弱监督子空间聚类方法中标签信息并不用于指导亲和矩阵构造的问题[34].Tuan 和 Ngan 针对牙科 X 射线图像分割问题,提出一种基于交互式模糊方法的弱监督模糊聚类算法[35].Tuan 等人提出了一种将弱监督模糊聚类算法应用于牙科 X 射线图像分割的新型协作方案[36].Alok 等人利用弱监督聚类的概念以尝试解决将卫星图像的像素分类为均匀区域的问题[37].Nie 等人提出一种基于自适应邻居点学习的方法,可以同时进行聚类、数据的局部结构学习及弱监督分类[38].Yang 等人提出了一种基于多密度信息的自适应弱监督聚类方法[39].Li 等人提出了一种基于弱监督聚类算法的云环境下的三维人脸识别方法[40].这些弱监督聚类方法相比完全无监督聚类算法,具有一定的优势,但与弱监督分类一样,很少有文献研究弱监督的标签数量需要的下界,也没有文献对之进行证明.

上面分别对弱监督分类和聚类进行了阐述,尽管对弱监督各方面的方法和模型进行了大量的研究和探索,但对弱监督学习的标签数量需要的下界的问题并没有进行论证.而如今的深度学习[41]基本上都需要含有真值

标签的大规模训练数据集,很多任务都很难获得标签信息以训练模型,因此也需要对弱监督深度学习进行探索与研究.本文结合目前的研究现状,探讨了弱监督领域中需要的带标签的训练数据数量的下界,对这个下界进行了证明.同时,结合深度学习,对受限玻尔兹曼机进行改进,使其能够进行弱监督学习.

本文的主要贡献如下.

(1) 提出了一种基于 k 个标记样本的弱监督学习框架,只需极少数的样本标记,用聚类及聚类置信度实现标记样本的扩展.相比于无监督学习,具有一定的优势,也解决了半监督学习中没有足够的先验知识的问题.

(2) 提出了一种基于 k 个标记样本的受限玻尔兹曼机学习模型,并对受限玻尔兹曼机的能量函数进行改进.

(3) 对基于 k 个标记样本的受限玻尔兹曼机学习模型进行推理、证明并设计相关算法,完成对该框架和模型的检验.

1 基于 k 个标记样本的弱监督学习框架

1.1 基于k个标记样本的弱监督学习理论分析

弱监督学习追求利用尽量少的标签去训练得到一个好的机器学习模型,但这个尽量少的下限是什么,这是一个值得探讨的理论问题.本文进行了一个初步的探讨.

命题 1. 用 $D(k)$表示一个 k 类(k 为大于 1 的自然数)的数据 D,用 $M(k)$表示与之相对应的弱监督学习模型,至少需要 k–1 数量的标记样本去训练模型 $M(k)$,使 $M(k)$可以把 $D(k)$中的数据分类正确.

使用归纳法.

(1) 当 k=2 时,用 $D(2)$表示一个 2 类的数据 D,训练一个弱监督学习模型 $M(2)$,至少需要 1 个标记样本去训练 $M(2)$,$M(2)$可以完全把 $D(2)$中的数据分类正确.如果 1 个标记样本都不需要,那么就是无监督学习模型.另外,2 分类问题,在很理想的情况下,只需要 1 个标记样本就完全可以把 2 分类问题的数据完全分类正确.

(2) 当 $2<k=n$ 时,可以把 k 分类问题分成 k–1 个 2 分类问题,由上面的(1)可知,至少需要 k–1 数量的标记样本去训练模型 $M(k)$,使 $M(k)$可以完全把 $D(k)$中的数据分类正确.

(3) 当 $k=n+1$ 时,综合上述(1)和(2)可知命题 1 成立.

在利用标记样本的弱监督学习中,需要标记样本数量的下界值是 k–1,这是把所有的样本分类正确的基本条件.基于 k 个标记样本的弱监督学习,比下界值多 1 的目的是为了更好地进行标记传播、模型建立及推理.同时,也是为了保证弱监督学习的效果.而且,也是得到这个理论下界值的指引,由 k–1 到 k,多增加 1,可以更容易地从理论到应用.基于 k 个标记样本的弱监督学习的理论是对命题 1 的继承与发展.

对命题 1 的证明实际上也间接证明了“基于 k 个标记样本的弱监督学习可以转化为强学习器”.命题 1 中的“至少需要 k–1 数量的标记样本去训练模型 $M(k)$,使 $M(k)$可以把 $D(k)$中的数据分类正确”实际上就表示,至少需要 k–1 数量的标记样本去训练模型 $M(k)$,这个 $M(k)$可能就是一个强学习器.那么,基于 k 个标记样本的弱监督学习即可以转化为强学习器.

1.2 基于k个标记样本的弱监督学习框架研究

在前一小节的分析与研究中,初步设计了基于 k 个标记样本的弱监督学习框架,如图 1 所示;然后,基于这个框架,设计了基于 k 个标记样本的受限玻尔兹曼机学习模型.本文所提框架使用受限玻尔兹曼机模型作为实例的优势主要有两个方面:第一是受限玻尔兹曼机模型对特征表示学习能力强,同时泛化能力比较好,以改进的受限玻尔兹曼机模型作为实例可以较好地证明本文所提框架;第二是对受限玻尔兹曼机模型的能量函数进行改进后,可解释性非常好,各部分的意义非常清晰.

从图 1 这个框架可知,数据首先包括 k 个标记样本和若干个无标记样本,通过标记扩展算法,对无标记样本进行标注;为了提高标注的精度,用标签置信度(本文采用的是与簇中心点距离比作为标签置信度标准)对标记的标签进行筛选,满足标签置信度的标签数据输出为有标签数据,反之则是无标记样本;用 τ表示标签置信度样本,x_i 的标签置信度如公式(1)所示.

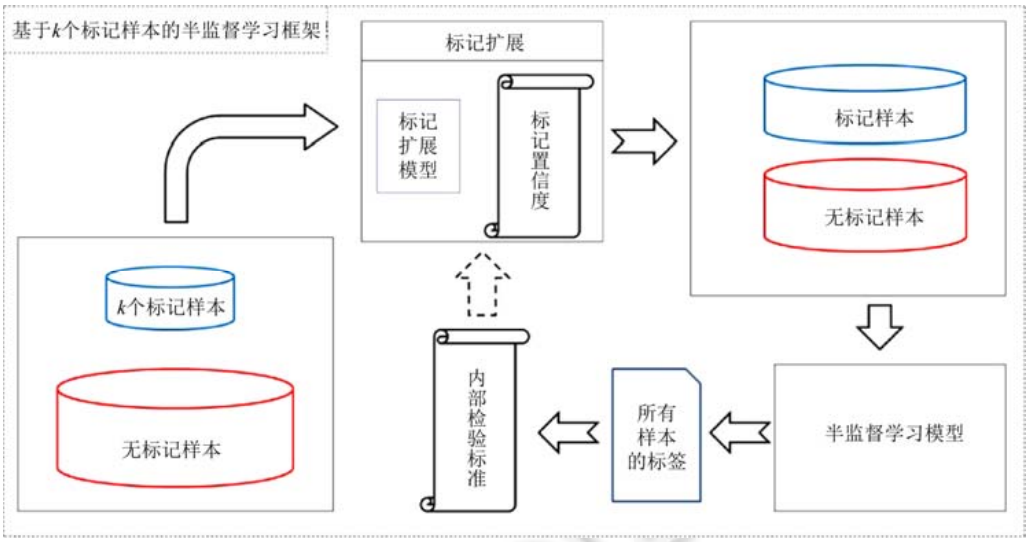

Fig.1 Weakly supervised learning framework based on k labeled samples

图 1 基于 k 个标记样本的弱监督学习框架图

$$\tau = 1 - \frac{\left\|x_i - c_j\right\|^2}{\max\left(\left\|x_q - c_j\right\|^2\right)} \quad \text{s.t.} x_i \in C_j, x_q \in C_j \tag{1}$$

其中,x_i 与 x_q 的类中心是 C_j,该式分母表示类簇中心 C_j 和与距其最远的样本点 x_q 之间的距离,该式表示,与类簇中心点的距离越近,该数据的标签置信度越高.在此框架中,为了避免 k 个标记样本中任意两个样本在聚类算法中被分配到同一类的问题,需要对聚类算法增加相应的约束改进,如在聚类算法的标签分配的步骤限制这 k 个样本任意两个分配在同一类.

通过标记拓展后即可设计弱监督学习模型对数据进行处理,输出标签结果;最后用内部检验标准(本文采用的是类内距离总和)对输出的结果进行总体上的评估,如果内部检验标准的值比上一次的值有所下降,则调整标记置信度,进行下一次的迭代;如果内部检验标准的值不下降,则该结果为基于 k 个标记样本的弱监督学习的输出结果.

命题 2. 用 $D(k)$表示一个 k 类(k 为大于 1 的自然数)的数据 D,用 $M(k)$表示与之相对应的弱监督学习模型,需要 k 数量的标记样本去训练模型 $M(k)$,且每类结果仅需标记一个标记样本,共标记 k 个样本,使 $M(k)$可以把 $D(k)$中的数据分类正确.

使用归纳法证明.

(1) 当 k=2 时,用 $D(2)$表示一个 2 类的数据 D,训练一个弱监督学习模型 $M(2)$,需要 2 个标记样本去训练 $M(2)$,$M(2)$可以完全把 $D(2)$中的数据分类正确.

(2) 当 $2<k=n$ 时,假设 $M(n)$可以把 $D(n)$中的数据分类正确.

(3) 当 $k=n+1$ 时,我们可以把数据分为 n 类数据组成的一个杂乱无序的大类和另一个有序的小类,此时,由上面的(1)可知,$M(n+1)$可以把 $D(n+1)$中的数据分类正确.

由第一数学归纳法可知,命题 2 成立.

1.3 基于k个标记样本的受限玻尔兹曼机学习模型

基于受限玻尔兹曼机的深度学习取得了很多的成功应用,本文重构基于 k 个标记样本的受限玻尔兹曼机学习模型.在对 k 个标记样本进行一定的标记传播后,建立的基于 k 个标记样本的受限玻尔兹曼机学习模型(weakly-supervised restricted Boltzmann machine,简称 W-RBM)[42−45]的目标函数如式(2)所示.

$$L(\theta,\mathrm{V})=-\frac{\mu}{n}\sum_{v_i\in V}\log p(v_i,\theta)+\left[\left(\frac{1-\mu}{N_M}\sum_{same}\left\|h_sW^T-h_tW^T\right\|^2-\frac{1-\mu}{N_C}\sum_{Diff}\left\|h_sW^T-h_tW^T\right\|^2\right)\right] \tag{2}$$

其中,$\theta=\{a,b,W\}$是模型参数,该目标函数包含两项,第 1 项是原始受限玻尔兹曼机的目标函数,第 2 项中的 N_M 和 N_C 分别是标记扩展后同类 *Same* 集合和标记扩展后异类 *Diff* 集合的数量,让 *Same* 集合中的数据尽量靠近,同时让 *Diff* 集合中的数据尽量远离,以此达到让数据更加可分的目的,提高弱监督的学习性能.总体模型训练过程中想要达到的效果是在数据尽量保持原编码不变的情况下,让编码后的数据能够更好地区分类别之间的界限.模型是一个二目标优化问题,第 1 个目标采用快速对比散度(contrastive divergence)算法,第 2 个目标采用随机梯度下降法,所以下面主要的任务是求解第 2 目标的梯度.

求解如下.

令:

$$J_M(W)=\frac{1-\mu}{N_M}\sum_{same}||h_sW^T-h_tW^T||^2 \tag{3}$$

$$J_C(W)=\frac{1-\mu}{N_C}\sum_{Diff}||h_sW^T-h_tW^T||^2 \tag{4}$$

则

$$\frac{\partial J_M(W)}{\partial w_{ij}}=\frac{1-\mu}{N_M}\sum_{Same}\left[(h_s-h_t)W^T(h')^T+h'W\left((h_s-h_t)\right)^T\right] \tag{5}$$

$$\frac{\partial J_C(W)}{\partial w_{ij}}=\frac{1-\mu}{N_C}\sum_{Diff}\left[(h_s-h_t)W^T(h')^T+h'W\left((h_s-h_t)\right)^T\right] \tag{6}$$

其中,$h'=(h_{s1}-h_{t1},\ldots,h_{sj}-h_{tj},\ldots,h_{sq}-h_{tq})$,且 $h_{sk}-h_{sk}=0,k\neq j$.那么最终的更新规则如下:

$$\begin{aligned}w_{ij}^{(\tau+1)}=w_{ij}^{(\tau)}+\mu\varepsilon\left(\langle v_ih_j\rangle_{data}-\langle v_ih_j\rangle_{recon}\right)+\frac{1-\mu}{N_M}\sum_{Same}\left[(h_s-h_t)W^T(h')^T+h'W\left((h_s-h_t)\right)^T\right]-\\ \frac{1-\mu}{N_c}\sum_{Diff}\left[(h_s-h_t)W^T(h')^T+h'W\left((h_s-h_t)\right)^T\right]\end{aligned} \tag{7}$$

单层 W-RBM 数据的输入输出可以通过上述推理进行迭代.

基于 k 个标记样本的弱监督深度学习模型是用多层 W-RBM 进行迭代,可以根据实际问题而进行多层 W-RBM 的构造,这样建立的弱监督深度学习模型具有良好的扩展性.

1.4 基于*k*个标记样本的算法框架描述

根据图 1 及 W-RBM 模型的求解过程,本文设计的算法框架见表 1.该算法框架中的置信度是一个比值,是占数据样本数量的比值,这个比值可确定标记拓展的数量;步骤 1 表示可以使用满足条件的聚类算法对数据进行聚类,如果使用 *k*-means 算法,需要对 *k*-means 分配各数据标签时加一个 *k* 个标记样本中任何两个不能在同一类的约束即可.

Table 1 Weakly supervised learning framework structure based on *k* labeled samples

表 1 基于 *k* 个标记样本的弱监督学习框架结构

算法.基于 *k* 个标记样本的弱监督学习框架结构.
输入:数据集 *D*、类别数量 *k*、置信度 $\tau\in[0,1]$、*k* 个标记样本;
1.使用基于简单约束的(*k* 个标记样本中任何两个不能在同一类)聚类算法对数据进行聚类,根据 *k* 个标记样本和标记置信度,拓展标记样本;或者直接在数据集中找出 *k* 个标记样本的邻居拓展标记样本,邻居的数量根据置信度来选取.
2.循环开始
3. 运行基于 *k* 个标记样本的弱监督算法 W-RBM;
4. 学习并标记所有标记样本;
5. 对数据集进行学习得到新的标记样本,将当前标记样本更新为新的标记样本,重复步骤 3~步骤 5,直到上次循环与本次循环得到的内部检验标准的值不再下降;
6.循环结束
输出:各数据对象类别标签.

2 实验研究

为了验证本文所提方法的有效性,采取对各种算法进行数值实验对比的方法.实验数据选取各类数据集中具有代表性的数据以及几个常用的数据集进行算法比较.为了保证实验的可靠性与准确性,本文采取十折交叉验证,即:将数据随机等分为 10 份,依次选取 1 份作为测试数据,剩余 9 份作为训练数据,对每个数据集,各种算法都要进行 10 次训练和测试.本文使用这种方法优化参数,每次训练和测试对于不同数据相比于不同的算法均进行多次训练和测试,取测试结果的平均值.运行环境为 Windows 10 操作系统,24G 内存,i7-4960X CPU 和 GeForce GTX 1080 显卡的 PC 机.全部算法均使用 Matlab 2018a 来实现和运行.在实验中,已知 k 个标记样本的给定方法为随机选取,其中置信度参数 τ 设置为 0.8,在实验中参数标记置信度的变化步长为 0.01,从 1 向 0 递减.

2.1 实验设计

本节实验选择了 3 种算法作为对比算法,分别是 k-means[46,47]、Affinity Propagation(AP)[48]和 Density Peaks (DP)[49].将每种算法结果的准确率以及纯度计算出来,作为评价受限玻尔兹曼机模型的有效性指标.

2.2 评价指标

为了分析比较基于 k 个标记样本的受限玻尔兹曼机模型的性能,本文将 k-means、Affinity Propagation(AP)算法和 Density Peaks(DP)作为对比算法,并将每个算法结果的准确率和纯度计算出来,作为评价每种算法的有效性指标.准确率定义如下:

$$Accuracy=\frac{1}{n}\sum_{i=1}^{K}\mathrm{Max}_{1\leqslant j\leqslant q}n_k^j \tag{8}$$

其中,m 表示样本数,C_i 表示第 i 个簇,L_j 表示第 j 个类别,$T(C_i,L_j)$表示属于类别 j 的数据点被分配给簇 i 的数量.准确率越大,则表示聚类效果越好.

纯度是集群质量的透明外部评估度量,它测量每个集群包含主要来自一个类的数据点的程度.聚类的纯度标准计算公式定义如下:

$$Purity=\frac{1}{n}\sum_{i=1}^{K}\mathrm{Max}_{1\leqslant j\leqslant q}n_k^j \tag{9}$$

其中, n_k^j 是数据样本中原本属于 j 类,但结果属于 k 类的样本数,q=K,纯度值越大,代表效果越好.

2.3 测试数据集

为了对上述基于 k 个标记样本的受限玻尔兹曼机算法的聚类效果进行评价检验,选择表 2 中所列出的数据集参与测试实验.数据集均来源于微软公开数据集[50].

Table 2 Summary of the datasets

表 2 测试数据集

数据集	样本数	样本维数	类别数
alphabet	814	892	3
aquarium	922	892	3
bed	888	892	3
blog	943	892	3
border	840	892	3
brain	891	892	3
ufo	889	899	3
ufo11	881	899	3
venus	891	899	3
video	936	899	3
voituretuning	879	899	3
weed	876	899	3

2.4 实验结果

本节采用经典聚类算法、经典聚类算法与受限玻尔兹曼机模型结合、经典聚类算法与弱监督受限玻尔兹曼机模型结合的方式,并用 *k*-means 算法、AP 算法和 DP 算法作为基本算法,与各个模型结合的方式,对数据样本进行聚类,得到聚类结果,然后将它们与原始标签对比得到,采用准确率和纯度标准进行评价.表 3 展示了每个数据集在不同算法下的准确率.

Table 3 Accuracies of different model algorithms (%)

表 3 不同模型算法准确率比较(%)

	k-means	DP	AP	*k*-means.RBM	DP.RBM	AP.RBM	*k*-means.W-RBM	DP.W-RBM	AP.W-RBM
alphabet	46.913 6	43.209 9	46.049 4	48.382 7	46.419 8	48.271 6	48.395 1	46.172 8	48.518 5
aquarium	45.164 8	41.428 6	47.802 2	48.901 1	41.758 2	46.593 4	49.120 9	44.395 6	47.033
bed	48.405 8	41.014 5	40.724 6	47.971 0	42.173 9	41.594 2	49.346 4	41.739 1	42.173 9
blog	55.842 7	42.584 3	51.910 1	53.932 6	44.606 7	54.157 3	55.955 1	43.707 8	53.932 6
border	43.222 2	39.666 7	43.222 2	42.666 7	41.000 0	43.444 4	42.333 3	43.555 5	43.333 3
brain	44.683 5	41.139 2	42.151 9	46.075 9	41.772 2	44.683 5	46.316 5	43.291 1	43.797 5
ufo	43.827 1	39.506 1	43.209 9	42.839 5	40.123 5	41.851 9	43.950 6	42.222 2	43.827 2
ufo11	44.141 4	41.010 1	47.474 7	46.262 6	39.899 0	46.969 7	47.474 7	40.101 0	47.575 7
venus	50.632 9	40.759 5	55.569 6	51.392 4	42.531 6	55.189 9	51.012 7	44.557 0	55.643
video	43.510 6	40.744 7	43.191 5	43.191 5	39.680 6	43.191 5	43.617 0	41.383 0	42.340 4
voituretuning	49.411 8	44.804 0	54.607 8	50.098 0	52.549 0	55.294 0	50.980 4	52.843 1	51.568 6
weed	46.018 5	38.518 5	45.833 3	47.129 6	41.111 1	46.481 4	47.481 5	40.370 4	45.740 7
average	46.814 6	41.198 8	46.812 3	47.403 6	42.802 1	47.310 2	47.998 7	43.694 9	47.123 7

表 3 中每个数据集上不同算法之间的最高准确率被粗化.由表 3 可以看出,本文提出的基于 k 个标记样本的受限玻尔兹曼机模型学习结果最好,在 aquarium 等 7 个数据集中,*k*-means.W-RBM 取得最佳正确率;在 alphabet 等 3 个数据集中,AP.W-RBM 取得了最佳正确率;在 border 数据集中,AP.W-RBM 取得了最佳正确率;且在统计算法平均正确率时,W-RBM 模型均相对有一定的提升.因此,W-RBM 模型在多数数据集上的准确率高于其他模型的准确率.

Table 4 Purities of different model algorithms (%)

表 4 不同模型算法纯度比较(%)

	k-means	DP	AP	*k*-means.RBM	DP.RBM	AP.RBM	*k*-means.W-RBM	DP.W-RBM	AP.W-RBM
alphabet	0.524 4	0.547 6	0.524 4	0.526 8	0.528 0	0.526 8	0.524 4	0.550 2	0.525 6
aquarium	0.680 9	0.680 9	0.680 9	0.685 1	0.680 9	0.680 9	0.686 9	0.680 9	0.680 9
bed	0.594 0	0.656 3	0.656 3	0.659 4	0.661 5	0.663 5	0.652 5	0.657 3	0.662 5
blog	0.682 3	0.672 2	0.670 9	0.678 5	0.672 2	0.696 2	0.674 7	0.675 9	0.702 5
border	0.539 1	0.512 0	0.526 1	0.538 0	0.530 4	0.528 2	0.531 5	0.514 1	0.540 2
brain	0.477 8	0.458 0	0.475 3	0.484 0	0.459 3	0.488 9	0.493 8	0.465 4	0.508 6
ufo	0.435 4	0.417 1	0.450 0	0.472 0	0.436 6	0.459 8	0.478 0	0.440 2	0.467 1
ufo11	0.440 5	0.416 5	0.457 0	0.448 1	0.431 6	0.464 6	0.451 9	0.411 4	0.472 2
venus	0.603 0	0.515 8	0.589 1	0.612 9	0.551 5	0.595 0	0.616 0	0.520 8	0.589 0
video	0.497 0	0.490 2	0.489 3	0.486 4	0.501 0	0.490 3	0.487 4	0.501 3	0.489 3
voituretuning	0.671 1	0.671 1	0.671 1	0.675 0	0.673 7	0.675 0	0.679 0	0.671 1	0.671 1
weed	0.487 8	0.435 6	0.461 1	0.464 4	0.442 2	0.504 4	0.453 3	0.436 7	0.502 2

表 4 展示了不同算法的纯度表现.表 4 中每个数据集上不同算法之间的最高纯度被粗化.由表 4 可以看出,本文提出的基于 k 个标记样本的受限玻尔兹曼机模型学习结果性能最好,在 aquarium 等 4 个数据集中,*k*-means.W-RBM 取得最佳纯度;在 blog 等 4 个数据集中,AP.W-RBM 取得了最佳纯度;在 alphabet 和 video 数据集中,AP.W-RBM 取得了最佳纯度;且在统计整体纯度时,W-RBM 模型均相对有一定的提升.因此,W-RBM 模型在多数数据集上纯度高于其他模型的纯度.

3 总结和展望

本文在对聚类与弱监督学习的研究中,提出了一种基于 k 个标记样本的弱监督学习框架,该框架运用聚类及聚类置信度实现了标记样本的扩展,并对受限玻尔兹曼机的能量函数进行了改进,提出了基于 k 个标记样本的受限玻尔兹曼机学习模型.完成了对该模型的推理并设计相关算法,并且通过使用 *k*-means、AP、DP 算法与

W-RBM 模型组合作为成组对比算法,大幅度地减少了标记样本的数量.其次,通过与主流聚类算法的对比实验可以看出,基于k个标记样本的弱监督学习框架及基于k个标记样本的受限玻尔兹曼机学习模型,在大部分数据集上优于现阶段主流的聚类算法.因此,本文提出的方案是可行且高效的.在接下来的研究中,我们还将继续深入探讨基于 k 个标记样本的弱监督学习框架与其他各类弱监督学习模型的相互关系.

References:

[1] Jain A, Dubes R. Algorithms for Clustering Data. Englewood Cliffs: Prentice-Hall, 1988.

[2] Hinton GE, Sejnowski TJ. Unsupervised Learning and Map Formation: Foundations of Neural Computation. Cambridge: MIT Press, 1999–3514.

[3] Tu EM, Yang J. A review of semi-supervised learning theories and recent advances. Journal of Shanghai Jiaotong University, 2018, 52(10):1280–1291 (in Chinese with English abstract).

[4] Liu JW, Liu Y, Luo XL. Semi-supervised learning methods. Chinese Journal of Computers, 2015,8:1592–1617 (in Chinese with English abstract).

[5] Liang JY, Gao JW, Chang Y. The research and advances on semi-supervised learning. Journal of Shanxi University (Natural Science Edition), 2009,32(4):528–534 (in Chinese with English abstract).

[6] 刘蓉,李红艳.半监督学习研究与应用.软件导刊,2010,9(8):6–7.

[7] Zhou ZH. A brief introduction to weakly supervised learning. National Science Review, 2018,5:44–53.

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

[9] Vapnik V, Sterin A. On structural risk minimization or overall risk in a problem of pattern recognition. Automation and Remote Control, 1977,10(3):1495–1503.

[10] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc. of the Conf. on Computational Learning Theory. 1998. 92–100.

[11] Zhou Y, Goldman S. Democratic co-learning. In: Proc. of the IEEE Int'l Conf. on TOOLS with Artificial Intelligence. 2004. 594–602.

[12] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning. IEEE Trans. on Neural Networks, 2009,20(3):542–542.

[13] Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Proc. of the 19th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers Inc., 2002. 307–314.

[14] Wang L, Bo LF, Jiao LC. Density-sensitive semi-supervised spectral clustering. Ruan Jian Xue Bao/Journal of Software, 2007, 18(10):2412–2422 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/jos182412]

[15] Gao Y, Liu DY, Qi H, Liu H. Semi-supervised *K*-means clustering algorithm for multi-type relational data. Ruan Jian Xue Bao/Journal of Software, 2008,19(11):2814–2821 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2814.htm [doi: 10.3724/SP.J.1001.2008.02814]

[16] Wang N, Li X. Active semi-supervised spectral clustering based on pairwise constraints. ACTA ELECTRONICA SINICA, 2010,38(1):172–176 (in Chinese with English abstract).

[17] Wang HJ, Li TR, Li T, Yang Y. Constraint neighborhood projections for semi-supervised clustering. IEEE Trans. on Cybernetics, 2014,44(5):636–643.

[18] Xiong S, Azimi J, Fern XZ. Active learning of constraints for semi-supervised clustering. IEEE Trans. on Knowledge and Data Engineering, 2014,26(1):43–54.

[19] Huang G, Song S, Gupta JND, Cheng W. Semi-supervised and unsupervised extreme learning machines. IEEE Trans. on Cybernetics, 2014,44(12):2405–2417.

[20] Portela NM, Cavalcanti GDC, Ren TI. Semi-supervised clustering for MR brain image segmentation. Expert Systems with Applications, 2014,41(4):1492–1497.

[21] Lad S, Parikh D. Interactively guiding semi-supervised clustering via attribute-based explanations. In: Proc. of the European Conf. on Computer Vision. Cham: Springer-Verlag, 2014. 333–349.

[22] Anand S, Mittal S, Tuzel O, Meer P. Semi-supervised kernel mean shift clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014,36(6):1201−1215.

[23] Lai HP, Visani M, Boucher A, Ogier J-M. A new interactive semi-supervised clustering model for large image database indexing. Pattern Recognition Letters, 2014,37:94−106.

[24] Ding SF, Jia HJ, Zhang LW, Jin FX. Research of semi-supervised spectral clustering algorithm based on pairwise constraints. Neural Computing and Applications, 2014,24(1):211−219.

[25] Huang SD, Wang HJ, Li T, Yang Y, Li TR. Constraint co-projections for semi-supervised co-clustering. IEEE Trans. on Cybernetics, 2015,46(12):3047−3058.

[26] Yi JF, Zhang LY, Yang TB, Liu W, Wang J. An efficient semi-supervised clustering algorithm with sequential constraints. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1405−1414.

[27] Triguero I, García S, Herrera F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. Knowledge and Information Systems, 2015,42(2):245−284.

[28] Mehrkanoon S, Agudelo OM, Suykens JAK. Incremental multi-class semi-supervised clustering regularized by Kalman filtering. Neural Networks, 2015,71:88−104.

[29] Honda K, Ubukata S, Notsu A, Takahashi N, Ishikawa Y. A semi-supervised fuzzy co-clustering framework and application to twitter data analysis. In: Proc. of the 2015 Int'l Conf. on Informatics, Electronics & Vision. 2015. 1−4.

[30] Nie FP, Li J, Li XL. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York, 2016. 1881−1887.

[31] Wang D, Gao XB, Wang XM. Semi-supervised nonnegative matrix factorization via constraint propagation. IEEE Trans. on Cybernetics, 2016,46(1):233.

[32] Saha S, Alok AK, Ekbal A. Brain image segmentation using semi-supervised clustering. Expert Systems with Applications, 2016, 52:50−63.

[33] Yu ZW, Luo PN, You J, Wong H, Leung H, Wu S, Zhang J, Han GQ. Incremental semi-supervised clustering ensemble for high dimensional data clustering. IEEE Trans. on Knowledge and Data Engineering, 2016,28(3):701−714.

[34] Fang XZ, Xu Y, Li XL, Lai ZH, Wong WK. Robust semi-supervised subspace clustering via non-negative low-rank representation. IEEE Trans. on Cybernetics, 2016,46(8):1828−1838.

[35] Tuan TM, Ngan TT. A novel semi-supervised fuzzy clustering method based on interactive fuzzy satisficing for dental X-ray image segmentation. Applied Intelligence, 2016,45(2):402−428.

[36] Tuan TM. A cooperative semi-supervised fuzzy clustering framework for dental X-ray image segmentation. Expert Systems with Applications, 2016,46:380−393.

[37] Alok AK, Saha S, Ekbal A. Multi-objective semi-supervised clustering for automatic pixel classification from remote sensing imagery. Soft Computing, 2016,20(12):4733−4751.

[38] Nie FP, Cai GH, Li XL. Multi-view clustering and semi-supervised classification with adaptive neighbours. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco, 2017. 2408−2414.

[39] Yang Y, Li ZZ, Wang W, Tao DP. An adaptive semi-supervised clustering approach via multiple density-based information. Neurocomputing, 2017,257:193−205.

[40] Li CX, Tan YJ, Wang DB, Ma PJ. Research on 3D face recognition method in cloud environment based on semi-supervised clustering algorithm. Multimedia Tools and Applications, 2017,76(16):17055−17073.

[41] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016.

[42] Chu JL, Wang HJ, Meng H, Jin P, Li TR. Restricted Boltzmann machines with Gaussian visible units guided by pairwise constraints. IEEE Trans. on Cybernetics, 2018. [doi: 10.1109/TCYB.2018.2863601]

[43] Yin J, Li WW, Yang DH, Run H. Improved model based on classification restricted Boltzmann machine. Journal of Chinese Computer Systems, 2018,39(7):41−45 (in Chinese with English abstract).

[44] Fu WB, Sun T, Liang J, Run BW, Fan FS. Review of principle and application of deep learning. Computer Science, 2018,45(S1): 24−28+53 (in Chinese with English abstract).

[45] Shi JR, Ma YY. Research progress and development of deep learning. Computer Engineering and Applications, 2018,54(10):1–10 (in Chinese with English abstract).

[46] Tao Y, Yang F, Liu Y, Dai B. Research and optimization of *K*-means clustering algorithm. Computer Technology and Development, 2018,28(6):90–92 (in Chinese with English abstract).

[47] Wang XZ, Wang YD, Zhan Y, Yuan F. Optimization of *K*-means clustering by feature weight learning. Journal of Computer Research and Development, 2003,40(6):869–873 (in Chinese with English abstract).

[48] Wang LM, Wang NB, Han XM, Wang YZ. Semi-supervised hierarchical optimization-based affinity propagation algorithm and its applications. Int'l Journal of Computers & Applications, 2018,(3):1–10.

[49] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014,344(6191):1492–1496.

[50] Wang M, Yang LJ, Hua XS. MSRA-MM: Bridging re-search and industrial societies for multimedia information retrieval. Technical Report, MSR-TR-2009-30 (March), Mi-440 Crosoft Research Asia, 2009.

附中文参考文献:

[3] 屠恩美,杨杰.半监督学习理论及其研究进展概述.上海交通大学学报(自然版),2018,52(10):1280–1291.

[4] 刘建伟,刘媛,罗雄麟.半监督学习方法.计算机学报,2015,8:1592–1617.

[5] 梁吉业,高嘉伟,常瑜.半监督学习研究进展.山西大学学报(自然科学版),2009,32(4):528–534.

[6] 刘蓉,李红艳.半监督学习研究与应用.软件导刊,2010,9(8):6–7.

[14] 王玲,薄列峰,焦李成.密度敏感的弱监督谱聚类.软件学报,2007,18(10):2412–2422. http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/ jos182412]

[15] 高滢,刘大有,齐红,刘赫.一种半监督 *k* 均值多关系数据聚类算法.软件学报,2008,19(11):2814–2821. http://www.jos.org.cn/1000-9825/19/2814.htm [doi: 10.3724/SP.J.1001.2008.02814]

[16] 王娜,李霞.基于监督信息特性的主动半监督谱聚类算法.电子学报,2010,38(1):172–176.

[43] 尹静,李唯唯,杨德红,闰河.一种新的分类受限玻尔兹曼机改进模型.小型微型计算机系统,2018,39(7):41–45.

[44] 付文博,孙涛,梁藉,闫宝伟,范福新.深度学习原理及应用综述.计算机科学,2018,45(S1):24–28+53.

[45] 史加荣,马媛媛.深度学习的研究进展与发展.计算机工程与应用,2018,54(10):1–10.

[46] 陶莹,杨锋,刘洋,戴兵.*K* 均值聚类算法的研究与优化.计算机技术与发展,2018,28(6):90–92.

[47] 王熙照,王亚东,湛燕,袁方.学习特征权值对 *K*-均值聚类算法的优化.计算机研究与发展,2003,40(6):869–873.

付治(1994－),男,湖北大悟人,学士,主要研究领域为机器学习,区块链.

滕飞(1984－),女,博士,副教授,CCF 专业会员,主要研究领域为云计算.

王红军(1977－),男,博士,副研究员,CCF 高级会员,主要研究领域为机器学习.

张继(1993－),男,学士,主要研究领域为深度学习,机器学习.

李天瑞(1969－),男,博士后,教授,博士生导师,CCF 杰出会员,主要研究领域为智能信息处理,数据挖掘,云计算,大数据,粗糙集,粒计算.