第 56 卷　第 1 期
2020 年 1 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 56, No. 1
Jan., 2020

DOI:10. 13232/j. cnki. jnju. 2020. 01. 002

半监督平面聚类算法设计

杨红鑫,杨绪兵*,张福全,业巧林
(南京林业大学信息科学技术学院,南京,210037)

摘　要:采用以平面为原型来拟合样本的思想设计学习机,已在机器学习和数据挖掘等领域引起广泛关注,然而,如何利用少量标记样本,兼顾平面原型特点实现聚类,鲜见报道. 以kPC(k-Plane Clustering)为切入点,在有标样本极端少的情况下,设计了半监督型平面聚类算法semi-kPC. 考虑到L1范数较L2范数更为鲁棒的事实,在已有工作L1kPC (L1 norm kPC)的基础上,提出基于L1范数的半监督聚类方法semi-L1kPC. 从每类仅有一个已标样本出发,在人工数据集和UCI数据集上的实验表明:(1)在XOR(Exclusive OR)问题上,平面型的聚类方法的聚类准确率均显著高于*k*-means算法,因为*k*-means无法利用平面特性;(2)在引入少量监督信息后,半监督型聚类方法semi-kPC和semi-L1kPC比其他聚类方法的聚类准确率更高;(3)采用L1范数的semi-L1kPC比semi-kPC的鲁棒性更好.

关键词:半监督聚类,平面分布,鲁棒性,L1范数度量
中图分类号:TP391　　**文献标识码**:A

Semi-supervised plane clustering algorithm

Yang Hongxin, Yang Xubing*, Zhang Fuquan, Ye Qiaolin
(School of Computer Science and Technology, Nanjing Forestry University, Nanjing, 210037, China)

Abstract: Unsupervised learning machine, typically generated from plane-shaped data clustering, has been widely applied in the fields of machine learning, data mining. Instead of point-prototype in classic *k*-means or FCM, the so-called plane clustering methods aim to seek multiple plane-prototype as cluster centers, and then group the data into clusters by minimizing the distance between fitting planes and data points in their corresponding clusters. There has been increasing interest in plane-based methods in the family of data clustering in last decades, including kPC (*k* Plane Clustering), PPC (Proximal Plane Clustering) and TWSVC (Twin Support Vector Clustering). However, it is rarely reported how to utilize plane characteristics to design semi-supervised plane clustering algorithms, even in the era of the rising of semi-supervised learning. Moreover, due to the fact that L1 norm is more robust than L2 norm, in this paper, we propose a robust semi-supervised plane clustering method based on L1 norm, termed as semi-L1kPC for shortly. Compared with the state-to-the-art methods, the advantages of our proposed lie in three folds: (1) similar to kPC, it has clear geometrical intuition. That is, the data is grouped into clusters by minimizing the point-to-plane distance measured by infinite norm, the dual of L1 norm. (2) The algorithm is designed on the small number of labeled samples, even only need ONE labeled sample per class. (3) The leading problem can be solved by linear programming, rather than eigenvalue problems in kPC, quadratic programming problems in TWSVC. The experimental comparisons on artificial and benchmark datasets show that: (1) on exclusive XOR problem, the clustering accuracies of plane-prototype methods are higher than that of point-prototype ones. (2) When introducing fewer supervised

基金项目:国家自然科学基金(31670554),江苏省自然科学基金(BK20171453),2019江苏省研究生科研创新项目(SJKY19_0907)
收稿日期:2019－07－02

*通讯联系人,E-mail: xbyang@njfu.edu.cn

information, i.e., labeled samples, our proposed semi-kPC and semi-L1kPC outperform the foresaid unsupervised methods in cluster accuracy. (3) As for kPC itself, semi-L1kPC receives more robustness than L2 norm based semi-kPC.

Key words: semi-supervised clustering, plane distribution, robustness, L1 norm metric

半监督学习是机器学习、数据挖掘等领域的重要研究方向,已广泛应用于自然语言处理、计算机视觉和手写数字识别等领域[1−2]. 科技发展使数据采集和数据存储更简便,人们可以很容易地获得无类别标记的海量数据. 但由于类别的标签难以获取(实验时间过长甚至是毁灭性的)或是代价昂贵(使用高精尖仪器设备等),如何充分利用少量的标记样本和大量无标记样本完成学习任务成为半监督学习所要解决的问题. 自 20 世纪 70 年代的自训练方法(Self-Training)问世以来,经过近 50 年的发展,按学习任务可将半监督学习[3−4]分为半监督分类[5]、半监督回归[6]、半监督聚类[7−8]和半监督降维[9]. 本文仅关注半监督聚类.

从监督信息利用上看现有的半监督聚类方法可分为三类[3]:(1)基于约束的;(2)基于距离度量的;(3)对约束和距离度量进行融合的. 第一类主要采用 must-link 和 cannot-link 成对约束关系作为监督信息,用于指导聚类. 第二类根据少量已标样本信息,通过相似性度量,如 k 近邻等,寻找一批与之相似的样本,并以已标样本的标号标记之. 当标记出的样本个数达到某条件后,再完成后续聚类. 第三类是在使用带约束的监督信息的同时进行度量学习的半监督聚类学习方法. 如经典 k-means,2002 年 Basu et al[10]设计了半监督的约束 k-means (Constrained k-means)和种子 k-means (Seeds k-means)算法,先使用标记样本组成种子集合,在种子集合中构造出 k 个簇中心用以替换 k-means 的随机代表点,再按 k-means 继续迭代. 以 k-means 为原型,吕峰等[7]提出同时利用特征和样本两个层面的信息来设计半监督聚类方法. 高滢等[11]改进了 k-means 聚类算法的初始聚类中心以及相似性度量方法,提出一种半监督 k-means 多关系数据聚类算法. 由于上述方法都是基于点原型的聚类算法进行设计,对于样本呈球形或椭球形分布较为有效,而当样本呈线性(包括 XOR(Exclusive OR)问题)或平面型分布时,上述算法难以胜任. 借鉴点原型思想,Bradley and Mangasarian[12]用平面原型替换 k-means 的点原型,提出 kPC(k-Plane Clustering)聚类算法,即对 k 簇聚类问题,先随机选择 k 个平面,根据样本到各个平面的距离实现“样本归簇”,再根据归簇的样本“更新平面”,交替使用“样本归簇”和“平面更新”直到所有簇中的样本归属不再变化,迭代中的拟合平面可通过计算特征方程获得. 这种采用平面拟合样本的思想,结合 SVM(Support Vector Machine),Fung and Mangasarian[13]和 Mangasarian and Wild[14]发表了二分类方法 PSVM(Proximal Support Vector Machine)[13]和 GEPSVM(PSVM via Generalized Eigenvalue)[14],导出的问题可分别通过线性方程组和广义特征方程求解. Jayadeva et al[15]用平面拟合样本的模型设计取代 SVM 的间隔,发表了 TWSVM(Twin SVM)二分类算法,两个拟合平面可通过求解两个较小规模的二次规划获得(占 SVM 求解规模的 1/4). 自此,平面型学习机受到了广泛关注,提出诸如聚类的 TWSVC(Twin SV Clustering)[16],FRTWSVC (L1 norm TWSVC)[17]和 MMC(Maximal Margin clustering)[18],半监督分类的 semi-SVM[18],semi-GEPSVM[19],半监督聚类的 semi-supervised WLL-TWSVC, fuzzy semi-supervised WLL-TWSVC[20]等. 据所知,除 Rastogi and Pal[20]报道将半监督方法应用在平面聚类外,未见文献采用平面拟合思想设计半监督聚类算法. fuzzy semi-supervised WLL-TWSVC 本身就是基于 L2 范数的 TWSVC 为原型构造目标函数,虽然 TWSVC 也借鉴平面拟合样本设计模型,但为了引入簇间信息取消了 kPC 的点到平面距离思想,这样做不仅使 TWSVC 失去了 kPC 原有的清晰的几何解释,另一方面,在聚类过程中产生的样本临时所属的簇信息是否一定有利于后续聚类仍值得商榷. 基于 TWSVC 的后续工作均存在此类问题. 因此本文的研究仍从保留 kPC 的几何解释开始.

本文沿袭kPC的几何解释,根据“样本距离其簇对应的拟合平面最近”的思想来实现“样本归簇”和“平面更新”,基于kPC和已有工作L1kPC设计半监督聚类算法,提出基于L2范数的semi-kPC和基于L1范数的semi-L1kPC. 所需求解的拟合平面,可通过求解一般特征方程和线性规划完成. 算法均在每簇仅有一个已标样本的极端情形下进行设计.

1　kPC和L1kPC聚类算法简介

1.1　kPC聚类　kPC算法用k个平面原型代替k-means的点原型聚类中心,其优化目标是要寻找k个超平面,记为$H=\{h_1,h_2,\cdots,h_k\}$,定义如下:

$$h_i=\{x\mid \boldsymbol{w}_i^T x+\gamma_i=0, x\in R^d\}\quad i=1,2,\cdots,k \tag{1}$$

其中,$\boldsymbol{w}_i$和γ_i分别表示超平面h_i对应的法向量和阈值. 在簇内样本距其对应超平面距离最近的目标指引下,问题归结为如下优化问题:

$$\min\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{w}_i^T x_j^{(i)}+\gamma_i\right)^2 \quad s.t.\left\|\boldsymbol{w}_i\right\|^2=1, i=1\sim k \tag{2}$$

其中,$\dfrac{\left(\boldsymbol{w}_i^T x_j^{(i)}+\gamma_i\right)^2}{\left\|\boldsymbol{w}_i\right\|^2}$可解释为第$i$簇的第$j$个样本$x_j^{(i)}$距其对应超平面$h_i$的距离平方,$n_i$为第$i$簇样本数. 进一步化简可知,超平面$h_i$可通过特征方程求解,可由该簇样本的协方差和均值决定.

1.2　L1kPC聚类　采用L1范数的L1kPC算法需要解决两个问题:(1)点到平面距离的L1解析度量;(2)L1范数约束下,可行域是非凸的. 由文献[21]可知,点v到平面$\boldsymbol{w}^T x+\gamma=0$的距离为:

$$s=\frac{\left|\boldsymbol{w}^T v+\gamma\right|}{\left\|\boldsymbol{w}\right\|_\infty} \tag{3}$$

其中,$\|\cdot\|_\infty$表示无穷范数,后续的$\|\cdot\|_1$表示1范数. 若依kPC方式处理,可令$\|\boldsymbol{w}\|_\infty=1$,其可行域同样是非凸的. 此时对于第$i$某簇样本

$$A_i=\begin{bmatrix}\left(x_1^{(i)}\right)^T\\ \vdots\\ \left(x_{n_i}^{(i)}\right)^T\end{bmatrix}$$

$\sum_{j=1}^{n_i}\left|\boldsymbol{w}_i{}^T x_j^{(i)}+\gamma_i\right|$恰可表示为$\left\|A_i\boldsymbol{w}_i+e\gamma_i\right\|_1$. 在$\left\|\boldsymbol{w}_i\right\|_\infty=1$作用下,$\left\|A_i\boldsymbol{w}_i+e\gamma_i\right\|_1$可表示L1度量下该簇样本距其拟合平面的距离之和,$\boldsymbol{e}$为分量全为1的列向量. 由于无穷范数与1范数互为对偶,若采用$\left\|\boldsymbol{w}_i\right\|_1=1$来代替$\|\boldsymbol{w}\|_\infty=1$,只需对$\boldsymbol{w}_i$作一尺度变化,即将$|\boldsymbol{w}_i|$的最大分量置为1,不影响上述的几何解释. 因该聚类的拟合平面求解也只需关注本簇样本,忽略下标,可得如下优化模型[9,22],如式(4)所示:

$$\min_{(\boldsymbol{w},\gamma)}\left\|A\boldsymbol{w}+e\gamma\right\|_1, s.t.\|\boldsymbol{w}\|_1=1 \tag{4}$$

该问题(式(4))是一个非凸优化问题,可将此非凸问题转化为有限个子集上凸问题来求解,这样可使问题转化为求解一个线性规划.

该算法启动同kPC,通过“样本归簇”和“平面更新”两个步骤实现聚类. 实验验证了该聚类较之kPC具有鲁棒性.

2　半监督kPC和L1kPC平面聚类算法

对呈平面型分布的数据,采用kPC类聚类方法非常有效,但此类算法通过随机初始化平面启动算法,实验中发现,这种随机启动会直接影响到聚类效果. 受k-means类型半监督聚类的启发,若是利用少量监督信息来启动算法,则聚类效果应该会有明显改善.

图1是一个典型的XOR问题的两类数据,图1a显示两类数据,抽样于两条相互交叉的直线(施加了部分噪声). 图1b中加入了两个标记样本,分别记为“x”(q1)和“x”(q2),在此监督信息作用下,半监督型k-means仍然有较多样本被错误聚类. 图1c和图1d表示本文提出的两个半监督聚类方法:semi-kPC和semi-L1kPC. 直观上,二者产生的两个拟合平面能够真实地反映样本的原来分布,绝大多数样本均获得了正确的聚类. 由于无法从一个已标样本生成拟合平面,图1c和图1d中标记为“□”的样本是根据q1和q2查询点k近邻寻找到的样本用以辅助求解拟合平面的.

下面介绍本文提出的半监督平面聚类算法.

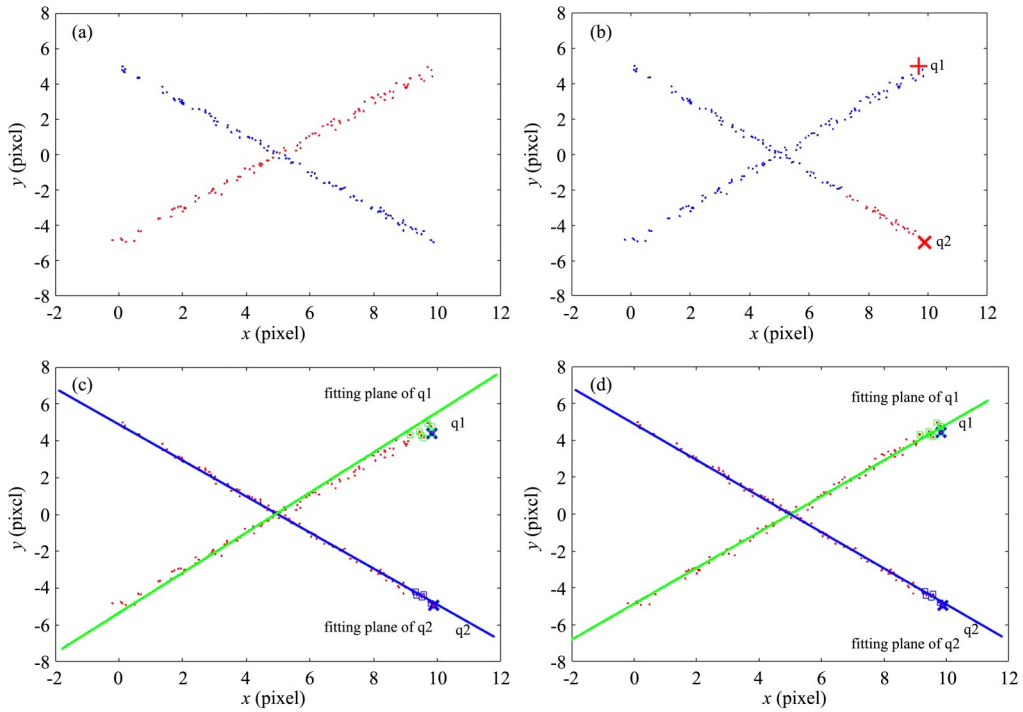

图 1　数据集的原始分布(a)和 k-means (b),semi-kPC (c),semi-L1kPC (d)三种方法在 XOR 上的聚类效果

Fig.1　Illustration for results of three methods on XOR dataset:(a) original distriburion, (b) k-means,(c) semi-kPC,(d) semi-L1kPC

2.1　半监督平面聚类算法　两个半监督聚类算法,其主要步骤仍遵循 kPC 的两步迭代:“样本归簇”和“平面更新”. 由于 Semi-kPC 采用人们熟悉的 L2 范数,为避免重复和限于篇幅,本节仅描述采用 L1 范数的 Semi-L1kPC 算法. 符号说明如下:给定 k 簇 n 个样本,包含两部分数据:有标样本集 x_{Labeled}(少量)和无标样本集 $x_{\text{Unlabeled}}$(大量),记每簇的有标样本集为 $S_j, j=1,2,\cdots,k$. 如前文所述,每簇中至少含一个有标样本,故 $S_j \neq \varnothing$, $S_j \cap S_i = \varnothing (i \neq j)$, $\cup S_j = X_{\text{Labeled}}$, $i,j=1,2,\cdots,k$. 迭代过程中产生的临时簇记为 $A_i, i=1,\cdots,k$, k 个平面标记同前文. 算法描述如下.

算法 1　Semi-L1kPC 算法

输入:样本集 x_{Labeled} 和 $x_{\text{Unlabeled}}$,簇数 k.

输出:簇集 $A_i, i=1,2,\cdots,k$

Step1　初始化,用有标样本初始化 S_i,并入簇集 $A_i \leftarrow S_i, i=1,2,\cdots,k$. 设置最大迭代次数 MaxIter,迭代计数器 iter←0.

Step2　检查每个 A_i 能否产生拟合平面,若能转 Step4,不能 Step3.

Step3　从 $x_{\text{Unlabeled}}$ 中寻找与 A_i 内的样本较为相似者,并入 A_i,转 Step2;

Step4　平面更新. 按式(4)的解更新超平面. //※

Step5　样本归簇. 根据公式(3)计算 L1 范数的点到平面距离,实现样本归簇.

$A_i \leftarrow A_i \cup S_i, i=1,2,\cdots,k$//※iter←iter+1.

Step6　检查停机条件.

条件 1:若 iter>MaxIter,停机,否则转 Step4;

条件 2:若 A_i 中的样本归属不再变化,停机,否则转 Step4.

几点说明:

(1)算法的 Step4 和 Step5 加注了标记“※”以示 Semi-kPC 与之有所不同,其平面更新可按下文的式(6)进行(见下文定理 1). 而 semi-L1kPC 则按下文的式(10)计算(见下文定理 4).

(2)Step2 中的检查能否产生拟合平面,可通过检查簇中线性无关的样本数是否超过样本维数来判断. 因为在 d 维线性空间中,唯一确定一个平面至少需要 d 个样本. 当然亦可通过点原型的半监督方法来启动平面聚类.

(3)两个聚类算法 kPC 和 L1kPC 得到的解都是非凸问题,无法保证全局最优解. k 个拟合平面

均在k个簇上计算,在各自的凸集上的优化是凸优化,存在最优解.为防止算法不收敛或收敛速度过慢,在停机条件中设置了最大迭代次数.由实验的经验可知设置次数为50时,大多数问题均有较好聚类效果.

(4)监督信息利用问题.算法的Step1和Step5都运用了监督信息.Step1中是为了避免原聚类算法的随机初始化问题,而Step5中是为了防止样本归簇过程中因拟合平面尚未稳定而可能产生的有标样本归簇错误,为强调有标样本的作用,在迭代过程中加以强化.

2.2　模型描述与相关证明　为方便阅读,给出几个定理并简证,为上述算法提供理论支撑.

定理1　可通过解决式(5)获得semi-kPC的更新平面:

$$\min_{(\boldsymbol{w}_i,\gamma_i)}\left\|A_i\boldsymbol{w}_i+\boldsymbol{e}\gamma_i\right\|_2^2, s.t.\left\|\boldsymbol{w}_i\right\|_2^2=1 \tag{5}$$

拟合平面的$\boldsymbol{w}_i$是特征方程$\boldsymbol{\psi}\boldsymbol{w}_i=\lambda\boldsymbol{w}_i$的最小特征值对应的特征向量,阈值$\gamma_i=-e^TA_i\boldsymbol{w}_i/n_i$,$\boldsymbol{\psi}$为样本协方差阵[17].

简证　由于L2范数可导,构造拉格朗日函数,并令其对$\boldsymbol{w}_i$和γ_i偏导数为0,可得:

$$\boldsymbol{\psi}\boldsymbol{w}_i=\lambda\boldsymbol{w}_i,\gamma_i=\frac{-\boldsymbol{e}^TA_i\boldsymbol{w}_i}{n_i} \tag{6}$$

其中,$\boldsymbol{\psi}=A^T\left(I-\dfrac{\boldsymbol{e}\boldsymbol{e}^T}{n_i}\right)A$,化简可知其为样本协方差,而$\dfrac{\boldsymbol{e}^TA_i}{n_i}$恰是样本均值(行向量表示).

下面介绍在L1范数下,semi-L1kPC算法需要的理论保证.符号说明同前.

定理2　L1范数下,可通过求解式(7)来解决semi-L1kPC中的平面更新:

$$\min_{(\boldsymbol{w}_i,\gamma_i)}\left\|A_i\boldsymbol{w}_i+\boldsymbol{e}\gamma_i\right\|_1, s.t.\left\|\boldsymbol{w}_i\right\|_\infty=1 \tag{7}$$

证明　由文献[22]可知,在$\left\|\boldsymbol{w}_i\right\|_\infty=1$作用下,$\left|\boldsymbol{w}_i{}^Tx_j^{(i)}+\gamma_i\right|$表示点到平面的L1距离.按L1范数定义知:

$$\left\|A_i\boldsymbol{w}_i+\boldsymbol{e}\gamma_i\right\|_1=\left|\left(x_j^{(i)}\right)^T\boldsymbol{w}_i+\gamma_i\right|$$

表示簇A_i中所有样本点到超平面的L1距离之和.

定理3　式(7)与式(8)同解.

$$\min_{(\boldsymbol{w}_i,\gamma_i)}\left\|A_i\boldsymbol{w}_i+\boldsymbol{e}\gamma_i\right\|_1, s.t.\left\|\boldsymbol{w}_i\right\|_1=1 \tag{8}$$

证明　比较式(7)和式(8),两者只在范数约束上存在不同.

先证(8)→(7).设$\boldsymbol{w}_i^*$是式(8)的最优解,其对应分量记为$\boldsymbol{w}_i^*=\left(w_{i1}^*,w_{i2}^*,\cdots,w_{id}^*\right)^T$,且$\sum_j\left|w_{ij}^*\right|=1$,不妨设其第$l$个分量的绝对值最大.若$w_{il}^*>0$则令$\boldsymbol{w}_i^*\leftarrow\boldsymbol{w}_i^*/w_{il}^*$;若$w_{il}^*<0$则令$\boldsymbol{w}_i^*\leftarrow\boldsymbol{w}_i^*/(-w_{il}^*)$,替换后的$\boldsymbol{w}_i^*$满足表示$\left\|\boldsymbol{w}_i^*\right\|_\infty=1$.对式(7)的目标函数同样进行放缩,可知式(7)仍能保持原有几何解释,且解不变.考虑在式(7)的目标中除以正数($|w_{il}^*|$)是为了保证问题在迭代过程中始终保持优化最小化目标.

证明(8)←(7),与证明(8)→(7)类似,只需对式(7)的最优解进行归一化处理即可.

证毕.

从证明过程知,令超平面法向量长度为1,一方面是为了保持几何解释,同时也起到了避免问题的解退化为平凡解(零解).

定理4　式(8)可通过式(9)求解:

$$\min_{\boldsymbol{\alpha}}\left\|(A-\boldsymbol{e}m)\boldsymbol{P}\alpha\right\|_1, s.t.\boldsymbol{\alpha}\geqslant 0, \boldsymbol{e}^T\alpha=1 \tag{9}$$

其中,$\boldsymbol{P}=(p_1,p_2,\cdots,p_d)$是该问题求解凸集上的一组标准正交基,该基由凸壳$\left\|\boldsymbol{w}_i\right\|_1=1$上的$d$个线性无关的向量(分量对应顶点坐标)组成.

限于篇幅,仅说明基本思路:因$\left\|\boldsymbol{w}_i\right\|_1=1$对应的是一凸壳,式(8)的可行域非凸.为解决非凸问题,先按凸壳顶点次序,将非凸集分解为多个凸集,再在凸集上进行优化.式(9)的目标函数,可通过两个非负向量差表示,问题可转化为线性规划问题.

由定理4得式(8)和式(9)的解有如下关系成立:

$$\boldsymbol{w}_i=\boldsymbol{P}\alpha \tag{10}$$

3　实验与分析

本文的实验环境为CPU 2.6 GHz,Intel(R) Core(TM) i5-3230 M,内存4.0 GB,Windows 7旗舰版,Matlab R2015b.在人工数据集和UCI标

准数据集上进行实验. 其中,人工数据集主要用于可视化、验证 semi-kPC 和 semi-L1kPC 算法在 XOR 问题上的平面特性以及比较两种范数的半监督聚类方法的鲁棒性;UCI 数据集用来测试半监督平面聚类算法的推广能力. 实验采用十折交叉验证. 训练集中除少量样本作为有标样本,对训练模型提供监督信息外,其余样本均作无标记处理. 采用聚类准确率[16−17,23]来评价聚类算法的性能. 假定 G 是聚类算法预测出的样本标签集合,Q 是样本真实标签集合. 对于样本集合中的任意一对数据点,用四个变量 f_{11},f_{10},f_{01},f_{00} 来计算聚类准确率. 聚类准确率(Accuracy,ACC)的定义如下:

$$ACC=\frac{f_{11}+f_{00}}{f_{11}+f_{01}+f_{10}+f_{00}} \tag{11}$$

其中,$f_{11}=|G\cap Q|$ 表示预测正确的样本个数,即既在集合 G 也在集合 Q 中,剩下三个变量. 同理,$f_{00}=\left|\overline{G}\cap\overline{Q}\right|$, $f_{10}=\left|G\cap\overline{Q}\right|$, $f_{01}=\left|\overline{G}\cap Q\right|$. 符号 $|\cdot|$ 表示集合的势,$\overline{G}$ 和 $\overline{Q}$ 表示集合 G 和 Q 的补集.

3.1 人工数据集

3.1.1 两类 XOR 问题实验 图 1 所示的两类人工数据集是从两条相互交叉的直线中各抽样 100 个点,在所有样本点的第二个分量上注入高斯噪声(噪声服从高斯分布,大小为 $N(0,0.1)$)构成. 从每类仅有的一个有标样本出发,此时仍无法产生拟合平面. 本文按 k 近邻方法寻找与之相似样本并加入到对应的临时簇集,当簇中样本数超过 2 时,计算拟合平面,不断重复算法 1 的“平面更新”和“样本归簇”,直到满足停机条件. 图 1b 的半监督 k-means 方法共有 70 个样本发生聚类错误,准确率为 65%,而 semi-kPC 和 semi-L1kPC 共有三个和一个样本未能获得正确归簇,准确率分别为 98.5% 和 99.5%. 实验表明,与点原型的 k-means 相比,在 XOR 问题上,平面型的半监督聚类方法更适用于平面型或线型分布的数据.

3.1.2 混有野值的人工数据集实验 大量文献说明,采用 L1 范数的学习机比采用 L2 范数能更有效地抑制野值(outlier)和噪声的影响. 已在实验上得到了验证,采用 L1 范数的 L1kPC 聚类方法,比采用 L2 范数的 kPC 抗野值能力更好. 人工数据集分布如图 2 所示,两类样本的第一分量(横轴)分别从区间[0,1.5]和[1.5,3]抽样,第二分量(纵轴)分别设为 1 和 1.3,按第一分量的次序,在两类的前六个样本中加上 10% 的高斯噪声. 图 2 所示的 q1 和 q2 同样表示两个有标记样本,按算法 1 生成两类拟合平面如图 2 所示的线段. 图 2a 至图 2f 分别展示了六种聚类算法的聚类结果及拟合平面的结果. 其中,图 2e 中的两条拟合直线,一条虽几乎与横轴平行,但发生了偏移;而另一条则明显偏离样本原始分布. 而在图 2f 中两类拟合平面几乎都与横轴平行,两类样本也都达到了正确归簇,即 semi-L1kPC 的两个拟合平面能够较好地反映样本分布. 需要注意的是,图 2e 至图 2f 中的有标样本选择了野值点,若选择正常样本作为有标样本,则两种方法获得的拟合平面均能较好地反映样本的线性分布特性,拟合效果上两者的差异不明显. 为节约篇幅,结果未在文中列出.

图 2 所示的人工数据集上,semi-L1kPC 算法和 semi-kPC 相比,其拟合效果更为符合人类直觉,这也正验证了 L1 范数比 L2 范数更鲁棒.

3.2 UCI 数据集 为了进一步检验平面型的半监督聚类算法的泛化能力,在 UCI 数据集上进行实验. 本文选取 Ecoli, Glass 和 Haberman 等九个数据集,数据基本信息如表 1 所示.

从每类仅有一个有标样本出发(监督信息),按算法 1,在不满足生成拟合平面的条件时采用 k 近邻方法寻找相似样本,当临时簇中样本数大于维数时按“平面更新”和“样本归簇”完成后续聚类. 按 10 折交叉验证,聚类准确率取 10 次的平均结果,汇报结果于表 2,表中黑体字为聚类准确率最高,下画线为聚类准确率的平均值较高.

从表 2 可见,和点原型的 k-means 聚类方法相比,平面型聚类方法在八个数据集上获得了最高的聚类准确率. 为检验引入的少量监督信息是否有利于提升聚类性能,表 2 中还分别对两个半监督方法和四个无监督方法的聚类结果进行了平均,结果显示:引入监督信息后,七个数据集上的聚类准确率均有提升,其中在 Ecoil 上有大幅度提升. 具体到各方法,在六个数据集上,semi-kPC 优于 kPC;在七个数据集上,semi-L1kPC 优于

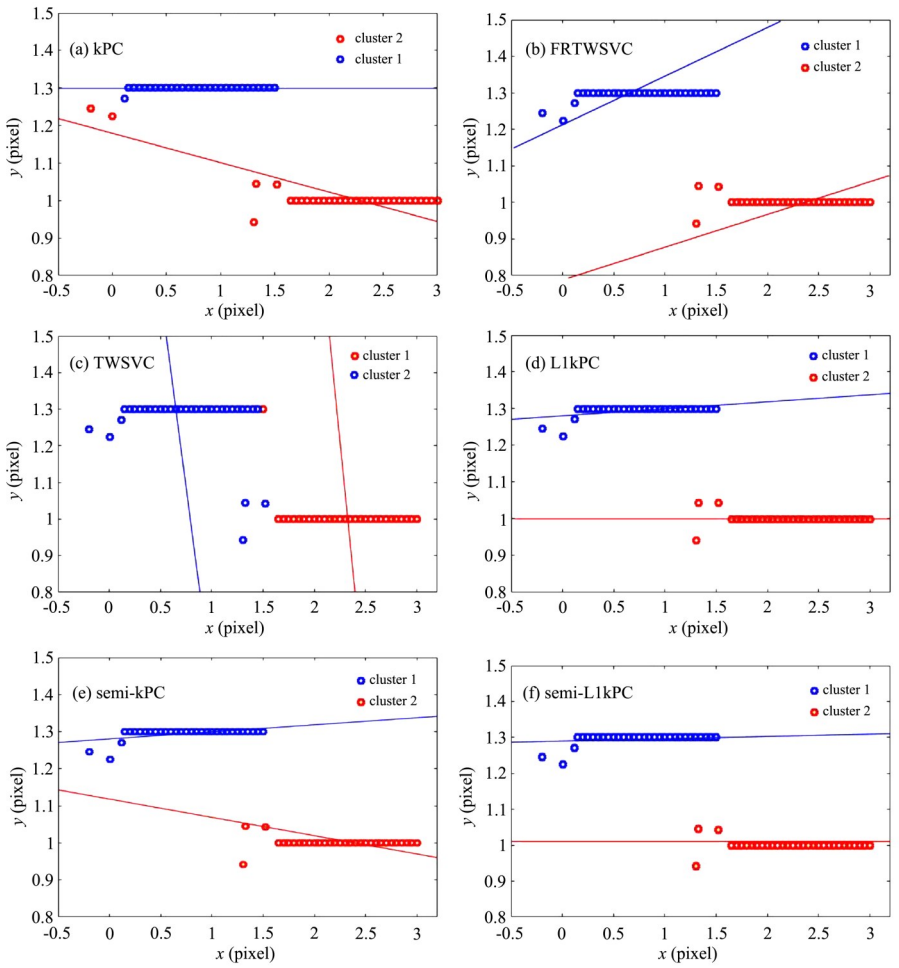

图2　六种算法的鲁棒性测试

Fig. 2　Robustness test of six algorithms

表1　UCI数据集的基本信息

Table 1　Main information of UCI datasets

数据集	样本数	维数	类别信息
Ecoil	332	6	8
Glass	214	9	6
Haberman	306	3	2
Vowel	528	10	11
Liver	345	6	2
Monk1	432	6	2
PID	768	8	2
User	403	5	4
Mushroom	8124	22	2

L1kPC,半监督的方法(semi-kPC和semi-L1kPC)也优于无监督(kPC,TWSVC,FRTWSVC和L1kPC).不同范数的聚类方法性能比较上,在八个数据集上FRTWSVC性能优于TWSVC,在七个数据集上L1kPC性能优于kPC,在六个数据集上semi-L1kPC优于semi-kPC.

综合以上信息:(1)平面型聚类方法整体优于k-means,说明平面型聚类方法对未知分布的数据同样有效;数据集Haberman上该数据(三维)的可视化结果呈椭球形分布,更适合k-means的点原型聚类方法.表2的实验结果也反映出在该数据集上,平面型聚类方法弱于k-means.(2)引入监督信息的半监督方法,其聚类性能整体上要优于无监督方法,说明少量监督信息的利用对通过随机取值启动算法的无监督聚类,聚类性能有明显提升.同时亦发现,监督信息的利用并非全部有效,若利用的监督信息自身有误或者不能反映类

表 2 七种聚类方法的准确率比较

Table 2 Accuracies of seven clustering methods

数据集	*k*-means	kPC	L1kPC	TWSVC	FRTWSVC	semi-kPC	semi-L1kPC	无监督平均值	半监督平均值
Ecoil	71.17	50.80	75.00	84.97	87.32	89.25	**94.87**	74.52	92.06
Glass	25.00	62.65	56.67	65.56	66.92	**72.45**	64.00	62.95	68.23
Haberman	**75.00**	54.57	70.81	61.26	62.54	73.50	63.33	62.30	68.42
Vowel	22.73	83.13	**87.50**	83.09	83.25	53.13	56.67	84.24	54.90
Liver	51.28	50.31	62.32	51.32	51.54	50.31	**63.64**	53.87	56.98
Monk1	50.00	49.88	54.55	50.50	52.36	57.68	**64.29**	51.82	60.99
PID	59.52	58.80	67.74	54.43	55.94	68.80	**72.62**	59.23	70.71
User	44.23	62.09	45.98	61.17	**68.36**	55.65	51.16	59.40	53.41
Mushroom	85.00	89.40	90.94	89.49	90.24	90.40	**92.45**	90.02	91.43

别信息,效果反而不如随机方法.此外,采用随机的方法在机器学习得到广泛应用,一种可行的解释是,随机方法更有利于跳出局部极小解问题.(3)采用L1范数的聚类方法较L2范数更鲁棒,实验结果也反映出在大多数据集上,L1范数的聚类性能更好.然而在面对真实场景时,数据在采集或传输过程中不可避免会受到污染,只是污染程度未知,这也是鲁棒学习普遍关心的问题.

为进一步探究监督信息的多少对聚类性能的影响,下面的实验通过逐步加大监督信息的比例完成.限于篇幅,只在三个数据集上做了实验,实验结果分别如图3a至图3c所示,横轴表示有标样本个数,用以描述监督信息的多少,纵轴表示聚类准确率.监督信息的多少根据有标样本的个数来决定:在样本维数 $d\leqslant 8$ 时,有标样本个数分别取 $\{2,4,8,16,20\}$;当 $d>8$ 时,取 $\{2,8,16,20,32\}$,两组样本个数分别对应图3中的横轴刻度 n_1, n_2, n_3, n_4, n_5.

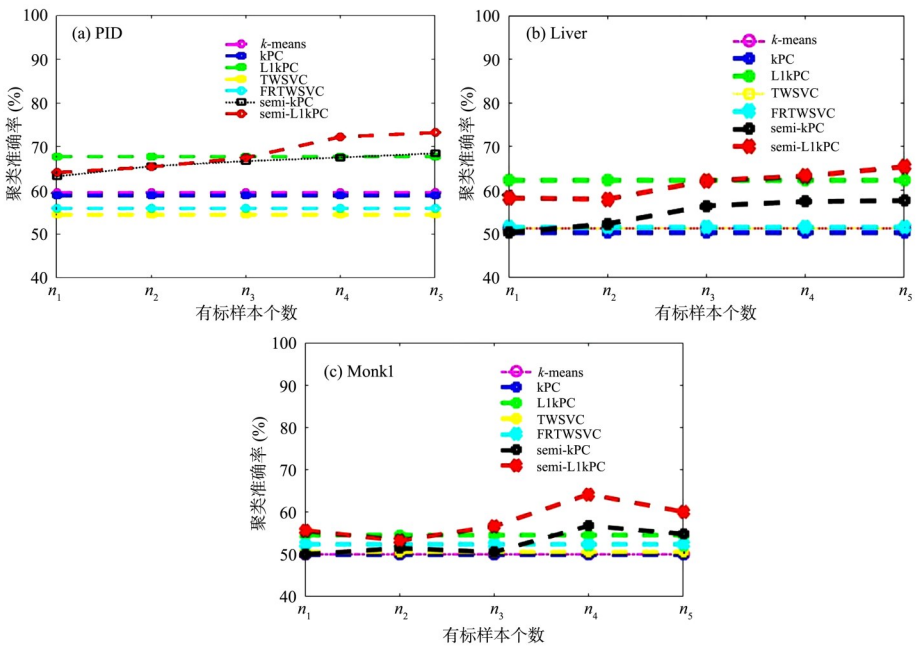

图 3 有标样本的个数对聚类准确率的影响

Fig.3 Accuracies on three datasets with labeled samples of different number

图3展示了七种算法在PID,Liver和Monk1数据集上的实验结果.整体上看,半监督学习的聚类结果随着有标样本的增加要优于无监督学习方法.此外,图3也反映出无监督学习在某些情况下已有可能优于半监督学习,其原因可能是样本的标签信息是错误的或者是有标样本选取的跨度不够大等原因造成的.

4 结　论

现有的半监督学习方法大多以点原型为基础设计,这类算法不能处理平面型数据分布.本文借鉴k平面聚类算法的思想,提出semi-kPC和semi-L1kPC.这两种聚类方法在处理平面分布型数据和XOR问题上大大提高了聚类准确率.针对semi-kPC算法不鲁棒的特点,本文进一步提出基于L1范数的semi-L1kPC.该方法在数据集中含有噪声或异常值时,聚类准确率比semi-kPC有明显的提升.但是在实验中也发现基于平面的半监督聚类算法依赖于查询点位置,当查询点的位置处于数据点的交叉处,则上述方法失效.今后的工作可以研究避免此类问题的优化算法.

参考文献

[1] 业巧林,许等平,张冬.基于深度学习特征和支持向量机的遥感图像分类.林业工程学报,2019,4(02):119－125.(Ye Q L,Xu D P,Zhang D. Remote sensing image classification based on deep learning features and support vector machine. Journal of Forestry Engineering,2019,4(02):119－125.)

[2] 许博鸣,刘晓峰,业巧林等.面向移动平台的深度学习复杂场景目标识别应用.陕西师范大学学报(自然科学版),2019,47(05):10－15.(Xu B M,Liu X F,Ye Q L,et al. A deep learning based object detection application for mobile platform in complex scenes. Journal of Shaanxi Normal University (Nature Science Edition),2019,47(5):10－15.)

[3] 刘建伟,刘媛,罗雄麟.半监督学习方法.计算机学报,2015,38(8):1592－1617.(Liu J W,Liu Y,Luo X L. Semi-supervised learning methods. Chinese Journal of Computers,2015,38(8):1592－1617.)

[4] Li Y F,Kwok J T,Zhou Z H. Semi-supervised learning using label mean//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM Press, 2009: 633－640.

[5] 张云斌,张春梅,周千琪等.基于L1范数和k近邻叠加图的半监督分类算法.模式识别与人工智能,2016,29(9):850－855.(Zhang Y B,Zhang C M,Zhou Q Q,et al. Semi-supervised classification algorithm based on L1-norm and KNN superposition graph. Pattern Recognition and Artificial Intelligence,2016,29(9):850－855.)

[6] 马蕾,汪西莉.基于支持向量机协同训练的半监督回归.计算机工程与应用,2011,47(3):177－180.(Ma L,Wang X L. Semi-supervised regression based on support vector machine co-training. Computer Engineering and Application,2011,47(3):177－180.)

[7] 吕峰,柴变芳,李文斌等.一种主动半监督K-means聚类算法的改进策略.南京师范大学学报(工程技术版),2018,18(2):56－62.(Lü F,Chai B F,Li W B,et al. An Improved strategy of active semi-supervision k-means clustering algorithm. Journal of Nanjing Normal University (Engineering and Technology Edition),2018,18(2):56－62.)

[8] 方玲,陈松灿.结合特征偏好的半监督聚类学习.计算机科学与探索,2015,9(1):105－111.(Fang L,Chen S C. Semi-supervised clustering learning combined with feature preferences. Journal of Frontiers of Computer Science & Technology, 2015, 9(1): 105－111.)

[9] 张春涛,郭皎,徐家良.基于稀疏表示的半监督降维方法.计算机工程与应用,2011,47(20):181－183,187.(Zhang C T,Guo J,Xu J L. Semi-supervised dimensionality reduction based on sparsity representation. Computer Engineering and Applications,2011,47(20):181－183,187.)

[10] Basu S,Banerjee A,Mooney R J. Semi-supervised clustering by seeding//Machine Learning, Proceedings of the Nineteenth International Conference. Sydney, Australia: University of New South Wales,2002.

[11] 高滢,刘大有,齐红等.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.(Gao Y,Liu D Y,Qi H,et al. Semi-supervised k-means clustering algorithm for multi-type relational data. Journal of Software, 2008, 19(11): 2814－2821.)

[12] Bradley P S, Mangasarian O L. K-plane clustering. Journal of Global Optimization, 2000, 16(1): 23－32.

[13] Fung G M, Mangasarian O L. Multicategory proximal support vector machine classifiers. Machine Learning, 2005, 59(1－2): 77－97.

[14] Mangasarian O L, Wild E W. Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 69－74.

[15] Jayadeva, Khemchandani R, Chandra S. Twin support vector machines for pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905－910.

[16] Wang Z, Shao Y H, Bai L, et al. Twin support vector machine for clustering. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2583－2588.

[17] Ye Q L, Zhao H H, Li Z C, et al. L1-norm distance minimization-based fast robust twin support vector *k*-plane clustering. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(9): 4494－4503.

[18] 徐庆伶, 汪西莉. 一种基于支持向量机的半监督分类方法. 计算机技术与发展, 2010, 20(10): 115－117, 121. (Xu Q L, Wang X L. A Novel semi-supervised classification method based on SVM. Computer Technology and Development, 2010, 20(10): 115－117, 121.)

[19] 杨绪兵, 潘志松, 陈松灿. 半监督型广义特征值最接近支持向量机. 模式识别与人工智能, 2009, 22(3): 349－353. (Yang X B, Pan Z S, Chen S C. Semi-supervised proximal support vector machine via generalized eigenvalues. Pattern Recognition and Artificial Intelligence, 2009, 22(3): 349－353.)

[20] Rastogi R, Pal A. Fuzzy semi-supervised weighted linear loss twin support vector clustering. Knowledge-Based Systems, 2019, 165: 132－148.

[21] 寇振宇, 杨绪兵, 张福全等. L1范数最大间隔分类器设计. 南京师范大学学报(自然科学版), 2018, 41(4): 59－64. (Kou Z Y, Yang X B, Zhang F Q, et al. Design of L1 norm Maximum margin classifier. Journal of Nanjing Normal University (Natural science Edition), 2018, 41(4): 59－64.)

[22] Yang H X, Yang X B, Zhang F Q, et al. Infinite norm large margin classifier. International Journal of Machine Learning and Cybernetics, 2019, doi: 10.1007/s13042-018-0881-y.

[23] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Journal of Intelligent Information Systems, 2001, 17(2－3): 107－145.

(责任编辑 杨可盛)