



文章编号:1001-9081(2019)12-3462-05

DOI:10.11772/j.issn.1001-9081.2019050813

结合支持向量机与半监督 K-means 的新型学习算法

杜 阳, 姜 震*, 冯路捷

(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

(* 通信作者电子邮箱: jiangz@ujs.edu.cn)

摘要: 半监督学习结合少量有标签样本和大量无标签样本, 可以有效提高算法的泛化性能。传统的半监督支持向量机(SVM)算法在目标函数中引入无标签样本的依赖项来推动决策面通过低密度区域, 但往往带来高计算复杂度和局部最优解等问题。同时, 半监督 K-means 算法面临着如何有效利用监督信息进行质心的初始化及更新等问题。针对上述问题, 提出了一种结合 SVM 和半监督 K-means 的新型学习算法(SKAS)。首先, 提出一种改进的半监督 K-means 算法, 从距离度量和质心迭代两个方面进行了改进; 然后, 设计了一种融合算法将半监督 K-means 算法与 SVM 相结合以进一步提升算法性能。在 6 个 UCI 数据集上的实验结果表明, 所提算法在其中 5 个数据集上的运行结果都优于当前先进的半监督 SVM 算法和半监督 K-means 算法, 且拥有最高的平均准确率。

关键词: 支持向量机; K-means; 半监督聚类; 分类; 融合

中图分类号: TP181; TP301.6 **文献标志码:**A

Novel learning algorithm combining support vector machine and semi-supervised K-means

DU Yang, JIANG Zhen*, FENG Lujie

(College of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang Jiangsu 212013, China)

Abstract: Semi-supervised learning can effectively improve the generalization performance of algorithm by combining a few labeled samples and large number of unlabeled samples. The traditional semi-supervised Support Vector Machine (SVM) algorithm introduces unlabeled sample dependencies into the objective function to drive the decision-making surface through the low-density region, but it often brings problems such as high computational complexity and local optimal solution. At the same time, semi-supervised K-means algorithm faces the problems of how to effectively use the supervised information to initialize and update the centroid. To solve these problems, a novel learning algorithm of Semi-supervised K-means Assisted SVM (SKAS) was proposed. Firstly, an improved semi-supervised K-means algorithm was proposed, which was improved from two aspects: distance measurement and centroid iteration. Then, a fusion algorithm was designed to combine semi-supervised K-means algorithm with SVM in order to further improve the performance of the algorithm. The experimental results on six UCI datasets show that, the proposed method outperforms the current advanced semi-supervised SVM and semi-supervised K-means algorithms on five datasets and has the highest average accuracy.

Key words: Support Vector Machine (SVM); K-means; semi-supervised clustering; classification; fusion

0 引言

传统的机器学习算法需要大量的有标签样本作为训练集, 但现实生活中大量数据往往是没有被标注的, 人工标注数据的代价太高。半监督学习^[1-3]则利用大量无标签样本和少量有标签样本来提高学习模型的泛化性能, 主要可分为两大类:

1) 半监督分类算法利用无标签样本结合有标签样本进行模型训练, 获得性能更优的分类器, 弥补有标签样本不足的缺陷。其中半监督支持向量机(Support Vector Machine, SVM)^[4-7]是目前应用较为广泛的一种半监督分类算法, 其主要思想是在同时考虑有标记样本和未标记样本的前提下, 找到最大间隔划分超平面并穿过数据低密度区域。大量无标签样本的引入提高了算法的复杂度, 并且容易陷入局部最优解。

半监督 SVM 集成是^[8-10]当前的一个研究热点, 通过集成多个半监督 SVM 基分类器来进一步提高泛化性能; 但仍面临着算法复杂性和局部最优解等问题。

2) 半监督聚类算法通过利用额外的监督信息来获得更好的聚类效果。目前所用的监督信息主要有两种形式: 第一种形式是“必连”(must-link)与“勿连”(cannot-link), 即两个样本属于同一类为“必连”, 不属于同一类则为“勿连”^[11]; 第二种形式是利用少量样本的类别标签, 即用有标签样本初始化 K 值和质心^[12]。但簇的个数不一定等于类别数以及质心迭代等问题依然对算法性能有着较大的影响。

半监督分类和聚类分别从不同的角度结合有标签样本和无标签样本进行样本的划分, 将二者结合是提高学习性能的一种可行方向, 但是当前类似的研究极少。本文提出了一种结合 SVM 和半监督 K-means 的新型学习算法(novel learning

收稿日期:2019-05-14; 修回日期:2019-07-23; 录用日期:2019-07-25。

基金项目: 国家自然科学基金资助项目(61672268); 江苏大学高级人才科研启动基金资助项目(14JDG036)。

作者简介: 杜阳(1994—), 男, 江苏扬州人, 硕士研究生, 主要研究方向: 机器学习; 姜震(1976—), 男, 山东烟台人, 副教授, 博士, 主要研究方向: 机器学习; 冯路捷(1996—), 女, 江苏淮安人, 硕士研究生, 主要研究方向: 机器学习。



algorithm of Semi-supervised K-means Assisted SVM, SKAS)。该算法融合了 SVM 和半监督 K-means (Semi-Supervised K-means, SSK) 的预测结果,通过二者的优势互补提升了算法的分类性能。特别地,从距离度量和质心迭代两个方面对半监督 K-means 算法进行了改进,进一步提高了算法的泛化性能。

1 相关工作

1.1 半监督 SVM

半监督 SVM 是目前半监督分类算法中较流行的一种分类算法。其中,半监督 SVM 的目标函数优化问题是一个混合整数规划问题,难以有效地解决。目前,针对该问题人们已经提出了各种方法,经典的方法有:Belkin 等^[4]提出的 Laplacian SVM 算法,Joachims 等^[5]提出的 Transductive SVM 算法,Chapelle 等^[6]提出的半监督支持向量机(Semi-Supervised Support Vector Machines, S3VMs)算法,以及 Li 等^[7]提出的安全半监督 SVM(Safe Semi-Supervised SVMs, S4VMs)算法等。

另一方面,一些研究者发现:半监督 SVM 与集成学习相结合可以进一步提高分类性能^[9-10]。Zhang 等^[8]提出了一种新的半监督 SVM 集成算法。该算法综合考虑了多种干扰因素对数据分布的影响,并提出了一种基于聚类评价方法的综合评价方法。

1.2 半监督聚类

目前,关于半监督聚类的研究主要基于约束信息^[13-16]。根据用户提供的约束信息,相应地修改聚类算法的目标函数来指导聚类过程。Wagstaff 等^[11]提出了 Constrained K-means 算法,根据样本集以及“必连”和“勿连”关系进行算法的迭代^[17-18]。Basu 等^[12]提出了 Constrained Seed K-means 算法,即将有标签样本作为“种子”,用它们初始化 K 个质心,并且在聚类簇迭代更新过程中不改变种子样本的簇隶属关系^[19-20]。Pelleg 等^[14]提出了线性时间约束向量化误差算法。Zeng 等^[15]引入有效损失函数克服了成对约束违反问题,提出了成对约束最大间隔聚类算法。何萍等^[16]研究成对约束对周围无约束样本点的影响,将在顶点上低层随机游走和在组件上高层随机游走相结合,提出了一种双层随机游走半监督聚类算法。

2 SKAS

本文提出了一种改进的半监督 K-means 算法,并结合 SVM 来提高分类算法的性能,其基本思想如图 1 所示。

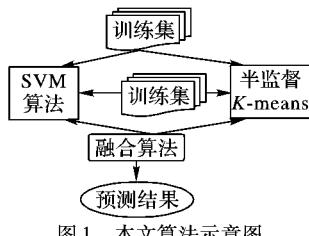


图 1 本文算法示意图

Fig. 1 Schematic diagram of proposed algorithm

设训练样本 D_i 、测试样本 D_u 、训练样本的标签 C 分别为:

$$\begin{aligned} D_i &= \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \\ D_u &= \{(x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \dots, (x_{m+l}, y_{m+l})\} \end{aligned}$$

$$C = \{C_1, C_2, \dots, C_K\}$$

其中: m 为训练样本的个数; l 为测试样本的个数; K 为类别个数。

2.1 SVM 算法

2.1.1 训练

基于训练集 D_i ,在样本空间中找到划分超平面,将不同类别的样本分开。得到基于 SVM 训练的模型。

$$\begin{aligned} \min_{w,b,\xi} &= \frac{1}{2} \|w\|^2 + c \sum_{i=1}^m \xi_i \\ \text{s. t. } & y_i((w^T x_i) + b) \geq 1 - \xi_i; \quad i = 1, 2, \dots, m \\ & \xi_i \geq 0; \quad i = 1, 2, \dots, m \end{aligned} \quad (1)$$

其中: w 是法向量,决定了超平面的方向; b 是位移项; m 是样本个数; ξ_i 为标准数据上的松弛变量; c 是给定的惩罚因子。

2.1.2 测试

SVM 的决策函数 $f(x)$ 为:

$$f(x) = \text{sgn}(w^T \psi(x) + b) = \text{sgn}\left(\sum_{i=1}^l y_i \alpha_i K(x_i, x) + b\right) \quad (2)$$

式(2)第二个等式右边括号里面的量是一个与超平面的距离成正比的量。这种算法的思想是离超平面越远的点认为分对的可能性越大。

基于上述原理,利用 sigmoid 函数将决策函数 $f(x)$ 投射到 $[0, 1]$ 上,得到 SVM 输出样本预测概率值的计算式为:

$$\text{Pr}(y = 1 | x) \approx P_{A,B}(f) = \frac{1}{1 + \exp(Af + B)} \quad (3)$$

其中 f 为式(2)中的 $f(x)$ 。

式(3)中的 A 和 B 值这两个参数是用来调整映射值的大小,这两个参数是未知的,需要估计,计算式如下:

$$\min \left\{ - \sum_i (t_i \ln(p_i) + (1 - t_i) \ln(1 - p_i)) \right\} \quad (4)$$

其中:

$$\begin{cases} P_i = \frac{1}{1 + \exp(Af_i + B)} \\ t_+ = \frac{N_+ + 1}{N_- + 2} \\ t_- = \frac{1}{N_- + 2} \end{cases} \quad (5)$$

式中: t_+ 表示样本属于正类; t_- 表示样本属于负类。

在处理多分类问题上采用 one-versus-one 法,在任意两类样本之间找到一个超平面,样本属于每个类有一个概率函数。因此 K 个类别的样本就需要设计 $K(K - 1)/2$ 个超平面。当对一个未知样本进行分类时,根据投票法原则,最后得票最多的类别即为该未知样本的类别。

2.1.3 置信度计算

为了计算预测样本的置信度,最直接的方法是将数据预测类别的概率作为权重,选择最大的类预测概率 $P_{\text{SVM}}(y = c_{\max_j} | x_j)$ 作为置信度 $C_{\text{SVM}}(x_j)$,即:

$$C_{\text{SVM}}(x_j) = P_{\text{SVM}}(y = c_{\max_j} | x_j) \quad (6)$$

但仅将类的最大预测概率作为置信度不够合理,因此采用一种新的置信度计算方法^[21],其通过类别最大的概率与第二大概率的差值来衡量置信度,即:

$$C_{\text{SVM}}(x_j) = P_{\text{SVM}}(y = c_{\max_j} | x_j) - P_{\text{SVM}}(y = c_{\text{sub_max}_j} | x_j) \quad (7)$$



这种置信度计算方法可以针对类重叠区域的数据,有效解决 SVM 在类重叠情况下性能下降的问题。

2.2 半监督 K-means 算法

2.2.1 初始化质心

K-means 算法有着 K 值和初始质心难以确定的问题,一般认为:同一个簇内的样本应该属于一个类,而同一个类的样本可能位于不同的簇。本文假定簇个数 K 等于类别数,若一个类对应多个簇,则将这些簇当作一个大簇的子簇进行处理,从而在寻找最优的 K 值的过程中实现算法简化。因此,本文首先根据训练集中的类别确定 K 值以及每个簇的标签。其次,根据训练集中每个样本的标签,把它们依次划分入每一个簇中,计算每个簇的初始质心:

$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{x_i \in C_i} \mathbf{x}_i \quad (8)$$

其中 C_i 表示当前样本属于的簇。

确定了 K 值并初始化质心后,计算样本与各个质心的距离,将样本划入相应的簇并更新质心,直到满足某个停止条件为止。

对于给定的质心 $\boldsymbol{\mu}$ 和样本 \mathbf{x} ,传统的距离计算公式为:

$$distance(\boldsymbol{\mu}, \mathbf{x}) = \sqrt{\sum_{d=1}^D (\boldsymbol{\mu}^d - \mathbf{x}^d)^2} \quad (9)$$

由于数据集中各类别之间的样本数会存在差异,训练过程会向样本数较多的类别倾斜。针对该问题,本文提出了一种基于权重的改进距离公式如下:

$$distance(\boldsymbol{\mu}, \mathbf{x}) = \frac{V_i}{V} \sqrt{\sum_{d=1}^D (\boldsymbol{\mu}^d - \mathbf{x}^d)^2} \quad (10)$$

其中: V_i 代表训练集中质心 i 所属的类别中样本的个数; V 代表训练集中所有样本的个数; D 代表数据集中样本的维度。

根据式(10),将样本划入相应的簇,并确定其所属类别。

2.2.2 质心迭代的终止条件

传统的聚类学习中,质心迭代的终止条件往往有两种:第一种是预先设置好迭代次数;第二种是计算迭代前后的误差,若小于某个值,则终止迭代。这种迭代的终止条件往往会造成迭代次数超过最优迭代次数时,算法的性能会急剧下降。特别地,在半监督 K-means 中,由于簇中的噪声会影响到质心的计算,并可能造成算法性能的下降。因此,本文提出一种新的迭代终止条件,根据 \mathbf{D}_l 上预测结果的准确率进行判断。

$$ACC(\mathbf{D}_l) < old_ACC(\mathbf{D}_l) \quad (11)$$

其中: ACC 为基于当前质心的预测准确率; old_ACC 为基于上一轮质心的准确率。当准确率下降即满足式(11)时,表明受簇内噪声的影响,继续迭代所产生的质心会降低算法性能。此时,停止迭代并恢复上一轮的质心。

该方法兼顾了聚类的传统指标误差平方和(Sum of Squares of Errors, SSE)和分类的准确度,在实验中表现出比较明显的优势。

2.2.3 置信度计算

$P[i]$ 代表样本 i 属于当前簇的概率,其计算式为:

$$P[i] = (1/d[cluster[i]])/sum \quad (12)$$

其中: $cluster[i]$ 代表样本 i 属于的簇标号; $d[j]$ 代表第 j 个簇中,当前样本 i 到达质心的距离; $sum = \sum_{j=1}^K \frac{1}{d[j]}$ 代表当前样

本 i 到达每个质心的距离的倒数和。

置信度的计算式如下:

$$\begin{aligned} C_{SKAS}(\mathbf{x}_j) &= P_{SKAS}(y = c_{max_j} | \mathbf{x}_j) - \\ &P_{SKAS}(y = c_{sub_max_j} | \mathbf{x}_j) \end{aligned} \quad (13)$$

2.3 融合算法

结合 2.1 节和 2.2 节中的算法,并为了进一步提高准确率,将 SVM 和半监督 K-means 结合起来进行最终的预测。SVM 和半监督 K-means 的预测结果都转化为概率的形式,但二者预测的概率并不在同一尺度上,直接把预测的结果结合起来并不能得到满意的结果。因此,对 SVM 和半监督 K-means 预测的置信度做了归一化处理,然后给出了最终的分类结果。

$$P(y_i | \mathbf{x}_i)_{SKAS} = \begin{cases} P(y_i | \mathbf{x}_i)_{SKK}, & \frac{\mu \cdot C_{SKK}(\mathbf{x}_i)}{\sum_{x_j \in U} C_{SKK}(\mathbf{x}_j)} > \frac{(1-\mu)C_{SVM}(\mathbf{x}_i)}{\sum_{x_j \in U} C_{SVM}(\mathbf{x}_j)} \\ P(y_i | \mathbf{x}_i)_{SVM}, & \text{其他} \end{cases} \quad (14)$$

其中, $\mu \in [0,1]$,是一个用来调节 SVM 和半监督 K-means 权重的参数。为了获得更好的效果,根据 SVM 和半监督 K-means 在训练集上的准确率来调节其权重,如式(15)所示:

$$\mu = W_1 / (W_1 + W_2) \quad (15)$$

其中, W_1 、 W_2 分别代表 SVM 和半监督 K-means 对有标签样本所属类别预测的准确率。

2.4 SKAS

SKAS 的流程如下:

输入 $\mathbf{D}_l = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}, \mathbf{D}_u = \{(\mathbf{x}_{m+1}, y_{m+1}), (\mathbf{x}_{m+2}, y_{m+2}), \dots, (\mathbf{x}_{m+l}, y_{m+l})\}$;

输出 \mathbf{D}_u 中每个样本的预测标签。

步骤 1 在 \mathbf{D}_l 上训练 SVM,然后分类 \mathbf{D}_u 中样本,根据式(7)得到每个样本的置信度。

步骤 2 根据 \mathbf{D}_l 中的有标签样本初始化 K 个质心,并根据距离公式(10)将 $\mathbf{D}_l \cup \mathbf{D}_u$ 中的所有样本划分到最近的簇中。

步骤 3 重复步骤 4 ~ 5 直到质心不再变化或满足式(11)。

步骤 4 根据式(8)更新每个簇里面的质心。

步骤 5 根据距离公式(10)重新把 $\mathbf{D}_l \cup \mathbf{D}_u$ 中所有的样本划分到最近的簇中。

步骤 6 根据迭代终止后每个簇的质心,把 \mathbf{D}_u 中样本重新划分到最近的簇中,根据式(13)得到每个样本的置信度。

步骤 7 对 SVM 和半监督 K-means 的预测结果进行融合,根据式(14)计算 \mathbf{D}_u 中样本所属类别及其概率。

3 实验与结果分析

3.1 数据集

针对本文提出的算法模型,使用来自 UCI 的六个数据集作为性能测试数据,随机选取 30% 作为训练集。同时,为了防止类别不平衡或样本数量较少导致训练集未能覆盖所有类别的情况,当随机选取的训练集中缺少某个类别的样本时,则向训练集中补充一个缺失类别的样本,从而保证 K 值等于训练集中类别的个数。数据集的详细信息如表 1 所示。



表 1 实验所用数据集
Tab. 1 Datasets for experiments

数据集	样本数	特征数	类别数	不平衡率
iris	150	4	3	1:2.00
wine	178	13	3	1:2.70
glass	214	9	7	1:22.80
thyroid	7200	21	2	1:42.40
balance	625	4	3	1:11.80
vehicle	846	18	4	1:3.25

3.2 结果分析

为了评估 SKAS 的分类性能,在标准 SVM 的基础上加入 S4VMs^[7]、EnsembleS3VM^[8] 和 Constrained Seed K-means 算法^[12]进行实验对比。对于每种算法,均使用与 SKAS 相同的训练预测方法,即基于 LIBSVM 使用五折交叉检验,所有算法均使用五次结果的平均值作为最终结果;其五折交叉验证通过调用 LIBSVM 软件包中的 grid 函数实现,并对特征值进行了归一化的处理,通过调用 svm-scale 来实现。

表 2 给出了四种不同算法对六个数据集进行训练预测的实验结果。实验采用跟文献[8]相同的参数设置,对比后发现:在所有数据集中,SKAS 中的五个数据集具有最高的准确率,剩下一个接近最好算法的准确率,并且 SKAS 的平均准确率为 75.77,优于其他三种算法。实验结果表明 SKAS 能够提高预测模型的准确率。

表 2 不同算法的准确率对比
Tab. 2 Accuracy comparison of different algorithms

数据集	SVM	S4VMs	EnsembleS3VM	Constrained Seed K-means	SKAS
iris	85.79	59.16	86.82	96.70	98.65
wine	83.72	65.88	88.01	96.90	97.20
glass	29.51	22.22	27.12	58.20	65.87
thyroid	60.17	67.75	67.85	52.08	77.75
balance	59.56	48.05	60.60	42.81	58.12
vehicle	47.85	39.30	48.05	38.2	57.05
平均值	61.11	50.39	63.08	64.15	75.77

选择其中三个数据集 iris、glass 和 thyroid,分别给出它们的准确率在 SVM、Constrained Seed K-means 和本文提出的 SKAS 迭代训练过程中的变化情况。

首先,由图 2 可以看出,本文提出的 SKAS 在迭代开始的准确率都有上升,并在到达峰值后开始下降,峰值点在图中已标出。根据 2.2.2 节中本文提出的新的迭代终止条件,发现图 2(a)至图 2(c)中 SKAS 的峰值即为迭代的终止点,进一步说明,根据新设置的迭代终止条件提前终止迭代可以取得更好的聚类效果。

其次,从图 2 可以发现,SKAS 的准确率均高于 SVM 算法和半监督 K-means 算法。这也表明了本文提出的融合算法综合了 SVM 和半监督 K-means 的预测结果,确实能有效地提高模型的泛化性能。

图 2(c)中,SKAS 的准确率远远高于其他两种算法,主要是因为 thyroid 的样本数量较大,且样本的不平衡率较高。本文提出的算法有效地解决了在样本数量较多以及类别不平衡时,SVM 算法分类性能下降的问题。此外,图 2(c)中半监督

K-means 的准确率低于 SVM,分析其原因可能是 thyroid 的特征数较多,类重叠现象较为严重。

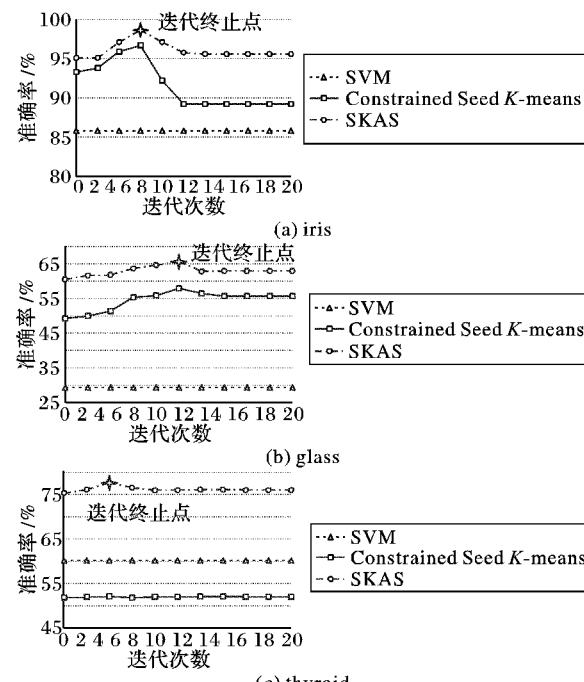


图 2 三个算法在不同数据集上迭代时准确率变化

Fig. 2 Accuracy change of three algorithms while iterating on different datasets

4 结语

本文对半监督 K-means 算法进行了相应改进,提出了一种结合 SVM 与半监督 K-means 算法的新型学习算法——SKAS,该算法可以实现半监督聚类和分类算法的优势互补。实验结果表明,SKAS 相较于对比算法取得了更好的性能结果,特别是在样本数量较大的情况下,本文算法的优势更为明显。

为进一步优化学习算法,我们后续工作将主要集中在半监督 K-means 算法的进一步改进上,特别是簇的数量与实际类别数量不一致的问题。此外,我们还将关注类别不平衡问题,研究通过改进算法的目标函数以提高小类别样本的查全率。

参考文献 (References)

- ZHU X, GOLDBERG A B. Introduction to Semi-Supervised Learning [M]. San Rafael: Morgan and Claypool Publishers, 2009: 130.
- ZHANG Z, SCHULLER B. Semi-supervised learning helps in sound event classification [C]// Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2012: 333 – 336.
- ZHU X. Semi-supervised learning [C]// Proceedings of the 2011 International Joint Conference on Artificial Intelligence. Menlo Park: AAAI, 2011: 1142 – 1147.
- BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006, 7: 2399 – 2434.
- JOACHIMS T. Transductive inference for text classification using support vector machines [C]// Advances in Large Margin Classifiers. Cambridge: MIT Press, 2000: 307 – 316.



- support vector machines [C] // Proceedings of the 1999 International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 200 – 209.
- [6] CHAPELLE O, CHI M, ZIEN A. A continuation method for semi-supervised SVMs [C] // Proceedings of the 2006 Twenty-Third International Conference on Machine Learning. New York: ACM, 2006: 185 – 192.
- [7] LI Y, ZHOU Z. Towards making unlabeled data never hurt [C] // Proceedings of the 28th International Conference on Machine Learning. Madison: Omnipress, 2011: 1081 – 1088.
- [8] ZHANG D, JIAO L, BAI X, et al. A robust semi-supervised SVM via ensemble learning [J]. Applied Soft Computing, 2018, 65: 632 – 643.
- [9] ZHOU Z. When semi-supervised learning meets ensemble learning [C] // Proceedings of the 8th International Workshop on Multiple Classifier Systems, LNCS 5519. Berlin: Springer, 2009: 529 – 538.
- [10] PLUMPTON C O, KUNCHEVA L I, OOSTERHOF N N, et al. Naive random subspace ensemble with linear classifiers for real-time classification of fMRI data [J]. Pattern Recognition, 2012, 45(6): 2101 – 2108.
- [11] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained K -means clustering with background knowledge [C] // Proceedings of the 8th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2001: 577 – 584.
- [12] BASU S, BANERJEE A, MOONEY R J. Semi-supervised clustering by seeding [C] // Proceedings of the 9th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 27 – 34.
- [13] DING S, JIA H, ZHANG L, et al. Research of semi-supervised spectral clustering algorithm based on pairwise constraints [J]. Neural Computing and Applications, 2014, 24(1): 211 – 219.
- [14] PELLEG D, BARAS D. K -means with large and noisy constraint sets [C] // Proceedings of the 18th European Conference on Machine Learning. Berlin: Springer, 2007: 674 – 682.
- [15] ZENG H, CHEUNG Y. Semi-supervised maximum margin clustering with pairwise constraints [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 926 – 939.
- [16] 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. (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.)
- [17] STEINLEY D, BRUSCO M J. K -means clustering and mixture model clustering: reply to McLachlan (2011) and Vermunt (2011) [J]. Psychological Methods, 2011, 16(1): 89 – 92.
- [18] HONG Y, KWONG S. Learning assignment order of instances for the constrained K -means clustering algorithm [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 568 – 574.
- [19] LI K, ZHANG C, CAO Z. Semi-supervised kernel clustering algorithm based on seed set [C] // Proceedings of the 2009 Asia-Pacific Conference on Information Processing. Piscataway: IEEE, 2009: 169 – 172.
- [20] GU L, SUN F. Two novel kernel-based semi-supervised clustering methods by seeding [C] // Proceedings of the 2009 Chinese Conference on Pattern Recognition. Piscataway: IEEE, 2009: 1 – 5.
- [21] 尹玉, 詹永照, 姜震. 伪标签置信选择的半监督集成学习视频语义检测 [J]. 计算机应用, 2019, 39(8): 2204 – 2209. (YIN Y, ZHAN Y Z, JIANG Z. Semi-supervised integrated learning video semantic detection with false label confidence selection [J]. Journal of Computer Applications, 2019, 39(8): 2204 – 2209.)

This work is partially supported by the National Natural Science Foundation of China (61672268), the Research Initiation Fund for Senior Talents of Jiangsu University (14JDG036).

DU Yang, born in 1994, M. S. candidate. His research interests include machine learning.

JIANG Zhen, born in 1976, Ph. D., associate professor. His research interests include machine learning.

FENG Lujie, born in 1996, M. S. candidate. Her research interests include machine learning.