计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200103

基于加权样本和共识率的标记传播算法

储 杰 张正军 汤鑫瑶 黄振生
南京理工大学理学院 南京 210094
(jiecnj@163.com)

摘 要 标记传播是使用最广泛的半监督分类方法之一。基于共识率的标记传播算法(Consensus Rate-based Label Propagation,CRLP)通过汇总多个聚类方法以合并数据各种属性得到的共识率来构造图。然而,CRLP算法与大多数基于图的半监督分类方法一样,在图中将每个标记样本视为同等重要,它们主要通过优化图的结构来提高算法的性能。事实上,样本不一定是均匀分布的,不同的样本在算法中的重要性也是不同的,并且CRLP算法容易受聚类数目和聚类方法的影响,对低维数据的适应性不足。针对这些问题,文中提出了一种基于加权样本和共识率的标记传播算法(Label Propagation Algorithm Based on Weighted Samples and Consensus-Rate,WSCRLP)。WSCRLP算法首先对数据集进行多次聚类,以探索样本的结构,并结合共识率和样本的局部信息构造图;然后为不同分布的标记样本分配不同的权重;最后基于构造的图和加权样本进行半监督分类。在真实数据集上的实验表明,WSCRLP算法对标记样本进行加权和构造图的方法可以显著提高分类准确率,在84%的实验中都优于对比方法。相比CRLP算法,WSCRLP算法不仅具有更好的性能,而且对输入参数具有鲁棒性。

关键词:加权样本;共识率;标记传播;半监督分类
中图法分类号 TP301.6

Label Propagation Algorithm Based on Weighted Samples and Consensus-rate

CHU Jie,ZHANG Zheng-jun,TANG Xin-yao and HUANG Zhen-sheng
School of Science,Nanjing University of Science and Technology,Nanjing 210094,China

Abstract Label Propagation is one of the most widely used semi-supervised classification methods. Consensus rate-based label propagation(CRLP) algorithm constructs the graph by summarizing multiple clustering solutions to incorporate various properties of the data. Like most graph-based semi-supervised classification method,CRLP focuses on optimizing the graph to improve the performance. In fact,samples are not always evenly distributed. The importance of different samples in the algorithm is different. CRLP algorithm is easily affected by the numbers of clustering and the clustering methods,and it is not adaptable to low-dimensional data. To deal with these problems,a label propagation algorithm based on weighted samples and consensus-rate (WSCRLP) is proposed. WSCRLP firstly clusters the dataset multiple times to explore the structure of sample and combines the consensus-rate and the local information of the sample to construct a graph. Secondly,different weights are assigned to labeled samples with different distributions. Finally,semi-supervised classification is performed based on constructed graph and weighted samples. Experiments on real datasets show that the WSCRLP of weighting and constructing graphs on labeled samples can significantly improve classification accuracy,and is superior to other compared methods in 84% of the experiments. Compared with CRLP,WSCRLP not only has better performance,but also is robust to input parameters.

Keywords Weighted samples,Consensus-rate,Label propagation,Semi-supervised classification

1 引言

监督学习算法在有充足的标记数据时是十分有效的。然而,在许多现实应用领域,如网页和文本分类、图像和视频检索、医学数据处理等,往往很难获得大量标记数据。半监督学习方法能够利用标记数据和大量的未标记数据进行建模。因此,半监督学习方法[1-2]在许多领域得到了广泛应用。在过去的几十年里,研究者们已经提出了很多半监督学习方法,主要有基于图的方法[3-7]、生成模型[8]、协同训练[7-10]、半监督支持向量机[11]和自我训练[12]等。其中,基于图的半监督学习方法凭借其计算速度快、准确率高、灵活和易于实现的特点,成为了近年来机器学习研究中最受关注的研究方向之一。

基于图的半监督学习方法将标记和未标记样本作为图形中的顶点,并使用边的权重来编码样本之间的相似性。然而,

到稿日期:2019-12-16 返修日期:2020-04-29
基金项目:全国统计科学研究重大项目(2018LD01)
This work was supported by the National Statistical Science Research Major Program of China(2018LD01).
通信作者:张正军(zzjnj@163.com)

构建一个能准确反映样本潜在分布的图比较困难。鉴于此，大多数研究人员希望借助优化图形结构和边的权重来提高基于图的半监督学习算法的性能。Zhu 等[5]在带有标记和未标记样本的 K 最近邻（K-Nearest Neighbor，KNN）图上应用高斯域和调和函数（Gaussian field harmonic function，GFHF）来预测图中的未标记样本。Zhou 等[13]提出了基于局部和全局一致性的方法（Local and Global Consistency，LGC）。LGC 算法利用样本的局部和全局结构，以及标签在图形上的传播，来预测未标记样本的标签。不同于 GFHF 算法，LGC 算法允许在迭代过程中更改初始标签信息，因此可以更有效地处理带标签的噪声样本。Wang 等[14]提出的线性邻域传播算法（Linear Neighborhood Propagation，LNP），通过最小化样本间的线性重构误差来调整样本与其 K 个邻域样本之间的线性重构（Local Linear Regression，LLR）权重。Zhao 等[15]提出了一种基于紧凑图的半监督学习算法（Compact Graph based Semi-Supervised Learning，CGSSL），通过最小化样本、邻域样本以及互邻样本的重构误差来优化图边缘的权重。为了加强算法的判别能力，Wu 等[16]利用判别正则化的优势来构造图。最近，Yu 等[17]提出了一种基于共识率的标记传播算法（CRLP）。CRLP 算法利用多种聚类解决方案建立的共识矩阵作为样本间的相似性矩阵，通过对在随机子空间中随机选择的 K 值进行聚类来获得样本之间的共识率。因此，CRLP 算法可以容纳数据集的各种属性，并能更好地揭示数据的内在结构。

我们观察到大部分基于图的半监督学习算法会对标记样本分配相同的权重。然而，在真实的数据集中，样本不总是均匀分布的，不同的样本在算法中的重要性也是不同的，尤其是在数据集存在噪声标签的情形下。最近的一些研究也表明，对样本或者样本特征进行加权可以提高半监督分类算法的性能。例如，Das 等[18]首先使用核函数构造加权图，该核函数随样本之间的距离而衰减。Ren 等[19]利用多个聚类为样本分配权重，然后利用分配的权重调整聚类或样本之间的距离。Chen 等[20]认为接近决策边界的标记样本比远离决策边界的样本更重要，并提出了加权样本的半监督分类算法。

本文提出的算法探究是否可以通过对标记样本分配不同权重，而不是优化所构造图形的边缘权重或结构，来提高基于图的半监督分类算法的性能。以此为目的，我们提出了一种基于加权样本和共识率的标记传播算法（WSCRLP）。与以往的算法不同，WSCRLP 利用了 CRLP 算法以及集成聚类[19]中建立共识矩阵的方法，使用 K 均值聚类算法在不同的特征子空间中产生聚类结果，然后通过一个由每个样本间的共识率构成的共识矩阵来总结这些聚类结果，最后通过分析标记样本进行聚类的难易程度来定义它们的权重。进一步，针对 CRLP 利用共识率构造图时对低维数据适应性不足的缺陷，提出结合共识率和样本的局部信息来构造图。

2　相关工作

2.1　基于局部和全局一致性（LGC）

LGC 算法是基于图的半监督分类算法中的代表算法，其步骤如下：

（1）构建一个 K 最近邻图，在 K 最近邻图中样本间的边权重 W_{ij} 定义为：

$$e_{ij}=\begin{cases}1, & o_j\in K(o_i)\ \text{or}\ o_i\in K(o_j)\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

其中，$K(o_i)$ 和 $K(o_j)$ 表示样本 o_i 和样本 o_j 的 K 个近邻集合。

$$W_{ij}=\begin{cases}\exp\left(-\dfrac{d(o_i,o_j)^2}{2\sigma^2}\right), & \text{if } e_{ij}=1\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

其中，$d(o_i,o_j)^2$ 和 σ 分别是样本 o_i 和样本 o_j 之间的欧氏距离和 RBF 内核的带宽。

（2）由（1）得到的亲和矩阵 $\boldsymbol{W}$ 构造标记传播矩阵 $\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$，其中 $\boldsymbol{D}$ 是一个对角矩阵，$\boldsymbol{D}_{ii}=\sum_{m=1}^{n}\boldsymbol{W}_{im}$。

（3）未标记样本的标签通过迭代标签过程进行拟合，迭代计算式为：

$$\boldsymbol{F}(t+1)=\alpha\boldsymbol{S}\boldsymbol{F}(t)+(1-\alpha)\boldsymbol{Y} \tag{3}$$

其中，$\alpha\in(0,1)$ 是用户指定的参数。

（4）F^* 表示序列 $\{F(t)\}$ 的极限，每个未标记样本的标签 $y_i=\arg\max_{j\leqslant c}F^*_{ij}$。

LGC 算法对应的正则化框架如下：

$$\arg\min_{F}\Phi(\boldsymbol{F})=\frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{W}_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}\boldsymbol{F}_i-\frac{1}{\sqrt{D_{jj}}}\boldsymbol{F}_j\right\|^2+\sum_{i=1}^{l}\mu\|\boldsymbol{F}_i-\boldsymbol{Y}_i\|^2 \tag{4}$$

其中，$\mu=(1-\alpha)/\alpha$ 控制等式右侧第一项和第二项的重要性。可以看出，最优的 F 主要依赖于 $\boldsymbol{W}$，因此 LGC 算法的性能主要取决于用 $\boldsymbol{W}$ 描述的图的边缘结构和权重。然而，由于样本分布不均匀，对于位于密集区域中的样本，较小的 k 是合适的，但对于稀疏区域中的样本，较大的 k 可能更好。因此，如何选择合适的 k 和 α 对于 LGC 算法十分重要。

2.2　基于共识率的标记传播算法（CRLP）

CRLP 算法是最近提出的标记传播算法，用于分类的目标函数，类似于 LGC 算法，其不同于 LGC 算法之处在于构图方法不同。CRLP 算法利用多个聚类解决方案构造了一个由数据驱动的自适应图。首先，其利用多种聚类解决方案构建共识矩阵，使用随机子空间方法和可随机选择 K 值的 K-均值聚类算法生成多个聚类解决方案，然后将聚类结果总结为共识矩阵。

共识矩阵 $\boldsymbol{W}$ 中的元素称为共识率[17]，样本 o_i 和样本 o_j 之间的共识率定义为：

$$\gamma_{ij}=\frac{1}{B}\sum_{l=1}^{B}\sum_{k=1}^{k(l)}I(o_i,o_j)^{(l,k)} \tag{5}$$

其中，B 和 $k(l)$ 分别是随机子空间的数目和随机子空间 l 中簇的数目。指示函数 $I(o_i,o_j)^{(l,k)}$ 的定义如下：

$$I(o_i,o_j)^{(l,k)}=\begin{cases}1, & \text{if } o_i,o_j\in C_k^l \text{ and } i\neq j\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

其中，C_k^l 是随机子空间 l 中的第 k 个簇。共识率表示两个样本被聚类方法分到同一簇的可能性。然后，CRLP 算法利用共识矩阵 $\boldsymbol{W}$ 来构造标记传播矩阵以进行标记传播，算法的其他过程与 LGC 算法类似。

3　基于加权样本和共识率的标记传播算法（WSCRLP）

假设 $X=[x_1;x_2;\cdots;x_n]\in\mathbb{R}^{n\times d}$ 是 d 维空间中的 n 个样

本；$\boldsymbol{Y}\in\mathbb{R}^{n\times c}$是已知的标签指示矩阵，其每行对应一个样本。不失一般性，我们假设前 l 个样本已标记，剩余 $u=n-l$ 个样本未标记。我们提出的 WSCRLP 算法由 3 部分组成。首先使用随机子空间和 K-均值聚类算法产生样本之间的共识矩阵，再结合样本的局部信息构造一个图；然后对标记样本分配不同的权重；最后基于加权样本和构造的图进行标记传播。

3.1 构造图

将 X 划分成 B 个随机子空间，每个随机子空间中样本的特征由从 d 个特征中随机采样的 ω 个特征组成（d 表示特征总数），随机子空间方法独立构造多个特征子集，可以反映数据的不同属性和内部结构。由于 K-均值聚类算法相比其他聚类算法更加简单且有效，因此我们选择 K-均值聚类算法对 B 个随机子空间执行聚类过程，其他聚类算法也可以在这里使用。

获得聚类结果后，将它们汇总为共识矩阵，共识矩阵中的元素称为共识率[17]。样本x_i和样本x_j之间的共识率A_{ij}的定义如下：

$$A_{ij}=\frac{1}{B}\sum_{l=1}^{B}\sum_{k=1}^{k(l)}I(x_i,x_j)^{(l,k)} \tag{7}$$

其中，B 和 $k(l)$分别是随机子空间的数目和随机子空间 l 中簇的数目。指示函数 $I(x_i,x_j)^{(l,k)}$ 的定义如下：

$$I(x_i,x_j)^{(l,k)}=\begin{cases}1, & \text{if } x_i,x_j\in C_k^l \text{ and } i\neq j\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

其中，C_k^l是随机子空间 l 中的第 k 个簇。

基于共识率构造图的方法的衡量标准过多地考虑样本的全局信息，容易受到划分的随机子空间和聚类方法的影响，这在低维数据集上的表现更为明显。因此，我们提出结合样本的局部信息之间的共识率来构造图。样本x_i和样本x_j之间边的权重定义为：

$$W_{ij}=\begin{cases}\left(1+\exp\left(-\dfrac{d(x_i,x_j)^2}{2\sigma^2}\right)\right)\times A_{ij}, & \text{if } A_{ij}\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

其中，根据文献[5]将高斯核函数的带宽设置为 1.25。W_{ij}结合了多个聚类中总结出的共识率和样本间的距离，受初始错误的单个聚类的影响较小，并且有可能识别异常值和降低异常值对分类的影响，因为异常值通常是自己形成聚类，与其他样本之间几乎没有联系。与 CRLP 算法只使用A_{ij}作为边的权重相比，WSCRLP 算法能够减小划分的随机子空间和使用的基聚类方法所导致的多样性不足的影响，从而提高算法的稳定性。

3.2 加权样本

在 B 个随机子空间中执行聚类得到样本之间的共识率 A_{ij}。一方面，当$A_{ij}\approx 0$ 时，意味着样本x_i和样本x_j在 B 次聚类中几乎从未将样本x_i和样本x_j分到同一簇。另一方面，当 $A_{ij}\approx 1$ 时，意味着样本x_i和样本x_j在 B 次聚类中几乎每次都将样本x_i和样本x_j分到同一簇。样本x_i和样本x_j在同一簇的次数越多意味着它们之间的相似性越高。然而，当$A_{ij}\approx 0.5$时，意味着约一半的聚类将样本x_i和样本x_j分到相同的簇，而其他聚类将样本x_i和样本x_j划分到不同的簇。B 个聚类解决方案在如何对样本x_i和样本x_j进行聚类方面没有达成共识，这种情况具有不确定性。

为了衡量抽取的两个样本应该被放到同一簇中的置信度，我们为这些成对样本引入混乱指数，通过映射A_{ij}来捕捉这种趋势。混乱指数的定义为：

$$CF(x_i,x_j)=\begin{cases}\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{(\tilde{A}_{ij}-\mu)^2}{2\sigma^2}\right), & \text{if } i\neq j\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

其中，$\tilde{A}_{ij}=(A_{ij}+\varepsilon)/(1+2\varepsilon)$，$\varepsilon\ll 1$ 是为了避免出现$A_{ij}=0$ 的情况而添加的一个平滑项。因为我们要使 $CF(x_i,x_j)\in[0,1]$，所以这里取 $\mu=0.5$，$\sigma=1/\sqrt{2\pi}$。当$A_{ij}=0.5$ 时，混乱指数达到它的最大值，即 1。

$CF(x_i,x_j)$仅测量了x_i和x_j之间的混乱指数。为了量化每个样本的混乱指数，我们将x_i和其他样本之间的混乱指数进行如下汇总：

$$\rho_i=\sum_{j=1}^{n}\text{confusion}(x_i,x_j) \tag{11}$$

很明显，一个较大的ρ_i意味着样本x_i相对于其他样本有更大的混乱指数，这也表明x_i在共识矩阵产生的过程中更难被聚类且更靠近不同簇的边界。因此，当x_i是一个标记样本时，与小于ρ_i的其他标记样本相比，x_i应该被给予更多的重视。根据以上分析，我们把ρ_i作为标记样本的权重。

3.3 基于加权样本和共识率的标记传播算法

在构造图和加权样本之后，WSCRLP 算法通过如下的目标函数来预测未标记样本的标签。

$$Q(F)=\frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{W}_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}\boldsymbol{F}_i-\frac{1}{\sqrt{D_{jj}}}\boldsymbol{F}_j\right\|^2+\sum_{i=1}^{l}\rho_i\|\boldsymbol{F}_i-\boldsymbol{Y}_i\|^2 \tag{12}$$

在 WSCRLP 算法中，每个样本被分配了一个不同的权重ρ_i，而在式(4)中，LGC 算法为每个标记样本分配了一个由用户指定的恒定值 μ。如果样本的混乱指数ρ_i不同，说明其在聚类过程中的困难程度不同，样本x_i的ρ_i越大，意味着样本越接近不同类别的边界，则应为其分配较大的权重。这样使得当x_i更新初始标签时会导致较大的损失，其标签可以保留在标签传播过程中，从而传播到其他样本。如果对所有的标记样本分配一个恒定的值 μ，一方面，对于不同的数据集，难以确定最佳的 μ；另一方面，更新样本x_i的标签导致的损失与其他样本相同，这样x_i可能会使用不正确的标签。

类似于 LGC 算法的优化求解，我们可以得到如下迭代计算式：

$$\frac{\partial Q(\boldsymbol{F})}{\partial \boldsymbol{F}}=\boldsymbol{F}-\boldsymbol{S}\boldsymbol{F}+\boldsymbol{\Lambda}(\boldsymbol{F}-\boldsymbol{Y}) \tag{13}$$

其中，$\boldsymbol{\Lambda}\in\mathbb{R}^{n\times n}$是一个对角矩阵，$\boldsymbol{\Lambda}_{ii}=\rho_i(i\leqslant l)$和$\boldsymbol{\Lambda}_{ii}=0(i>l)$，令$\frac{\partial Q(\boldsymbol{F})}{\partial \boldsymbol{F}}=0$，可以得到 F 的解析解如下：

$$\boldsymbol{F}=(\boldsymbol{I}-\boldsymbol{S}+\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\boldsymbol{Y} \tag{14}$$

最终，样本x_i的预测标签y_i为：

$$y_i=\arg\max_{1\leqslant j\leqslant c}\boldsymbol{F}_{ij} \tag{15}$$

因为式(12)的优化求解过程类似于 LGC 算法，与 LGC 算法类似，能够很容易地证明式(12)具有收敛性。WSCRLP 算法的流程如算法 1 所示。

算法 1　WSCRLP 算法

输入：标记样本和未标记样本

$\mathbf{X}=[\mathbf{x}_1;\mathbf{x}_2;\cdots;\mathbf{x}_l;\mathbf{x}_{l+1};\cdots;\mathbf{x}_n]$

基聚类方法的数目 T；每个聚类中的簇数 k

输出：未标记样本的预测标签 y_i

1. 基于式(9)构造一个图；

2. 基于式(11)为标记样本分配权重；

3. 利用式(12)和式(14)基于 LGC 框架做迭代式标记传播；

4. 利用式(15)确定样本 x_i 最终的预测标签 y_i。

4　实验结果与分析

本节将提出的 WSCRLP 算法与 3 种现有的标记传播算法 LGC[13]，CRLP[17]，WS3C1[20] 进行实验比较。所有比较方法均基于 MATLAB2016a 实现，实验环境为 Intel Core i7-8750HCPU，8GB 内存的笔记本电脑。我们使用了 5 个 UCI 常用的数据集，所有这些数据集都可以从开放的源代码存储库 UCI 中获得。表 1 列出了它们的特征。

表 1　实验数据集

Table 1　Experimental datasets

数据集	样本数	特征数	类别数
Wine	178	14	2
Wdbc-2	699	9	2
Vertebral(V3C)	310	6	3
Bupa	345	6	2
Balance scale	675	147	9

4.1　不同数据集上的准确率

为了公平比较，实验中 LGC 算法的参数 k 从 2 开始按步长为 2 变化到 20，μ 的值[13]为 0.99。CRLP 和 WS3C1 算法中的实验参数均使用文献[17,20]提供的数据。所提出的 WSCRLP 算法中，B 设置为 60，每次聚类的簇数 k 设置为 $10C$(C 为目标数据集的类别数)，每个随机子空间中随机选取原始特征的 70%。为了避免随机性，我们在一个固定 ratio(标记样本的比例)下对每种方法在每个数据集上都进行了 10 次实验。

表 2 列出了各算法在不同数据集上的分类准确率和标准差。从表 2 的实验结果可以看出：

(1)在大部分数据集上，尤其是 Bupa，Wdbc-2，Vertebral 和 Wine 这几个低维数据集上，本文提出的 WSCRLP 算法的准确率都优于 LGC，CRLP，WS3C1，在 25 种情况下(5 个数据集和 5 个 Ratio)，WSCRLP 算法有 21 次都优于对比的其他方法。相比其他算法，WSCRLP 结合共识率和样本间的局部信息来构造图，不仅准确地捕获了数据的内在结构，而且降低了低维情况下由于划分的随机子空间多样性不足导致算法准确率较低的风险。

(2)WSCRLP 算法和 WS3C1 算法在大多数情况下的准确率和稳定性都优于另外两种方法，说明对标记样本进行加权的方法确实能提高基于图的半监督分类算法的性能。

(3)随着标记样本数量的增多，各分类算法的准确率均有所提高。当给定的标记样本较多时，样本包含的类别信息更丰富，各算法的表现均较好。

(4)Wine 数据集中，在 Ratio 为 0.05 和 0.1 的情况下，WSCRLP，WS3C1 和 CRLP 这 3 种从聚类共识的角度构造图的算法的准确率都低于 LGC 算法。原因可能是在这种情况下 Wine 数据集的标记样本过少，划分的随机子空间多样性不足等，从而导致这 3 种算法所构造的图不能完全地反映样本的全部信息。

表 2　各算法在不同 Ratio 下的分类准确率和标准差

Table 2　Classification accuracy and standard deviation of each algorithm with different Ratios

Datasets	ratio	LGC	CRLP	WS3C1	WSCRLP
Bupa	0.05	0.482±0.058	0.537±0.085	**0.539±0.038**	0.526±0.013
	0.1	0.564±0.033	0.525±0.006	0.550±0.044	**0.589±0.042**
	0.2	0.579±0.030	0.577±0.078	0.556±0.046	**0.593±0.030**
	0.3	0.580±0.028	0.598±0.021	0.581±0.043	**0.601±0.050**
	0.4	0.583±0.039	0.588±0.015	0.583±0.033	**0.649±0.015**
Wdbc-2	0.05	0.767±0.025	0.860±0.081	0.883±0.018	**0.904±0.013**
	0.1	0.864±0.020	0.892±0.054	0.893±0.014	**0.901±0.016**
	0.2	0.892±0.006	0.887±0.030	**0.903±0.009**	0.902±0.014
	0.3	0.899±0.009	0.896±0.041	0.902±0.014	**0.909±0.014**
	0.4	0.900±0.011	0.890±0.022	0.907±0.009	**0.913±0.021**
Vertebral(V3C)	0.05	0.749±0.105	0.573±0.136	0.664±0.086	**0.770±0.053**
	0.1	0.748±0.035	0.689±0.034	0.768±0.054	**0.781±0.044**
	0.2	0.750±0.031	0.672±0.029	0.782±0.032	**0.787±0.029**
	0.3	0.754±0.019	0.735±0.042	0.797±0.036	**0.802±0.031**
	0.4	0.745±0.033	0.742±0.037	0.815±0.032	**0.817±0.037**
Balance	0.05	0.543±0.105	0.706±0.112	0.541±0.124	**0.715±0.090**
	0.1	0.586±0.003	0.770±0.160	0.707±0.120	**0.816±0.046**
	0.2	0.631±0.133	0.803±0.129	0.824±0.024	**0.835±0.030**
	0.3	0.619±0.133	0.813±0.075	0.842±0.033	**0.848±0.024**
	0.4	0.762±0.162	0.822±0.050	0.846±0.017	**0.865±0.023**
Wine	0.05	0.580±**0.072**	0.433±0.059	0.479±0.131	0.549±0.103
	0.1	0.636±**0.036**	0.460±0.052	0.610±0.084	0.618±0.045
	0.2	0.648±0.028	0.585±0.068	0.701±0.052	**0.712±0.050**
	0.3	0.688±0.047	0.653±0.027	0.727±0.036	**0.755±0.050**
	0.4	0.693±0.055	0.694±0.023	0.731±0.042	**0.762±0.043**
Average accuracy		0.638	0.662	0.725	**0.757**

4.2 随机子空间数目的灵敏度分析

WSCRLP 算法是基于 CRLP 算法的改进，两种算法都需要提前确定随机子空间 B 的数目。因此我们对比在固定的 Ratio 下不同的随机子空间的数目对算法分类准确率的影响，这里我们固定 Ratio 为 20%。如图 1—图 5 所示，随着随机子空间数目的增加，两种算法的分类准确率均有所上升，但总体上 WSCRLP 算法优于 CRLP 算法；在稳定性方面，WSCRLP 也好于 CRLP 算法。考虑到随着随机子空间数量的增加，算法的运行时间也会增加，因此本文中随机子空间的数目取 60 是合理的选择。

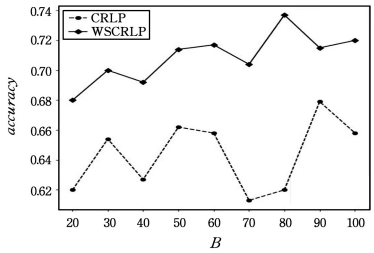

图 1 Wine 数据集下准确率与随机子空间数目的关系

Fig. 1 Relationship between accuracy rate and number of random subspaces on Wine dataset

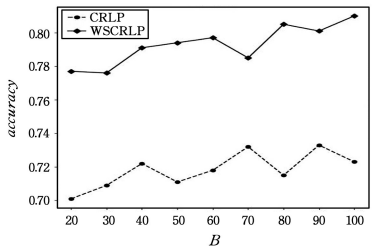

图 2 Vertebral 数据集下准确率与随机子空间数目的关系

Fig. 2 Relationship between accuracy rate and number of random subspaces on Vertebral dataset

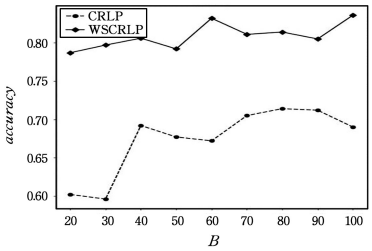

图 3 balance 数据集下准确率与随机子空间数目的关系

Fig. 3 Relationship between accuracy rate and number of random subspaces on Balance dataset

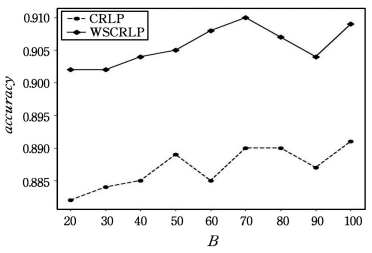

图 4 Wdbc-2 数据集下准确率与随机子空间数目的关系

Fig. 4 Relationship between accuracy rate and number of random subspaces on Wdbc-2 dataset

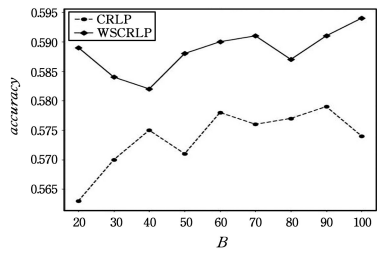

图 5 Bupa 数据集下准确率与随机子空间数目的关系

Fig. 5 Relationship between accuracy rate and number of random subspaces on Bupa dataset

4.3 运行时间分析

WSCRLP 算法的时间复杂度主要包含 3 个部分。第一部分，WSCRLP 分别重复运行 B 次 K-均值聚类，产生 B 个基聚类结果，其时间复杂度为 $O(BKndt)$，这里 K 是聚类的簇数，n 是样本的数目，d 是样本的维度，t 是迭代的次数。计算标记样本权重的时间复杂度为 $O(B\ln)$，这里 l 是标记样本的数目，因此整个第一部分的时间复杂度为 $O(Bn(kdt+l))$。第二部分，WSCRLP 在 B 个基聚类的基础上结合样本的局部信息构造一个图，其时间复杂度为 $O(B\,n^2)$。第三部分，WSCRLP 执行基于图的半监督分类并得到样本最终的预测标签，其时间复杂度为 $O(n^3)$。因此，整个 WSCRLP 的计算复杂度为 $O(Bn\times(K\times d\times n)+Bn^2+n^3)$。WSCRLP 算法的时间复杂度与 CRLP 算法相差较小，但都高于 LGC 等传统的基于图的半监督分类算法。实际上，WSCRLP 算法达到较高分类准确率和稳定性所需要执行的基聚类次数低于 CRLP 算法，因此，在大多数情况下 WSCRLP 算法的运行时间都低于 CRLP 算法。

表 3 列出了 WSCRLP 算法和其他对比方法的运行时间。表 3 中的每个数据是在 Ratio 固定为 20%的情况下 10 次独立实验结果的平均值。

表 3 LGC，CRLP，WS3C1 和 WSCRLP 的运行时间

Table 3 Runtime of LGC, CRLP, WS3C1 and WSCRLP

(单位：s)

Datasets	LGC	CRLP	WS3C1	WSCRLP
Wine	0.0058	0.6860	0.4014	0.4173
Wdbc-2	0.0490	1.0424	0.6430	0.6361
Vertebral(V3C)	0.0092	0.7978	0.4376	0.4813
Bupa	0.0101	0.7551	0.6253	0.6368
Balance Scale	0.0418	0.6862	0.5411	0.4828

从表 3 可以看出，LGC 算法的运行速度比其他的标记传播算法(CRLP，WS3C1，WSCRLP)都更快，原因是 LGC 采用了原始 KNN 图，而另外 3 种算法中构造图的方法需要对数据集进行 B 次基聚类。WSCRLP 算法和 WS3C1 算法在这 5 个数据集上的运行效率相当且都优于 CRLP 算法。这是因为 CRLP 若要达到较高分类准确率和稳定性所需要执行基聚类的次数多于 WSCRLP 和 WS3C1。

结束语 基于已有的研究，本文提出基于加权样本和共识率的标记传播算法(WSCRLP)。WSCRLP 算法首先利用标记样本在聚类过程中难以聚类的不确定程度，对标记样本分配不同的权重；其次，在基于共识率构造图的基础上结合了

样本的局部信息。在6个真实数据集上的实验结果表明，WSCRLP对低维数据的适应性更好，分类准确率更高，对随机子空间的数目具有鲁棒性，并且算法的运行效率优于CRLP算法，这说明对标记样本进行加权和构造图的方法是有效的。但是，WSCRLP算法与CRLP，WS3C1类似，没有改善运行时间较长这个缺陷。未来我们将研究如何在保证分类准确率的情况下，降低划分的随机子空间的数量，或者选择其他更有效的加权方式，来提高算法的运行效率。

参 考 文 献

[1] BLUM A，CHAWLA S. Learning from labeled and unlabeled data using graph mincuts[C]// Processdings of 18th International Conference on Machine Learning. San Francisco：Morgan Kaufman Publishers Inc，2001：19-26.

[2] LI J N，ZHU Q S. Semi-Supervised self-training method based on an optimum-path forest[J]. IEEE Access，2019，7：36388-36399.

[3] GAO Y，MA J，ALAN L，et al. Semi-Supervised sparse representation based classification for face recognition with insufficient labeled samples[J]. IEEE Transactions Image Processing，2017，26(5)：2545-2560.

[4] BELKIN M，NIYOGI P，SINDHWANI V. Manifold Regularization：A geometric framework for learning from labeled and unlabeled examples[J]. The Journal of Machine Learning Research，2006，7：2399-2434.

[5] ZHU X，GHAHRAMANI Z，LAFFERTYJ D. Semi-supervised learning using Gaussian fields and harmonic functions[C]// Processdings of the Twentieth International Conference on Machine Learning. Washington：AAAI Press，2003：912-919.

[6] TAO G H，HUA L Z，WU W，et al. Safety-aware graph-based semi-supervised learning[J]. Expert Systems with Applications，2018，107：243-254.

[7] WANG J，YAO G J，YU Z W. Semi-supervised classification by discriminative regularization[J]. Applied Soft Computing，2017，58：245-255.

[8] NIGAM K，MCCALLLUM A K，et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning，2000，39：103-134.

[9] WANG S，WU L，JIAO L，et al. Improve the performance of co-training by committee with refinement of class probability estimations[J]. Neurocomputing，2014，136：30-40.

[10] HONG Y，ZHU W. Spatial co-training for semi-supervised image classification[J]. Pattern Recognition Letters，2015，63：59-65.

[11] LI Y C，WANG Y L，BI C，et al. Revisiting transductive support vector machines with margin distribution embedding[J]. Knowledge-based Systems，2018，152：200-214.

[12] JURIC L，CECI M，KOCEV D，et al. Self-training for multi-target regression with tree ensembles[J]. Knowledge-based Systems，2017，123：41-60.

[13] ZHOU D，BOUSQUET O，LAL T N，et al. Learning with local and global consistency[C]// Proceedings of the Sixteenth Advance in Neural Information Processing Systems. Whistler：MIT Press，2003：321-328.

[14] WANG F，ZHANG C. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering，2008，20(1)：55-67.

[15] ZHAO M，CHOW T W S，ZHANG Z，et al. Automatic image annotation via compact graph based semi-supervised learning[J]. Knowledge-Based Systems，2015，76：148-165.

[16] WU F，WANG W，YANG Y，et al. Classification by semi-supervised discriminative regularization[J]. Neurcomputing，2010，73(10)：1641-1651.

[17] YU J，SB K. Consensus rate-based label propagation for semi-supervised classification[J]. Information Sciences，2018，465：265-284.

[18] DAS S，MOORE T，WONG W K，et al. End-user feature labeling：Supervised and semi-supervised approaches based on locally-weighted logistic regression[J]. Artificial Intelligence，2013，204：56-74.

[19] REN Y，DOMENICONI C，ZHANG G，et al. Weighted-object ensemble clustering[C]// IEEE International Conference on Data Mining. Dallas：IEEE Press，2013：627-636.

[20] CHEN X，YU G X，TAN Q Y，et al. Weighted samples based semi-supervised classification[J]. Applied Soft Computing，2019，79：46-58.

CHU Jie，born in 1996，postgraduate. His main research interests include data mining and machine learning.

ZHANG Zheng-jun，born in 1965，Ph.D，associate professor. His main research interests include data mining and graphic image.