Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(1), 81-89
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131010

一种基于最小最大邻域阶构图的半监督分类法

包婉莹，姚　欢

呼和浩特职业学院计算机系，内蒙古　呼和浩特

收稿日期：2023年12月13日；录用日期：2024年2月23日；发布日期：2024年2月29日

摘　要

为克服K近邻图边的对称问题及互K近邻图的连通性的不足，并且针对局部全局一致性学习(LLGC)算法的分类精度在很大程度上取决于控制参数α的设置，设置不合理可能造成分类的准确率较低，聚类的结果不准确的情况，研究提出一种半监督学习分类算法，将最小最大邻域阶构图法(KMM)结合少参数的简洁局部全局一致性学习(BB-LLGC)，即KMM-BB-LLGC算法，兼顾边的对称及整个图的连通，简化图上的目标函数，使其不受参数α的影响。用UCI数据库中的数据集进行实验，与KNN-LLGC、KNN-BB-LLGC、KMM-LLGC几种分类方法进行对比，实验表明，提出的方法能可以带来更高的分类准确率，达到较高的分类精度，算法效率更高，可以实现对样本精确、快速的分类。

关键词

图构建，局部全局一致性学习，半监督学习

A Semi Supervised Classification Algorithm Based on Minimum and Maximum Neighborhood Order Composition

Wanying Bao, Huan Yao

Department of Computer, Hohhot Vocational College, Hohhot Inner Mongolia

Received: Dec. 13th, 2023; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

Abstract

In order to overcome the problem of edge symmetry of the K-Nearest Neighbor Graph and the lack of connectivity of mutual K-Nearest Neighbor Graph, and the classification accuracy of local-global consistency learning (LLGC) algorithm largely depends on the setting of control parameters α, Unreasonable setting may result in low accuracy of a classification and inaccurate results of clus-

文章引用：包婉莹，姚欢．一种基于最小最大邻域阶构图的半监督分类法[J]．人工智能与机器人研究，2024，13(1)：81-89. DOI: 10.12677/airr.2024.131010

tering. A semi-supervised learning classification algorithm is proposed, which combines the minimum and maximum neighborhood order composition method (KMM) with a kind of barebones LLGC (BB-LLGC) algorithm with fewer parameters, that is, KMM-BB-LLGC algorithm, considering the symmetry of the edge and the connectivity of the whole graph, simplifies the objective function on the graph and make it independent of parameters α, was used in experiments with data sets in UCI database. Compared with KNN-LLGC, KNN-BB-LLGC, KMM-LLGC, experiments show that the proposed method can bring higher clustering accuracy and achieve higher classification accuracy. It is more efficient and can realize the accurate and fast classification of samples.

Keywords

Graph Construction, Local Global Consistency Learning, Semi Supervised Learning

1. 引言

机器学习的核心是学习数据，从数据中获取未知规律，利用规律对未知样本进行预测和分析[1]。半监督学习突破了传统方法只考虑一种样本类型的局限，综合利用有标签与无标签样本[2] [3] [4]，对于样本空间描述的更详细并且分类性能良好。以图表为基础的半监督分类，直接用标示、识别的方法，通过邻域图的结构将标签信息传递到无标签的数据上，具有很好的直观性和解释性[5]。常用的 K 近邻图在连接关系中存在不对称性的问题，互 K 近邻图对边的连接严格要求对称，难以保证图的连通。针对以上不足，研究使用 KMM 构图方法，连边时考虑到相对的对称性，目的是既得到容易连通的图，又能保证连接关系的可靠性[6]。

简洁局部全局一致性学习算法(BB-LLGC)是在 LLGC 算法的基础上提出的一种半监督学习算法，具有精确度高、计算速度快等优点[7]。本文将二者结合，提出 KMM-BB-LLGC 算法，算法在构图上可以达到更高的相对几何对称性又保证边紧密连接，并且控制参数少、使用简单，收敛速度快。用 UCI 数据库中的数据集进行实验，验证算法的有效性。

2. 算法相关描述

通常基于图模型的半监督学习算法大致包括三个步骤，首先选择某种构图方式构造图，接下来定义目标函数，然后进行目标函数最小化，得到最佳分类效果。

2.1. 构图

常用的 K 近邻构图本质上是一种有向图，通常的做法是简单的忽略掉边的方向，图中每个样本点强制性的与它最近的 K 个邻居连接，如 x_i 是 x_j 的 K 近邻，则两点之间就存在一条边，而不考虑 x_j 是否是 x_i 的 K 近邻，从而可能导致连接关系的不对称。K 近邻图不能反映样本之间边的对称性，造成分类的准确率较低，聚类的结果不准确。

互 K 近邻图是要求必须把每条边都以对称性的方式连接起来，即 x_i，x_j 必须互为对方的 K 近邻才会存在边的连接，这是方式往往难以保证连通性[8]。针对两种构图方法的不足，需要一个既能得到容易连通的图，又能保证连接关系的可靠性的构图方法来解决。

最小最大邻域阶构图(KMM)方法：

首先介绍几个定义：

邻域阶(Neighboring order)，设集合 X 有 n 个数据点，$p, q \in X$, p 是 q 的第 i 个邻居，q 是 p 的第 j 个邻居。从 q 到 p 的邻域阶记作：$ord(q, p) = i$；从 p 到 q 的邻域阶记作：$ord(p, q) = j$。

邻域阶和(neighboring order summation)：

$$nos\left(x_p, x_q\right) = nos\left(x_q, x_p\right) = ord\left(x_q, x_p\right) + ord\left(x_p, x_q\right) \tag{1}$$

粗略度量样本之间的相似度[6]。

邻域阶差(neighboring order difference)：

$$nod\left(x_p, x_q\right) = nod\left(x_q, x_p\right) = \left|ord\left(x_q, x_p\right) - ord\left(x_p, x_q\right)\right| \tag{2}$$

粗略度量样本之间的对称程度[6]，

$|\bullet|$ 表示取绝对值。合并可得：

$$nos\left(x_p, x_q\right) + nod\left(x_p, x_q\right) = 2 \times \max\left\{ord\left(x_q, x_p\right), ord\left(x_p, x_q\right)\right\} \tag{3}$$

基于这个表达的构图方法，即每个节点仅与 K 个最小的最大邻域阶点相连接。

将图中各顶点的邻域阶和矩阵定义为 $\boldsymbol{M_{nos}}$，邻域阶差矩阵定义为 $\boldsymbol{M_{nod}}$。则最大邻域阶矩阵 $\boldsymbol{M}$ 表示为：

$$M = M_{nos} + M_{nod} \tag{4}$$

其中，“+”表示两个矩阵的元素对应相加，表达式忽略了常量因子 2。为使 $\boldsymbol{M}$ 中的元素尽可能唯一，将其修正为：

$$\widetilde{M} = M_{nos} + M_{nod} + W / \max(W) \tag{5}$$

$\boldsymbol{max}\ (\boldsymbol{W})$是距离矩阵 $\boldsymbol{W}$ 中的最大元素，对于计算得到的矩阵 $\boldsymbol{M}$，解如下的优化问题来构造 KMM 图的邻域：$P = \arg\min \sum_{p \in B} P_{ij} \widetilde{M}_{ij}$

$$s.t. \sum_j P_{ij} = K$$

$$P_{ii} = 0, \forall_{i,j} \in 1, 2, \cdots, n \tag{6}$$

$\boldsymbol{P}$ 是只取 0 和 1 的二值矩阵，$P_{ij} = 0$ 即 i 和 j 之间不连边，值为 1 则有一条边相连。K 近邻图采用距离矩阵，而这里采用最大邻域阶矩阵。当邻域选择确定之后，利用式(7)采用非负局部线性重构来给产生的边加权。

$$w* = \arg\min_w \left\| x_i - \sum_j w_j x_j \right\|_2$$

$$s.t. \sum_j w_j = 1, w_j \geq 0 \tag{7}$$

x_i 表示被重构的点，$x_j (0 < j \leq k)$表示参与重构的数据点，$\boldsymbol{W}$ 表示长度为 $\boldsymbol{K}$ 的重构系数向量。从相似度度量的角度来讲，W_j 表示的是从 x_i 到 x_j 的有向边的权值[6]。

2.2. 半监督学习算法

最具代表性的是 Zhou 等人在 2004 年提出了一种基于局部与全局一致性的算法(LLGC)。该算法将类别标签通过样本的近邻传递到整个图中，将优化目标项的权值取值范围约束到限定范围内，使算法允许出现一定的错误标注，但是算法对控制参数 α 的设置比较敏感，并且在标签传递过程中使已标记样本的

标签随着循环传递而改变，导致传递源改变[9]，以至于分类的准确性无法保证。

LLGC 算法的思想：

假设 $X=\{x_{i'}\}_{xi'=1}^{n}$ 数据集中有 n 个样本，$C=\{c_j\}_{j=1}^{c}$ 标签集中有 c 类，c_j 表示某一样本的类别。将样本数据集合 X 分为已标记样本集 $X_L=\{(x_1,y_1)(x_2,y_2)\cdots(x_l,y_l)\}$ 和未标记样本集 $X_U=\{x_{l+1},\cdots,x_n\}$，其中 $y_i\in$C，$Y_L=\{y_l\}_{i=1}^{l}$，学习的目标是根据 X 和 Y_L 预测未标记样本集 X_U 的类别标签 Y_U。

Step1：利用样本集构造图并计算图中样本的相似度矩阵 $\boldsymbol{W}$，β 是带宽，控制径向作用范围：

$$w_{i',j'}=e^{\frac{-\|x_{i'}-x_j\|^2}{\beta^2}} \tag{8}$$

Step2：构造函数：

函数要满足将有标签样本准确分类，并且无标签样本可以在图上平滑分布。图的半监督学习的框架可以用损失项和光滑正则项两项相加表示[10]。对于已标记样本，令 $El(f)$为损失函数，表示预测标签与真实标签间的误差，令 $Es(f)$为目标函数的调整项，通常采用引入正则项的方法来确保标签分布的平滑性。

Step3：最小化目标函数：

$$\text{即}\min_f E(f)=E_l(f)+E_s(f) \tag{9}$$

LLGC 算法的目标函数如下所示：

$$\min_f E(f)=\frac{1}{2}\sum_{i',j'=1}^{n}\omega_{i'j'}\left(f_{i'}-f_{j'}\right|_l)^2+\mu\sum\nolimits_{i=1}^{l}\left\|(f_i-y_i)\right\|^2 \tag{10}$$

通过推导目标函可以写为：

$$F^*=(1-\alpha)(I-\alpha S)^{-1}\tilde{Y} \tag{11}$$

其中，$\boldsymbol{S}$ 是正则化的图拉普拉斯矩阵，$S=D^{-1/2}WD^{-1/2}$，$W=\{\omega_{i'j'}\}_{n\times n}$，$\boldsymbol{D}$ 是一个对角矩阵，对角线元素为 $D_{i'i'}=\sum_{j'=1}^{n}\omega_{i'j'}$，$\tilde{Y}$ 是一个 $n\times c$ 的矩阵，$\tilde{Y}=\begin{cases}1, x_i\in C, 1\le i\le l\\ 0, otherwise\end{cases}$，参数 $\alpha\in(0,1)$，需根据情况事先确定。然后，通过 $f_{i'}=\arg\max_{j\le c}F_{i'j}$（$F_{i'j}$ 表示预测样本 i' 属于类别 j 的概率，$F=\{F_{i'j}\}_{n\times c}$）来预测未标记样本标签。

公式(11)中，α 对 LLGC 算法影响较大，决定相对较多的信息是来自于近邻还是初始标记样本，所以算法性能的好坏很大程度上取决于参数的选择。

简洁全局一致性学习算法(BB-LLGC)：修改了目标函数以避免参数 α 的影响，将图中样本的标签根据其与相邻样本的相似度大小传递给其近邻，标签以此方式反复传递；同时不改变已知的标记样本标签，直至全局稳定。

BB-LLGC 算法定义了新目标函数：

$$\min\frac{1}{2}\sum_{x_j'\in N(x_i')}w_{i'j'}\left(f_{i'}-f_{j'}\right)^2 \tag{12}$$

其中 $N(x_i')$表示未标记样本 x_i'的 $\boldsymbol{K}$ 个近邻组成的数据集合。每次迭代，已标记样本的标签不变，保证标记源头准确，以此为起点，未标记样本标签逐次传播到其近邻的未标记样本，直到所有样本类别不变为止。

定义一个 $n\times c$ 的矩阵 $\boldsymbol{F}$ 表示每个样本的标注概率，将其分为 $\boldsymbol{F_L(0)}$与 $\boldsymbol{F_U(0)}$，分别对应 $\boldsymbol{X_L}$ 和 $\boldsymbol{X_U}$。其中，$\boldsymbol{F_L(0)}$为 $l\times c$ 的矩阵，$\boldsymbol{F_U(0)}$为$(n-l)\times c$ 的矩阵：

$$F_{i'j}=\begin{cases}1, x_{i'}\in c_j\\ 0, x_{i'}\notin c_j\end{cases} \tag{13}$$

x_i'属于 $\boldsymbol{X_L}$，初始设定 $\boldsymbol{F_U(0)}$每一行元素的值为 0。

根据公式(14)预测每个未标记样本的类别标签：

$$f_{i'}=\arg\max_{j\le c} F_{i'j} \tag{14}$$

3. 算法 KMM-BB-LLGC

本文提出将最小最大邻域阶构图法(KMM)结合少参数的简洁局部全局一致性学习(BB-LLGC)，即 KMM-BB-LLGC 算法。

这种半监督学习算法既兼顾边的对称及整个图的连通又简化图上的目标函数，使其不受参数 α 的影响。

KMM-BB-LLGC 算法的核心步骤：

Step1：根据公式(6)的优化问题，计算每个样本的邻域并构造 KMM 图；

Step2：构造相似度近邻矩阵 $\boldsymbol{W}$，近邻点按照公式(8)来计算，非近邻点设为 0；

Step3：计算图的正则化拉普拉斯矩阵 $\boldsymbol{S}$；

Step4：根据 $F(t+1)=SF(t)$ 更新样本点标签概率，

Step 4.1: $t=0$, $F(0)=[F_L(0); F_U(0)]$;

Step 4.2: $t=t+1$, $F(t+1)=SF(t)$;

Step 4.3：限制

$$F_L(t)=F_L(0); \tag{15}$$

Step 5：重复 Step 4.2 和 Step 4.3，直到 $\boldsymbol{F}$ 收敛到一个确定的值 $\boldsymbol{F^*_U}$ 为止。使得标注矩阵 $\boldsymbol{F}$ 收敛至确定的

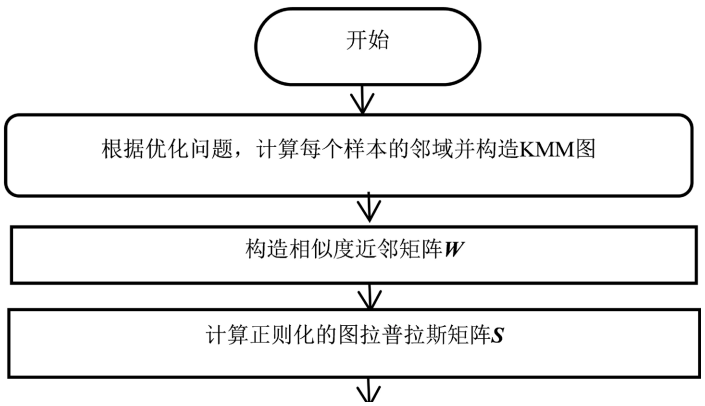
$$F^*_U=\lim_{t\to\infty}F_U(t+1)=(I-S_{UU})^{-1}S_{UL}F_L(0), \tag{16}$$

其中 $\boldsymbol{S_{UU}}$ 和 $\mathrm{S}_{\boldsymbol{UL}}$ 是正则化拉普拉斯矩阵 $\boldsymbol{S}$ 中的两个部分：

$$S=\begin{bmatrix}S_{LL} & S_{LU}\\ S_{UL} & S_{UL}\end{bmatrix}$$

Step6：利用公式(14)预测每个未标记样本的类别标签。

算法流程如图 1 所示：



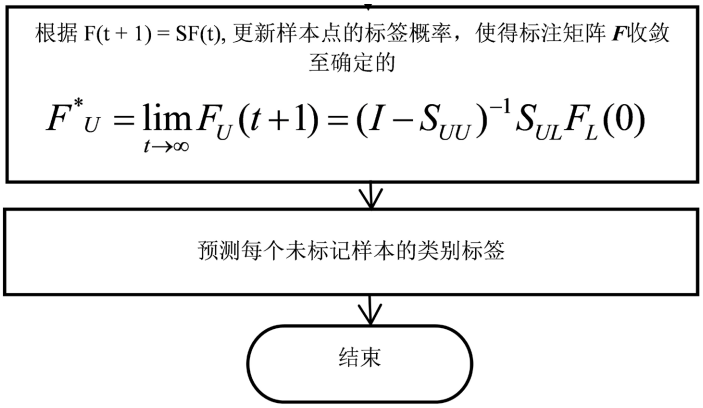


Figure 1. KMM-BB-LLGC algorithm flow chart
图 1. KMM-BB-LLGC 算法流程图

4. 数据与实验

4.1. 数据

为验证基于最小最大邻域阶构图的简洁局部全局一致性学习方法 KMM-BB-LLGC 在数据分析上的有效性，本文采用如下数据集进行测试分析：

实验使用 UCI 数据集，数据已做好预处理及空间向量化[11]。选取其中 4 个数据集，如表 1 所示。

Table 1. Dataset information
表 1. 数据集信息

数据集	样本总数	特征维数	类别
Iirs	150	4	3
Ionosphere	351	34	2
VoIirswel	990	10	11
Imagesegmentation	2310	19	7

4.2. 实验

实验用计算机的硬件配置如下：实验采用操作系统为 Windows10，内存 1 GB，采用 Matlab2016b 软件编程。

在参数选择中，将每个算法的共有参数设为一致，通过反复实验，试凑对 LLGC 中的 α 和 KMM 中的 K 进行选择，具体对于每一个数据集参数取值情况如表 2 所示。

Table 2. Parameter setting
表 2. 参数设置情况

数据集	KNN	KMM		LLGC		BB-LLGC
	K	K	β	α	β	β
Iris	1	4	180.5	0.6	180.5	180.5
Ionosphere	1	15	312.5	0.4	312.5	312.5
VoIirswel	1	25	0.5	0.99	0.5	0.5
Image	1	20	512	0.99	512	512

对于每个实验样本集在选择已标签样本时，已标签的样本数分别取 10%，20%，30%，40%，50%，

60%独立重复上述样本选择过程 20 次，作为随机实验的输入样本数据集。将这 20 次独立重复实验结果的平均值作为评价算法效果的最终依据。实验对测试数据的采用 10 折交叉验证法进行测试，主要采用分类精度(ACC)和迭代时间作为评价标准。

4.3. 实验结果

在实验数据集上测试了不同分类方法及不同标记比例上的分类准确率，结果如图 2~5 所示；Iirs 数据集及 VoIirswel 数据集在不同方法中迭代时间的对比如表 3~4 所示。

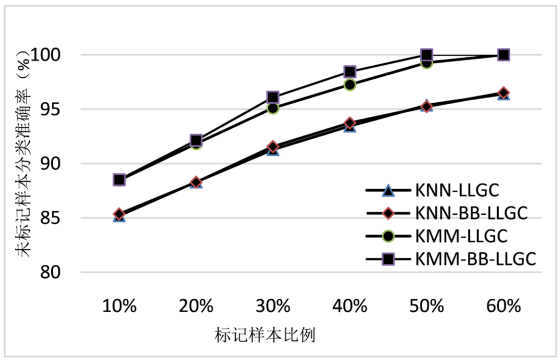


Figure 2. Iris dataset
图 2. Iris 数据集

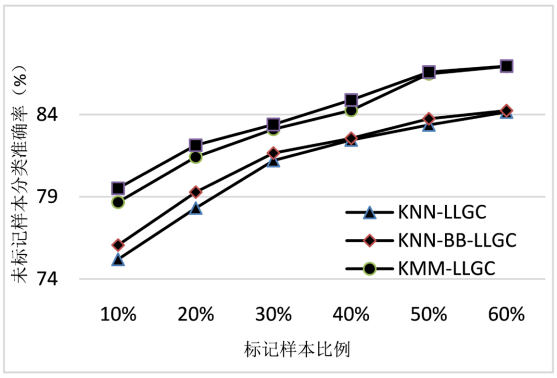


Figure 3. Ionosphere dataset
图 3. Ionosphere 数据集

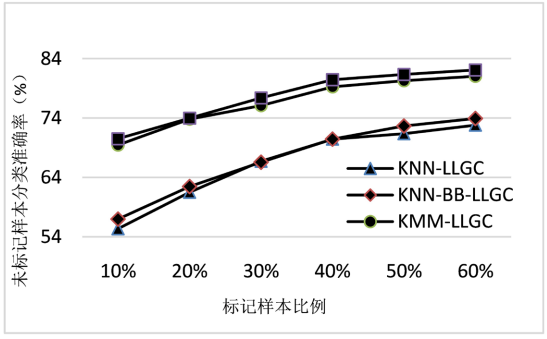


Figure 4. VoIirswel dataset
图 4. VoIirswel 数据集

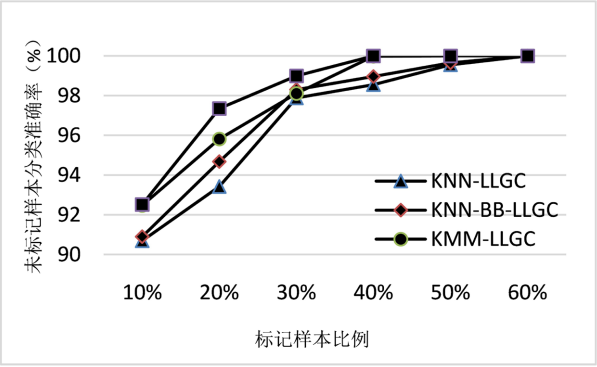


Figure 5. Image segmentation dataset
图 5. Image segmentation 数据集

Table 3. IriS data set
表 3. Iris 数据集

标记比例 / 算法 / 迭代时间	10%	20%	30%	40%	50%	60%
KNN-LLGC	2.3622	2.3175	2.2936	2.1462	1.9732	1.9621
KNN-BB-LLGC	1.8335	1.7856	1.7823	1.6746	1.6731	1.6714
KMM-LLGC	3.8466	3.6308	3.3543	3.3109	3.0874	3.0311
KMM-BB-LLGC	2.4952	2.3814	2.3309	2.2331	2.1423	1.9868

Table 4. Volirswel datase
表 4. VoIirswel 数据集

标记比例 / 算法 / 迭代时间	10%	20%	30%	40%	50%	60%
KNN-LLGC	4.6835	4.2875	4.0936	4.0521	3.824	3.6650
KNN-BB-LLGC	3.8335	3.3824	2.9863	2.6746	2.5798	2.4684
KMM-LLGC	6.3378	6.2384	6.0914	5.8769	5.6865	5.5302
KMM-BB-LLGC	4.7924	4.3906	4.3418	4.2432	4.1024	3.7543

5. 结果分析

5.1. 从分类准确率的角度看

(1) 整体上对于每种分类算法，当标记数据比例 ≤ 40%时，随着有标记样本比例的增加未标记样本的分类准确率有明显的上升，之后有标签数据增加，分类准确率的上升趋势趋于平稳。

(2) 当标记样本数较少时，KNN-LLGC 与 KMM-LLGC 算法及 KNN-BB-LLGC 与 KMM-BB-LLGC 相比准确率差距较大；当标记的样本数量增大时，KNN-LLGC 与 KMM-LLGC 算法及 KNN-BB-LLGC 与 KMM-BB-LLGC 准确率之间的差异逐渐减小。基于 KMM 构图的方法相较于 KNN 构图的方法分类准确率明显增加，尤其在只有很少的有标记数据时显示出更强的优越性能。

(3) 无论是在数据维数达到 34 的 Ionosphere 数据集还是样本总数达到 2310 的 Image segmentation 数据集上，KMM-BB-LLGC 算法均表现出准确率最高，说明结合的改进算法 KMM-BB-LLGC 的分类性能较理想，分类精度较好。

5.2. 从收敛时间上的角度看

(1) 实验中，在同一数据集、相同半监督学习算法的条件下，不同构图方式的收敛时间不同。KNN-LLGC < KMM-LLGC，KNN-BB-LLGC < KMM-BB-LLGC。原因是在构建 KMM 邻域图的过程中需要构建三个矩阵 $\boldsymbol{M}_{nos}$，$\boldsymbol{M}_{nod}$ 及 ω，但是基于 KMM 构图的方法的分类准确率更高。

(2) 同一数据集，基于同种构图方法的条件下，KNN-LLGC > KNN-BB-LLGC，KMM-LLGC > KMM-BBB-LLGC，原因是 BB-LLGC 使已标记样本的标签在标签传递过程中确保不变，因此其分类准确率高于传统 LLGC 且收敛速度快；

(3) 四种方法中，KNN-BB-LLGC 算法的迭代时间最短，是由于算法仅需要构造一个距离矩阵并且参数少，已标记样本的标签在传递过程中不变；KMM-BB-LLGC 算法与 KNN-LLGC 算法的迭代时间相差很小，但却表现出更高的准确率，说明结合后的算法可以得到更加理想和稳定的分类效果。

6. 结论

研究使用 KMM 构图方法，避免了原有 K 近邻图法与互 K 近邻图法各自的不足，通过解决优化问题样本点的邻域，将其应用于 BB-LLGC 算法中，为避免受参数 α 的影响，建立简化的目标函数并遵循简单的标签传递过程，可以带来更高的聚类准确性，具有更好的分类效果[12] [13]。虽然基于 KMM 构图牺牲了一点收敛速度，但是对样本的分布特征描述的更加全面，并且 BB-LLGC 比传统 LLGC 收敛更快，准确率更高，将二者结合的 KMM-BB-LLGC 算法是一种有效的半监督分类算法。下一步工作准备将算法应用于高光谱图像分类中，验证其适应性，利用高光谱图像信息和光谱信息进行分类，改善高光谱图像的分类效果[14] [15]。

参考文献

[1] 韩嵩, 韩秋弘. 半监督学习研究的述评[J]. 计算机工程与应用, 2020, 56(6): 19-27.

[2] 高翠芳, 吴小俊, 张松顺. 改进的半监督模糊聚类算法[J]. 控制与决策, 2010, 25(1): 115-120.

[3] 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述[J]. 计算机科学, 2017, 44(1): 7-13.

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

[5] 韩灵珊. 基于两种不同构图方法的半监督分类算法研究[D]: [硕士学位论文]. 重庆: 重庆师范大学, 2016.

[6] 张钧伟, 齐鸣鸣, 许淑华. 最小最大邻域阶构图方法[J]. 计算机工程与应用, 2012, 48(12): 202-205.

[7] 祝磊, 曹凯敏, 游晓璐, 等. 基于聚类分析和半监督学习的蛋白质质谱数据分类[J]. 航天医学与医学工程, 2014, 27(5): 367-372.

[8] Luxburg, U. (2007) A Tutorial on Spectral Clustering. *Statistics and Computing*, **17**, 395-416.

[9] 王雪松, 张晓丽, 程玉虎. 一种简洁局部全局一致性学习[J]. 控制与决策, 2011, 26(11): 1727-1734.

[10] 王君言. 基于稀疏图的小样本高光谱图像半监督分类算法研究[D]: [硕士学位论文]. 银川: 北方民族大学, 2017.

[11] 贺松林, 张晖. 基于 K-means 和 LabelPropagation 的半监督网页分类[J]. 软件导刊, 2011, 10(2): 49-51.

[12] 蔡先发. 基于图的半监督算法及其应用研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2013.

[13] 宗鸣, 龚永红, 文国秋, 等. 基于稀疏学习的 KNN 分类[J]. 广西师范大学学报(自然科学版), 2016, 34(3): 39-45.

[14] 刘丹. 基于稀疏表示的高光谱图像分类算法研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2016.

[15] 罗甫林. 高光谱图像稀疏流形学习方法研究[D]: [硕士学位论文]. 重庆: 重庆大学, 2016.