计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于标签进行度量学习的图半监督学习算法
作者：吕亚丽，苗钧重，胡玮昕
收稿日期：2020-06-12
网络首发日期：2020-10-23
引用格式：吕亚丽，苗钧重，胡玮昕．基于标签进行度量学习的图半监督学习算法[J/OL]．计算机应用．
https://kns.cnki.net/kcms/detail/51.1307.TP.20201022.1005.002.html

Journal of Computer Applications　ISSN 1001-9081　2020-10-18
计算机应用　CODEN JYIIDU　http://www.joca.cn

文章编号:1001-9081(****)**-0000-00　DOI:10.11772/j.issn.1001-9081.2020060893

基于标签进行度量学习的图半监督学习算法

吕亚丽[1,2*]，苗钧重[1]，胡玮昕[1]

(1.山西财经大学 信息学院，太原 030006)

(2.计算智能与中文信息处理教育部重点实验室(山西大学)，太原 030006)

(*通信作者电子邮箱 sxlvyali@126.com)

摘 要: 大多基于图的半监督学习方法，在样本间相似性度量时没有用到已有的和标签传播过程中得到的标签信息。同时，度量方式相对固定，不能有效度量分布结构复杂多样的数据样本间的相似性。针对上述问题，提出了基于标签进行度量学习的图的半监督学习算法。首先，给定样本间相似性度量方式，构建相似度矩阵。然后，基于相似度矩阵进行标签传播，筛选出 k 个低熵数据作为新确定的标签信息。最后，充分利用所有标签信息更新相似性度量方式，重复迭代优化，直至学出所有标签信息。所提算法不仅利用标签信息改进了样本间相似性度量方式，而且充分利用中间结果以降低半监督学习对标签数据的需求量。在 6 个真实数据集上的实验结果表明，该算法在超过 95%的情况下比三种传统的基于图的半监督学习算法取得了更高的分类准确率。

关键词: 机器学习；图半监督学习；度量学习；标签传播；相似度矩阵

中图分类号:TP181　　**文献标志码: A**

Semi-supervised learning algorithm of graph based on label-based metric learning

LYU Yali[1,2*], MIAO Junzhong[1], HU Weixin[1]

(1 School of information, Shanxi University of Finance and Economics, Taiyuan Shanxi 030006, China;

2 Key Laboratory of Computational and Chinese Information Processing of Ministry of Education (Shanxi University), Taiyuan Shanxi 030006, China)

Abstract: Most graph-based semi-supervised learning methods do not use the known label information and the obtained label information from the label propagation process when measuring the similarity between samples. At the same time, the measurement methods are relatively fixed, which cannot effectively measure the similarity between samples with complex and diversified distribution structures. To avoid these disadvantages, a semi-supervised learning algorithm of graph based on label-based metric learning was proposed. Firstly, the similarity measure was given and then the similarity matrix was constructed. Secondly, labels are propagated based on the similarity matrix and k samples with low entropy are selected as the new obtained labels. Next, the similarity measures are updated based on all label information, this process is repeated until all labels are learned. The algorithm proposed in this paper, not only uses label information to improve the measurement between samples, but also, makes full use of intermediate results to reduce the demand for labeled data in semi-supervised learning. The experimental results on the six real data set show that, compared with three traditional graph-based semi-supervised learning methods, the proposed algorithm achieves the higher classification accuracy in more than 95% of cases.

Keywords: machine learning; based-graph semi-supervised learning; metric learning; label propagation; similarity matrix

0 引言

随着机器学习算法的迅猛发展，其应用领域也越来越广泛，需要处理的数据也越来越复杂。由于标签数据有标准或最优的输出，所以在算法中可以很好地构建目标函数用于求解模型参数。然而，大数据环境下，标签信息有限，许多无标签数据唾手可得，但要想获得它们的标签信息却需要付出高昂的人工成本[1]。因此，如何同时利用少量标签数据与大量无标签数据提高模型的分类性能这一问题显得越来越重要。尤其是在学习过程中，如何降低模型学习对标签样本的需求量，同时又可以提高学习性能，成为了一个非常重要的研究问题[2-3]。近些年，涌现出大量关于半监督学习的研究工作并取得了较好效果[4]，其中包括较为热点的图半监督学习

收稿日期: 2020-06-12;　修回日期:2020-08-20;　录用日期: 2020-09-11。

基金项目:山西省自然科学基金项目(201801D121115)；山西省回国留学人员科研资助项目(2020-095)。

作者简介: 吕亚丽(1975—)，女，山西临汾人，副教授，博士研究生，CCF 会员，主要研究方向：数据挖掘、机器学习、概率推理；苗钧重(1993—)，男，山西晋中人，硕士研究生，主要研究方向：数据挖掘、机器学习；胡玮昕(1996—)，女，山西晋中人，硕士研究生，主要研究方向：数据挖掘、机器学习。

算法，其任务可以转换为一个凸优化问题，从而可以求得全局最优解[5]。

半监督学习算法大致可以分为直推式学习和归纳学习两类[4]。直推式学习是指将标签数据作为训练集用于预测无标签样本的类别的算法。归纳学习是指同时利用标签样本和无标签样本学习出一个分类器，再将其用于分类无标签样本的算法。

基于图的半监督学习属于归纳学习的一种，是基于局部假设和全局假设来进行的，局部假设为邻近的样本应该具有相同的类标签[6]，全局假设是在同一结构中的样本应该具有相同的类标签。基于图的半监督学习可以总结为将数据中少数的标签进行传播，利用大量的无标签数据进行样本空间的结构识别[7]。

大多数基于图的半监督学习算法包含两个步骤：一是通过计算样本间的距离或相似性度量来构建相似度矩阵。每个样本可以看作是图中的一个节点，样本间的相似度可以看作是图中节点间连边的权重[8]。权重越大表示这两个样本具有相同标签的概率就越大。二是根据得出的相似度矩阵来预测无标签样本的所属类别。

目前，第一步已有工作在构建相似度矩阵方面的算法有：k-近邻(k-Nearest Neighbor, KNN)、ε近邻(ε-neighbor)[9]、热核方法(heat kernel)[10]、局部线性标识(local linear representation)[8,11]、低秩表示(Subspace Segmentation by Low Rank Representation, S^{2}LRR)[12]以及稀疏表示(sparse representation)[13-15]等。其中，KNN方法在构建相似度矩阵时，选择距离每个样本点最近的k个样本点作为其邻居，据此构建样本间相似度矩阵。在该方法中，通常测量的是样本间的欧氏距离，超参数k的选择非常重要。k过大过小均不能正确反映出样本间的相似性；在ε-neighbor方法中，通过设定阈值来筛选对应样本的邻居来构建相似度矩阵，该方法中阈值的选择尤为重要；而heat kernel[10]、local linear representation[8,11]、low rank representation[12]以及sparse representation[13-15]，四种方法基本原理是通过不同的核方法来度量样本间相似性，对于不同的核方法，均涉及超参数，这些参数决定了模型复杂度从而决定样本间的相似性度量是否合适。此外，文献[16]采用非负矩阵分解与概念分解提出了一种基于数据表示的图半监督学习算法。然而，上述算法基本采用了欧氏距离作为样本间相似性度量的核心方法，且其度量方式相对固定，这样对于不同数据采用不同的度量灵活性和适应性相对较差。

第二步预测标签方面的标签传播算法有：高斯场与调和函数(Gaussian Fields and Harmonic Function, GFHF)[17]、局部和全局一致性方法(the Local and Global Consistency, LGC)[6]以及特殊标签传播算法(Special Label Propagation, SLP)[18]等。其中，GFHF方法是利用高斯核来度量样本间的相似度，使用调和函数来进行标签传播，在调和函数中，对于标签数据，函数值为其标签值；对于无标签数据，函数值为标签其类别归属的权重平均值。这种方法的优点在于优化问题具有良好的数学性质，且具有闭式解。LGC方法基于流形正则化思想，通过构造一个相对平滑的分类目标函数，来实现标签传播过程中尽可能使得处于同一类簇结构中的样本具有相同的标签。此外，文献[19]提出了一种基于流形的可解释性的图半监督学习算法，文献[20]从图形信号处理的角度来考虑了标签传播，提出了一种广义标签传播算法。而这些标签传播算法的标签传播过程与第一阶段的样本相似性度量过程是分离的，且对于中间结果的利用不够充分。

基于上述问题，本文提出了基于标签进行度量学习的图半监督学习算法。具体地，将在构建相似度矩阵与标签传播过程中均充分利用宝贵的标签信息。同时，利用在标签传播过程中的中间结果来不断更新相似性度量方式，通过不断迭代优化调整相似性度量与标签传播，进而提升标签预测性能，提高分类准确率。本文提出的算法不仅使得样本间相似性度量更加准确，而且充分利用中间结果大大降低了对标签数据的需求量。最后，通过实验验证了本文所提算法在标签数据占比极小的情况也可以取得较高的分类准确率。

1 预备知识

本章主要介绍基于图的半监督学习算法以及度量学习的相关知识。

1.1 基于图的半监督学习方法

在基于图的半监督学习中，第一步计算样本间相似度时，已有工作经常采用高斯核函数来进行[17-18,21]，如样本$\boldsymbol{x}_i$与$\boldsymbol{x}_j$的相似度被定义为：

$$s_{ij}=\mathrm{e}^{-\frac{\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2}{2\sigma^2}} \tag{1}$$

这种方法将样本间的欧氏距离视作样本间的相似度，其核心还是欧氏距离，这种度量方式并未用到宝贵的标签信息，这样不仅导致了标签信息的浪费，而且使得相似性度量不精确。

基于图的半监督学习算法在进行标签传播时采用的思想是：相似度大的样本应该具有相同的标签。通常会先构建标签向量，如某一任务中共有K个类别的数据，其中某一样本属于第m类，那么该样本所对应的标签向量为(0,0,...,1,...,0)，即向量中除第m个元素为1外，其余元素均为0，同时还规定标签向量中所有元素之和为1。可以把标签向量中每一个位置上的元素看成是对应样本属于某一类的概率，当两个样本间的相似度s_{ij}较大时，样本$\boldsymbol{x}_i$与$\boldsymbol{x}_j$所对应的标签向量应尽可能的相似，即其欧氏距离$\left\|\boldsymbol{f}_i-\boldsymbol{f}_j\right\|$要尽可能小。也

就是将每一个样本点看作图中的一个节点，样本间的相似度看作是节点间连边的权重。然后，找到最合适的切割方式把整个图分成 K 个子图，使得各个子图所包含的边的权重之和最大，同时使得被切割掉的边的权重之和最小。

1.2 度量学习

许多机器学习算法很大程度上依赖于样本间的度量方式，一个合适的度量方式不仅可以使得学习的结果更加准确，而且可以使得学习过程更加便捷。大多数算法采用了固定的度量方式，常见的度量方式有欧氏距离、曼哈顿距离、推土机距离、切比雪夫距离等。还有一些算法是首先在原始样本空间上进行特征选择，然后在特征空间上进行固定形式的距离度量。

那么，如何根据实际问题或面对不同的数据进行不同方式的度量？面对这一问题，研究者们提出了很多度量学习方法：大边界最近邻算法[22]、基于密度加权的大边界最近邻分类算法[23]、基于余弦距离的度量学习算法等。

许多度量学习方法中采用了马氏距离作为度量的函数形式，根据样本相似性计算具体的度量参数值。其度量公式被定义为：

$$d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_{\boldsymbol{A}}=\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})} \quad (2)$$

其中，矩阵 $\boldsymbol{A}$ 满足半正定。当 $\boldsymbol{A}$ 为单位矩阵时，该距离就变成了欧氏距离。

接下来根据样本间的相似度来计算 $\boldsymbol{A}$ 矩阵[24]。设 M 为相似样本对集合，D 为不相似样本对集合。按相似样本之间的距离应尽可能小的原则，构建如下优化问题：

$$\begin{aligned}&\min_{\boldsymbol{A}}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\\&\text{s.t.}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}\geq 1\\&\qquad A\succeq 0\end{aligned} \quad (3)$$

其中的约束条件是为了避免所有的数据都集中到一个点这种极端情况的出现。该问题为一个凸优化问题，可以求得全局最优解。

若采用拉格朗日对偶性进行求解，其时间复杂度为 $O(n^6)$、空间复杂度为 $O(n^2)$。上述问题也可被转化为如下等价问题来求解[24]：

$$\begin{aligned}&\max_{\boldsymbol{A}}\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i,\boldsymbol{x}_j\|_{\boldsymbol{A}}\\&s.t\ \ f(\boldsymbol{A})=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M}\|\boldsymbol{x}_i,\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\leq 1\\&\qquad\boldsymbol{A}\geq 0\end{aligned} \quad (4)$$

此时，可以使用梯度下降法对目标函数进行求解。用迭代优化的方式来使得 $\boldsymbol{A}$ 满足约束条件。

具体的距离度量学习（Distance Metric Learning，DML）求解思路如算法 1 所示。

算法 1: DML。

1) begin

2) While $g(\boldsymbol{A})$不收敛 do

3) While $\boldsymbol{A}$ 不收敛 do

4) $\boldsymbol{A}:=\arg\min_{\boldsymbol{A}'}\left\{\|\boldsymbol{A}'-\boldsymbol{A}\|_F:\boldsymbol{A}'\in C_1\right\}$；

5) $\boldsymbol{A}:=\arg\min_{\boldsymbol{A}'}\left\{\|\boldsymbol{A}'-\boldsymbol{A}\|_F:\boldsymbol{A}'\in C_2\right\}$；

6) end

7) $\boldsymbol{A}:=\boldsymbol{A}+\alpha(\nabla_{\boldsymbol{A}}g(\boldsymbol{A}))_{\perp\nabla_{\boldsymbol{A}}f}$；

8) end

9) return $\boldsymbol{A}$；

10) end

其中，$C_1=\{\boldsymbol{A}:\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\leq 1\}$，$\boldsymbol{A}$ 为距离度量矩阵，$C_2=\{\boldsymbol{A}:A\succeq 0\}$。$g(\boldsymbol{A})=-\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i,\boldsymbol{x}_j\|_{\boldsymbol{A}}$ 为优化问题的目标函数。算法 1 采用的是梯度下降法。算法 1 中的第 7）行为梯度下降迭代公式，α 为学习率，$\nabla_{\boldsymbol{A}}g(\boldsymbol{A})$ 为目标函数的梯度。第 4）、5）行的两步是为了使得迭代后的矩阵满足约束条件 $f(\boldsymbol{A})=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M}\|\boldsymbol{x}_i,\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\leq 1$ 和 $\boldsymbol{A}\geq 0$。由于式（4）的优化问题与式（3）的等价，同样也是一个凸优化问题，这样就保证了算法 1 可以求得全局最优解且保证了算法的收敛性，详细理论分析见文献[24]。

2 基于 DML 的图半监督学习方法

本文主要利用标签信息进行相似性的度量学习，进而提出基于标签进行度量学习的图半监督学习算法。因此，本章首先给定相似性度量方式，进而构建相似度矩阵；其次，基于该相似性矩阵进行标签传播；接着，基于信息熵确定前 k 个相对确定的样本标签；然后，再加入新学出的标签信息进行相似性度量学习；最后，构建相似性矩阵和标签传播等，如此迭代，直至学出所有标签信息。

2.1 相似度矩阵的构建

给定一个包含 n 个样本的数据集 $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\ldots,\boldsymbol{x}_n\}$，具体包含 l 个具有标签的数据和 $u=n-l$ 个无标签数据。给定一个标签集

$\mathcal{Y}=\{1,2,\ldots,C\}$ 表示有 C 个类。其中 $\boldsymbol{x}_i \in \mathrm{R}^d$，$i=1,2,\ldots,n$，这里 d 表示数据的维度。

为了构建相似度矩阵 $S=\{s_{ij}\}_{n\times n}$，定义样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的相似度为：

$$s_{ij}=\frac{\mathrm{e}^{-dis_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}}{\sum_{k=1}^{n}\mathrm{e}^{-dis_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_k)}} \tag{5}$$

其中，$dis_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_j}$。当 $\boldsymbol{A}$ 为单位矩阵时，$dis_{\boldsymbol{I}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为欧式距离。由于本文考虑的是一对样本之间的相似性，所以规定 $s_{ii}=0$。这时可以理解为：要找到一种映射 $f:X\mapsto Y$，其中 X 为输入空间，Y 为特征空间。满足 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}=\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_{\boldsymbol{I}}$，即输入空间以 $\boldsymbol{A}$ 为矩阵的马氏距离等于特征空间的欧氏距离。

从概率角度看，s_{ij} 可看作是 $\boldsymbol{x}_i$ 选择 $\boldsymbol{x}_j$ 作为自己邻居的概率，若记 $p_{j|i}=s_{ij}$，则 $p_{j|i}$ 为以 $\boldsymbol{x}_i$ 为中心，单位矩阵 $\boldsymbol{I}$ 为协方差矩阵的高斯分布的概率密度[25]。

2.2 基于相似度矩阵的标签传播

当确定了样本间的相似度矩阵后，接下来根据相似度矩阵进行标签传播，给每个样本 $\boldsymbol{x}_i$ 一个标签向量 $\boldsymbol{f}_i$，若 $i\le l$，则：

$$\boldsymbol{f}_{ij}=\begin{cases}1, & y_i=j\\ 0, & \text{其他}\end{cases} \tag{6}$$

即若样本 $\boldsymbol{x}_i$ 属于第 j 类，则标签向量第 j 个元素为 1，其余均为 0。若 $i>l$，则 $\boldsymbol{f}_i$ 为零向量。根据相似度大的样本的标签向量应比相似度小的样本的标签向量更相似的原则，本文也将标签传播定义为下述的优化问题：

$$\min\frac{1}{2}\sum_{i,j=1}^{n}s_{ij}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2 \tag{7}$$

这个最优化问题与下述的问题等价[17]：

$$\min_{\boldsymbol{F}_u} Tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_{\boldsymbol{S}}\boldsymbol{F}) \tag{8}$$

式中：$\boldsymbol{F}\in\mathrm{R}^{n\times c}$，其第 i 行表示第 i 个样本的标签向量。为方便起见，定义 $\boldsymbol{F}=\{\boldsymbol{F}_l,\boldsymbol{F}_u\}$，$\boldsymbol{F}_l$ 表示标签数据所对应的标签矩阵，$\boldsymbol{F}_u$ 表示无标签数据的标签矩阵，目标是求解 $\boldsymbol{F}_u$。$\boldsymbol{L}_{\boldsymbol{S}}=\boldsymbol{D}-\boldsymbol{S}$ 为一个拉普拉斯矩阵，$\boldsymbol{S}$ 为相似度矩阵，$\boldsymbol{D}$ 为对角矩阵，其第 i 个对角元素 $\boldsymbol{D}_{ii}=\sum_j s_{ij}$。

现对 $\boldsymbol{L}$ 矩阵进行分块：

$$\boldsymbol{L}_{\boldsymbol{s}}=\begin{bmatrix}\boldsymbol{L}_{ll} & \boldsymbol{L}_{lu}\\ \boldsymbol{L}_{ul} & \boldsymbol{L}_{uu}\end{bmatrix} \tag{9}$$

设 $H=\min\frac{1}{2}\sum_{i,j=1}^{n}s_{ij}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2^2$，通过求解 $\frac{\partial H}{\partial \boldsymbol{F}_{\boldsymbol{u}}}=0$ 得到 $\boldsymbol{F}_u$。详细证明见文献[17]。

$$\boldsymbol{F}_u=-\boldsymbol{L}_{uu}^{-1}\boldsymbol{L}_{ul}\boldsymbol{F}_l \tag{10}$$

对于样本 $\boldsymbol{x}_i$ 的标签向量 $\boldsymbol{f}_i=\{p_{i1},p_{i2},\ldots,p_{iC}\}$，其中 p_{ij} 可以看作 $\boldsymbol{x}_i$ 属于第 j 类的概率，在标签传播过程中，样本的标签向量可以体现出样本所属类别的不确定性程度，用标签向量的熵来作为这种不确定性的度量，即：

$$H(\boldsymbol{f}_i)=-\sum_{k=1}^{c}p_{ik}\log_2 p_{ik} \tag{11}$$

熵值越小则说明所对应样本的所属类别更加明确。低熵值样本将在后续的标签传播过程中改进样本间的度量，使得传播更加准确。

2.3 迭代更新

在半监督标签传播过程中，每次从学出的标签中筛选出分类准确率高的前 k 个新标签样本，即前 k 个低熵样本信息，利用它们进行距离度量矩阵的更新，以此来进行下一轮标签传播。在此过程中，样本间相似性度量不断地向着更加准确的方向变化着。从另一方面考虑，不同的相似性度量方式体现了不同的样本结构的分布形式。如果样本的分布更加明确，算法的分类效果就会有大幅提升。本文正是基于这一点设计了迭代优化的算法不断加强样本空间的结构识别，从而提升学习效果。

2.4 算法描述

基于上述内容，本节给出迭代优化的算法—基于标签进行度量学习的图半监督学习算法(Semi-Supervised Learning algorithm of graph based on label-based Metric Learning ML-SSL)，具体伪代码描述如算法 2 所示。

算法 2: ML-SSL。

输入：数据集 D，循环次数 n，确定性样本数 k；

输出：无标签样本的类标签。

1) begin

2) 初始化 $\boldsymbol{A}$ 为单位矩阵，计算相似度矩阵 $\boldsymbol{S}$；

3) 根据标签数据构建相似对集合 M，通过解决

如下问题求得无标签样本的标签向量矩阵 $\boldsymbol{F}_u$

$$\boldsymbol{F}_u = -\boldsymbol{L}_{uu}^{-1}\boldsymbol{L}_{ul}\boldsymbol{F}_l ;$$

4) 根据标签向量计算熵值，选出前 k 个熵最小的样本加入 M ；

5) while 循环次数小于 n 或 $\boldsymbol{A}$ 不收敛 do

6) 根据 M ，调用算法 1 计算距离度量矩阵 $\boldsymbol{A}$ ；

7) 根据 $\boldsymbol{A}$ 与如下公式，计算相似度矩阵:

$$S_{ij} = \frac{e^{-dis_{\boldsymbol{A}}(x_i, x_j)}}{\sum_{k=1}^{n} e^{-dis_{\boldsymbol{A}}(x_i, x_k)}} ;$$

8) 根据 $\boldsymbol{S}$ ，求出标签向量矩阵 $\boldsymbol{F}_u$ ；

9) 根据标签向量计算熵值，选出前 k 个熵最小的样本，根据其最可能的类归属加入 M ；

10) end

11) 根据 $\boldsymbol{F}_u$ ，得到无标签样本的标签 labels；

12) return 标签 labels；

13) end

算法 2 中，2)~4)是初始的各个量的计算，包括初始化距离度量矩阵 $\boldsymbol{A}$ 、相似样本对集合 M 、标签向量矩阵 $\boldsymbol{F}_u$ 以及计算无标签样本的熵值。接着，算法进行迭代优化，通过低熵值样本与距离度量矩阵的相互作用影响彼此。循环终止的条件可以是距离度量矩阵 $\boldsymbol{A}$ 收敛，此时说明样本间的相似度已达到了最佳稳定状态，也可以是根据实际情况设定迭代次数。

3 实验与结果分析

本章设计如下实验验证本文所提算法的可行性和分类性能。实验主要分为三部分：首先，详细分析 k 值的选取情况；然后，在人造数据集上进行分类性能验证与分析；最后，在真实数据集上进行对比分析和验证。本文采用的对比算法包括：LGC、GFHF 和 S^2LRR 三种，这三种方法均为半监督学习领域的经典算法。然而，其在算法执行中均未利用中间结果，且在样本间相似性度量时未用到标签信息。通过实验结果可以看出，本文算法具有很大的竞争优势。

3.1 实验环境与数据集

本文实验所用的硬件环境为：windows10、CPU 主频为 2.00GHz、8GB 运行内存、CPU 型号为 AMD A8-6410。

本文采用的人造数据集为一个双月型数据集，该数据集由上下两个半圆形分别表示由两个类组成。每类包含 100 个二维样本，其中每个样本由两个实数描述其特征。该数据集属于非凸型数据集。从算法的二维结果可以展示其对数据空间结构的识别能力，具体如图 1 所示，其中横轴表示第一个特征，纵轴表示第二个特征，图 2~图 6 横纵坐标也是如此。

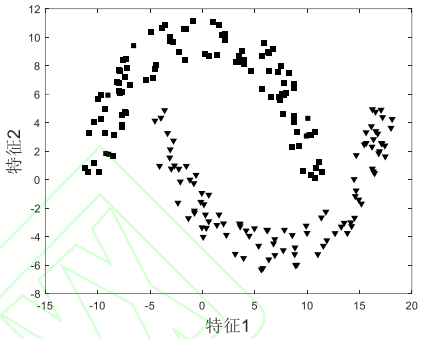


图 1 双月型数据集

Fig. 1 Two moons dataset

采用的真实数据集有 Breast、German、Ionosphere、Vote、Heart、Monkl 6 个，全部来自于 UCI。详细信息如表 1 所示，具体包括数据集名称、样本量、维数和类别数。

表 1 来自于 UCI 的真实数据集

Tab. 1 Real dataset from UCI

数据集	样本数	维数	类别数
Breast	683	9	2
German	1000	20	2
Ionosphere	351	34	2
Vote	435	16	2
Heart	270	13	2
Monkl	432	6	2

3.2 度量指标和参数

实验采用分类准确率为评价指标，即将数据集中无标签数据的真实标签 y_i 与算法学习的预测标签 $\hat{y}_i$ 做比较，即分类准确率 acc 为：

$$acc = \frac{\sum_{x_i \in X_u} I(\hat{y}_i = y_i)}{|X_u|} \quad (12)$$

其中：X_u 为无标签数据集；$|X_u|$ 为该集合所包含的样本量。

除了分类准确率外，还增加了一项标签样本占比的数据。实验通过这两项指标来说明本文所提算法在提高分类准确率方面的优势，从不同数据集的对比中可以体现所提算法的鲁棒性。

实验中，算法的参数设置为最大迭代次数 $n=20$、判断度量矩阵 $\boldsymbol{A}$ 收敛的 $\varepsilon=0.01$，即 $\left\|\boldsymbol{A}^{(t)}-\boldsymbol{A}^{(t-1)}\right\|_2 \le \varepsilon$，则认为 $\boldsymbol{A}$ 收敛、设置低熵样本个数 $k=5$。

3.3 k 值对分类结果的影响

本文对 k 值的选择是通过在 6 个真实数据集上进行交叉验证得出，具体是分别在每个数据集上，随机分配 12 个标签，分别取 $k=2$、$k=5$、$k=10$ 进行交叉验证，具体结果如表 2 所示。

表 2 交叉验证实验结果

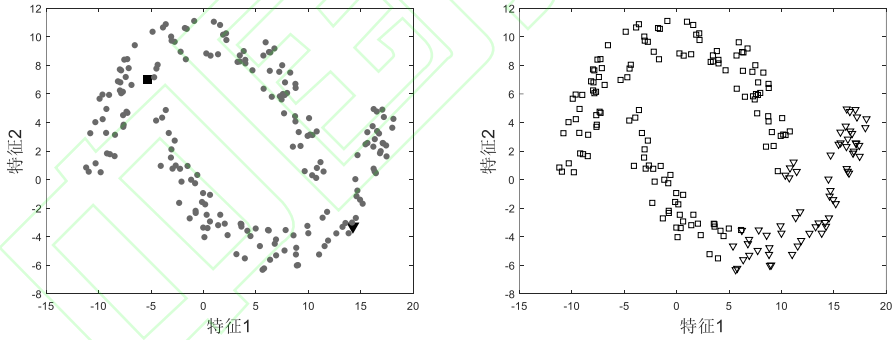
Tab. 2 Cross-validation of experimental results

数据集	标签数	k	准确率
Breast	12	2	0.8450
		5	**0.9449**
		10	0.8459
German	12	2	0.5941
		5	**0.7267**
		10	0.6711
Ionosphere	12	2	0.8032
		5	**0.8289**
		10	0.6519
Vote	12	2	0.9054
		5	**0.9362**
		10	0.8751
Heart	12	2	0.7752
		5	**0.8023**
		10	0.7209
Monkl	12	2	0.5952
		5	**0.6571**
		10	0.5000

从表 2 可以看出，当 $k=5$ 时分类效果最佳，具体见表中加粗部分。这是由于 k 的选择会对低熵值样本改进距离度量矩阵产生影响。具体地，当 $k=2$，即取较小的值时，算法将会退化成固定度量方法，每次迭代所选出的确定性样本基本保持不变。当 $k=10$，即取较大的值时，学习过程将会引入确定性较低的无标签样本，这样的样本对于距离度量矩阵的学习来说属于噪声影响。

3.4 人造数据集上的实验验证与分析

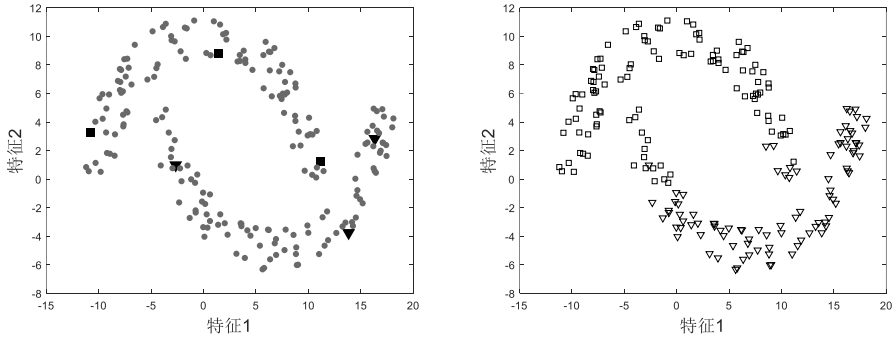
在人造数据集上，通过随机地为每类样本分配一定数量的标签来进行对比实验。这里分别随机地给每个类分配 1、3、5、7、9 个标签，使用本文所提算法 2 进行其余标签的学习，具体学习结果如图 2~6 所示。每张图左边是初始数据集(initial data)，图中正方形和倒三角分别表示两类标签数据点，圆形表示无标签数据点，右边为运用所提算法 2 的分类结果(result)。



(a) initial data (b) result

图 2 每类样本具有 1 个标签的分类结果

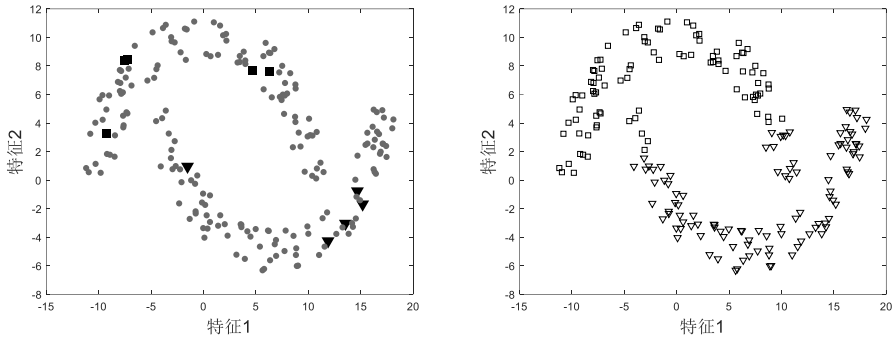
Fig. 2 Classification results on sample data that each class has one label



(a) initial data (b) result

图 3 每类样本具有 3 个标签的分类结果

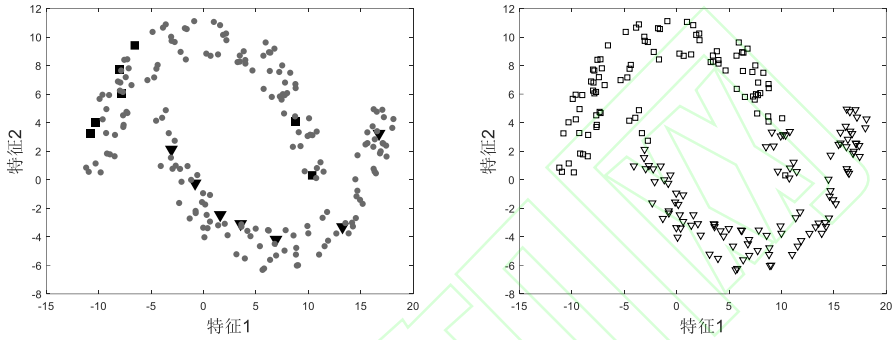
Fig. 3 Classification results on sample data that each class has three labels



(a) initial data (b) result

图 4 每类样本具有 5 个标签的分类结果

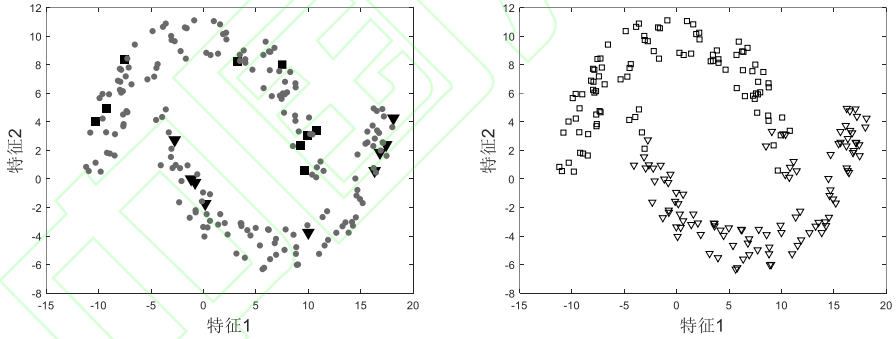
Fig. 4 Classification results on sample data that each class has five labels



(a) initial data (b) result

图 5 每类样本具有 7 个标签的分类结果

Fig. 5 Classification results on sample data that each class has seven labels



(a) initial data (b) result

图 6 每类样本具有 9 个标签的分类结果

Fig. 6 Classification results on sample data that each class has nine labels

由图 2~图 6，可以得出：

1)本文所提算法在标签样本占比很小的情况下就可以得出较好的分类结果。

2)本文算法随标签样本数量的增加分类效果明显增强，尤其是对每类数据尾部的样本分类，即从实验结果可以看出，分类错误的样本主要集中在每个类的尾部。随着标签样本数量的提升，本文的算法加强了对尾部数据的分类能力。

3.5 真实数据集上的实验及其结果分析

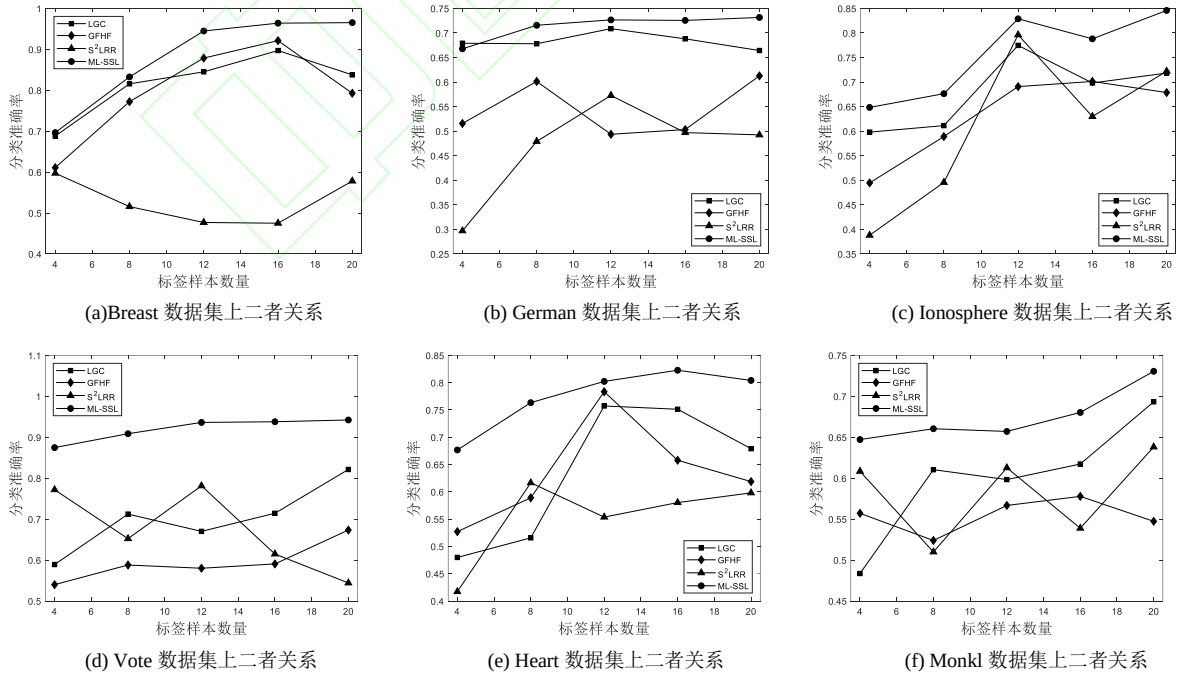
本小节在 6 个真实数据集上进行实验验证，并与 LGC[6]、GFHF[17]以及 S^2LRR[2] 3 种半监督学习算法进行实验对比。每个数据集上随机地给每个类分配 2、4、6、8、10 个标签。4 种算法在不同数据集、不同标签数量下的分类结果具体如表 3 所示。其中，粗体表示每种情况下取得的最高分类准确率。

另外，随着标签样本数量地增加，四种算法取得的分类准确率变化如图 7 所示，子图(a)-子图(f)分别代表 Breast、German、Inonsphere、Vote、Heart 以及 Monkl 数据集上二者间的关系。其中，横轴表示标签数量，纵轴表示分类准确率。

表 3 真实数据集上的分类准确率

Tab. 3 Classification accuracy on real data sets

数据集	标签数量	标签数占比	LGC	GFHF	S^2LRR	DMLSSL
Breast	4	0.0059	0.6882	0.6109	0.5973	**0.6966**
	8	0.0017	0.8157	0.7721	0.5157	**0.8326**
	12	0.0176	0.8451	0.8787	0.4769	**0.9449**
	16	0.0234	0.8971	0.9213	0.4751	**0.9640**
	20	0.0293	0.8377	0.7927	0.5781	**0.9653**
German	4	0.0040	**0.6791**	0.5157	0.2971	0.6678
	8	0.0080	0.6781	0.6013	0.4791	**0.7157**
	12	0.0120	0.7088	0.4937	0.5727	**0.7267**
	16	0.0160	0.6881	0.5033	0.4971	**0.7256**
	20	0.020	0.6643	0.6127	0.4925	**0.7316**
Ionosphere	4	0.0114	0.5981	0.4947	0.3881	**0.6484**
	8	0.0228	0.6114	0.5891	0.4957	**0.6764**
	12	0.0342	0.7747	0.6908	0.7962	**0.8289**
	16	0.0456	0.6982	0.7013	0.6299	**0.7881**
	20	0.0570	0.7184	0.6787	0.7221	**0.8459**
Vote	4	0.0092	0.5891	0.5399	0.7719	**0.8747**
	8	0.0184	0.7123	0.5879	0.6522	**0.9087**
	12	0.0276	0.6701	0.5800	0.7813	**0.9362**
	16	0.0386	0.7143	0.5906	0.6148	**0.9379**
	20	0.046	0.8213	0.6733	0.5441	**0.9422**
Heart	4	0.0148	0.4798	0.5271	0.4172	**0.6767**
	8	0.0296	0.5157	0.5891	0.6166	**0.7634**
	12	0.0444	0.7573	0.7835	0.5537	**0.8023**
	16	0.0593	0.7513	0.6578	0.5803	**0.8228**
	20	0.0741	0.6791	0.6188	0.5981	**0.8041**
Monkl	4	0.0093	0.4836	0.5571	0.6084	**0.6472**
	8	0.0185	0.6104	0.5239	0.5099	**0.6604**
	12	0.0278	0.5983	0.5667	0.6127	**0.6571**
	16	0.0370	0.6173	0.5778	0.5389	**0.6803**
	20	0.0463	0.6933	0.5473	0.6381	**0.7306**



(a)Breast 数据集上二者关系 (b) German 数据集上二者关系 (c) Ionosphere 数据集上二者关系

(d) Vote 数据集上二者关系 (e) Heart 数据集上二者关系 (f) Monkl 数据集上二者关系

图 7 不同数据集上标签数量与不同算法分类准确率的关系

Fig. 7 Relationship between number of labels on data set and classification accuracy with different algorithm

由表 3 和图 7 可以得出:

1)除 German 数据集上已知标签数是 4 的情况外，其余情况下本文所提算法均取得了最高的分类准确率，即本文所提算法在 6 个数据集上准确率最高的情况占比达到 96.7%(29/30)。具体地，在 Breast 数据上，本文算法比其他 3 种算法在准确率上平均提高了 16.72 个百分点；在 German 数据集上提高了 14.79 个百分点；在 Ionosphere 数据集上提高了 11.84 个百分点；在 Vote 数据集上提高了 26.37 个百分点；在 Heart 数据集上提高了 16.55 个百分点；在 Monkl 数据集上提高了 9.62 个百分点。由此可见，本文所提算法在绝大多数情况下的分类准确率均优于其他 3 种算法。

2)每个数据集上已知标签比例范围为[0.0017, 0.0741]，可见本文所提算法在已知标签比例很低的情况下便可取得相比于其他 3 种算法更高的分类准确率。在 Breast 数据集上当标签数为 20，标签占比仅为 0.0293 时，分类准确率达到了 0.9653；在 German 数据集中，最高分类准确率达到了 0.7316，此时标签占比仅为 0.02；在 Votes 数据中，在标签占比仅为 0.0092 即不到百分之一的情况下，本文所提算法的分类准确率达到了 0.8747。从上述分析可知，在标签数极少的情况下，本文所提算法也能实现较高准确率的分类效果。

3)从图 7 的子图(a)、(c)、(e)、(f)可以看出，本文所提算法在 Breast 数据集、Ionosphere 数据集、Heart 数据集以及 Monkl 数据集上的分类准确率随标签数的增加而增加。而其他 3 种半监督学习算法，分类准确率并不随着标签数量的增加而直线提高，尤其 S2LRR 算法，起伏很大。在 German 数据集和 Vote 数据集上，通过对比 4 种算法所对应的曲线也可以明显看出，本文所提算法比其他 3 个算法更加平稳，表明了本文算法具有较好的鲁棒性。

4 结语

为更准确地反应样本间相似性关系以及充分利用中间结果来提高半监督学习的分类准确率，本文提出了一种基于标签进行度量学习的图半监督学习算法，利用标签传播过程中确定性标签样本来不断修正样本间的相似性度量方式。然后，通过迭代算法使得以样本为节点、相似度为边权重的图不断得以优化，从而使得标签传播更加准确，并通过实验验证了所提算法的良好分类性能。接下来，我们进一步的研究工作包括该算法的合理性理论分析、计算效率的提高、算法中参数 k 的选取方法以及该算法的实际应用研究等方面。

参考文献

[1] LI C G, LIN Z C, ZHANG H G, et al. Learning semi-supervised representation towards a unified optimization framework for semi-supervised learning [C] // Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015, 2767-2775.

[2] ZHOU Z H. A brief introduction to weakly supervised learning [J]. National Science Review, 2017, 5(1): 48-57.

[3] MEY A, LOOG M. Improvability through semi-supervised learning: a survey of theoretical results [EB/OL]. [2020-05-09]. https://arxiv.org/pdf/1908.09574.pdf.

[4] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法 [J]. 计算机学报, 2015, 000(8): 1592-1617.（LIU J W, LIU Y, LUO X L. Semi-supervised learning methods [J]. Journal of Computer Science, 2015, 000(8): 1592-1617.）

[5] ZHANG Y M, ZHANG X Y, YUAN X T, et al. Large-scale graph-based semi-supervised learning via tree Laplacian solver [J]. State Key Laboratory of Pattern Recognition, 2016, 13(7): 2344-2350.

[6] ZHOU D Y, BOUSQUET O, LAL T N. Learning with local and global consistency [J]. Advances in Neural Information Processing Systems, 2003, 16(3): 321-328.

[7] NIE F P, SHI S J, LI X L. Semi-supervised learning with auto-weighting feature and adaptive graph [J]. IEEE Transactions on Knowledge and Data Engineering, 2019: 1-1.

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

[9] BELKIN M, NIYOGI P. Laplacian Eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2003, 15(6): 1373-1396.

[10] BELKIN M, NIYOGI P. Semi-supervised learning on Riemannian manifolds [J]. Machine Learning, 2004, 56(1): 209-239.

[11] SAUL L K, ROWEIS S T. Think globally, fit locally: unsupervised learning of low dimensional manifolds [J]. Journal of Machine Learning Research, 2003, :4(2) 119-155.

[12] LIU G C, LIN Z C, YU Y. Robust subspace segmentation by low-rank representation[C] // Proceedings of the 2010 International Conference on Machine Learning. Washington, DC: IEEE Computer Society, 2010: 663-670.

[13] CHENG H, LIU Z C, YANG J. Sparsity induced similarity measure for label propagation [C]// Proceedings of the 2009 International Conference on Computer Vision. Piscataway: IEEE, 2009: 317-324.

[14] CHENG B, YANG J C, YAN S C, et al. Learning with-graph for image analysis[J]. IEEE Transactions on Image Processing, 2010, 19(4): 858-866.

[15] HE R, ZHENG W S, HU B G, et al. Nonnegative sparse coding for discriminative semi-supervised learning [C] // Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2011: 2849-2856.

[16] LI H, ZHANG J, HU J, et al. Graph-based discriminative concept factorization for data representation[J]. Knowledge Based Systems, 2017, 118: 70-79.

[17] ZHU X J, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions [C] // Proceedings of the 2003 International Conference on Machine Learning. Washington, DC: IEEE Computer Society, 2003: 912-919.

[18] NIE F P, XIANG S M, LIU Y, et al. A general graph-based semi-supervised learning with novel class discovery[J]. Neural Computing & Applications, 2010, 19(4): 549-555.

[19] RUSTAMOV R M, KLOSOWSKI J T. Interpretable graph-based semi-supervised learning via flows [C]// Proceedings of the 2018 AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018: 3976-3983.

[20] LI Q M, WU X M, GUAN Z C. Generalized label propagation methods for semi-supervised learning [EB/OL]. [2020-05-09]. https://arxiv.org/pdf/1901.09993.pdf.

[21] SZUMMER M, JAAKKOLA T S. Partially labeled classification with Markov random walks [C] // Proceedings of the 2001 International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2001: 945-952.

[22] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10: 207-244.

[23] SONG K, NIE F P, HAN J W, et al. Parameter free large margin nearest neighbor for distance metric learning [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park, CA:AAAI Press, 2017: 2555-2561.

[24] XING E P, NG A Y, JORDAN M, et al. Distance metric learning with application to clustering with side-information [C]// Proceedings of the 2002 International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 521-528.

[25] LAURENS V D M, HINTON G. Visualizing high-dimensional data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.

This work is partially supported by the Natural Science Foundation of Shanxi Province (201801D121115), and the Postdoctoral Science Foundation of China (2020-095).

LYU Yali, born in 1975, Ph.D., associate professor, CCF member. Her research interests include probabilistic reasoning, data mining and machine learning.

MIAO Junzhong, born in 1993, M.S. candidate. His research interests include machine learning and data mining.

HU Weixin, born in 1996, M.S. candidate. Her research interests include machine learning and data mining.