DOI: 10.11992/tis.201809019
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181225.1017.002.html

结合谱聚类的标记分布学习

王一宾[1,2], 李田力[1], 程玉胜[1,2]
(1. 安庆师范大学 计算机与信息学院, 安徽 安庆 246011; 2. 安徽省高校智能感知与计算重点实验室, 安徽 安庆 246011)

摘 要: 标记分布是一种新的学习范式, 现有算法大多数直接使用条件概率建立参数模型, 未充分考虑样本之间的相关性, 导致计算复杂度增大。基于此, 引入谱聚类算法, 通过样本之间相似性关系将聚类问题转化为图的全局最优划分问题, 进而提出一种结合谱聚类的标记分布学习算法 (label distribution learning with spectral clustering, SC-LDL)。首先, 计算样本相似度矩阵; 然后, 对矩阵进行拉普拉斯变换, 构造特征向量空间; 最后, 通过 K-means 算法对数据进行聚类建立参数模型, 预测未知样本的标记分布。与现有算法在多个数据集上的实验表明, 本算法优于多个对比算法, 统计假设检验进一步说明算法的有效性和优越性。

关键词: 谱聚类; 标记分布学习; 相似度矩阵; 拉普拉斯变换; K-均值; 参数模型; 标记分布; 机器学习
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673−4785(2019)05−0966−08

中文引用格式: 王一宾, 李田力, 程玉胜. 结合谱聚类的标记分布学习 [J]. 智能系统学报, 2019, 14(5): 966–973.
英文引用格式: WANG Yibin, LI Tianli, CHENG Yusheng. Label distribution learning based on spectral clustering[J]. CAAI transactions on intelligent systems, 2019, 14(5): 966–973.

Label distribution learning based on spectral clustering

WANG Yibin[1,2], LI Tianli[1], CHENG Yusheng[1,2]
(1. School of Computer and Information, Anqing Normal University, Anqing 246011, China; 2. Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing 246011, China)

Abstract: Label distribution is a new learning paradigm. Most of the existing algorithms use conditional probability to build parametric models but do not consider the links between samples fully, which increases computational complexity. On this basis, the spectral clustering algorithm is introduced to transform the clustering problem into the global optimum graph partitioning problem based on the similarity relation between samples. Thus, a label distribution learning algorithm combined with spectral clustering (SC-LDL) is proposed. First, we calculate the similarity matrix of the samples. Then, we transform the matrix using the Laplace transform to construct the feature vector space. Finally, we cluster the data to establish the parameter model with K-means algorithm and use this new model to predict the label distribution of unknown samples. The comparison between SC-LDL and the existing algorithm on multiple data sets shows that this algorithm is superior to multiple contrast algorithms. Furthermore, statistical hypothesis testing illustrates the effectiveness and superiority of the SC-LDL algorithm.

Keywords: spectral clustering; label distribution learning; similarity matrix; Laplace transform; K-means; parametric model; label distribution; machine learning

标记多义性学习作为机器学习领域中一个热门方向, 目前单标记学习与多标记学习已经较为成熟[1-2]。其中多标记学习是单标记学习的拓展, 许多已有的研究成果也证实了多标记学习是一种有效的学习范式[3]。传统的多标记学习虽然可以有效地处理实例和标记之间的关系, 但却将标记进行了简单的定义, 即存在为“1”不存在为“−1”,

收稿日期: 2018−09−13. **网络出版日期**: 2018−12−26.
基金项目: 安徽省高校重点科研项目 (KJ2017A352).
通信作者: 程玉胜. E-mail: chengyshaq@163.com.

忽略了标记对实例的描述有着不同重要程度这一问题, 因此并不适用于真实世界中某些实例, 例如人的情感分析。人的情感可以通过面部表情来表达, 而面部表情往往是一种混合了多种基本情绪的综合表情, 例如: 惊讶和恐惧可能会同时出现, 高兴和感动可能会在同一时刻表现出来; 这些基本情绪共同组成了一种情绪分布, 在表达的情感中各自占有不同的比重。针对这类实例, Geng 等[4] 提出了一种标记分布学习范式 (label distribution learning, LDL)。

基于此, 为了更好地研究标记分布学习, 研究者们提出了许多预测标记分布的算法。目前已有的算法主要分为 3 类: 1) 为了使用现有学习范式的相关公式定义, 提出了将标记分布学习退化成单标记学习的 PT-Bayes(problem transformation bayes) 算法和 PT-SVM(problem transformation support vector machine) 算法[5]; 2) 将机器学习领域相关算法引入到标记分布学习中, 以提高预测精度, 如 AA-KNN 算法 (algorithm adaptation k nearest neighbors) 和 AA-BP 算法 (algorithm adaptation back propagation)[6]; 3) 为了切合标记分布学习自身的特性, 而专门设计的 CPNN(conditional probability neural network) 算法[7]、SA-IIS(specialized algorithms improved iterative scaling) 和 SA-BFGS 算法 (specialized algorithms broyden fletcher goldfarb shanno)[8]。然而上述算法均未考虑样本之间联系, 导致样本数目过多, 计算时间过长, 甚至受异常样本数据的影响使得预测精度下降。为了解决这种问题, 本文考虑引入聚类方法处理。

聚类是一种无监督的处理数据并寻找内部关联的分类方式[9]。目前聚类算法以及改进算法很多, 如 K-means 和模糊聚类等[10-11]。聚类对简化训练模型效果优秀, 其中, 张敏灵[12] 在多标记学习中对输入空间进行 K-means 聚类, 构建类属属性, 然后训练对应的分类模型; 邵东恒等[13] 将 K-means 引入标记分布学习, 取得了较好的效果。但是这些传统聚类算法依托于凸球形样本空间, 并且由于采取迭代更新, 算法容易陷入局部最优, 同时对数据适应性也不够[14]。为了避免传统聚类这些弊端导致算法可能不具备一般性[15], 故而引入谱聚类方法。谱聚类近年被引入了机器学习领域, 并迅速成为热点之一[16]。谱聚类算法中最经典的算法是 Ng 等[17] 提出的 NJW (Ng Jordan Weiss) 算法, NJW 算法聚类效果也较为理想。

因此, 为了解决以上传统聚类的问题, 将谱聚类中 NJW 算法引入到标记分布学习中, 提出了一种结合谱聚类的标记分布学习算法 (label distribution learning with spectral clustering, SC-LDL)。特征相似的样本对应的标记分布理论上也是相似的, 将这些聚类中心作为新的输入空间, 可以得到新的预测标记分布。相比于已有算法, 本算法考虑了样本之间的联系, 并且通过谱聚类预处理后的样本数据有效减少了相似数据和异常数据对模型复杂度及预测精度的影响, 相比于其他聚类算法既不限定样本空间, 也不需要迭代更新, 同时本文通过在 12 个数据集上和 5 个经典算法进行对比实验, 证明了本文所提算法的有效性。

1　相关知识

1.1　谱聚类算法

谱聚类 (spectral clustering)[18] 是一种基于谱图理论的算法, 该算法是将数据聚类问题转换成寻求图的最优分割, 它是一种点对聚类算法, 并且适用于非凸数据[19]。谱聚类基本步骤可以描述为: 通过高斯核函数计算原始数据集对应的相似矩阵 $\boldsymbol{W}$, 进而得到一个度矩阵 $\boldsymbol{D}$, 之后进行拉普拉斯变换, 对转换后的拉普拉斯矩阵 $\boldsymbol{L}$ 进行特征值分解, 找出前 m 个特征值对应的特征向量, 并将特征向量按列存储得到新的特征矩阵, 对新的特征矩阵进行 K-means 聚类处理[20]。谱聚类算法一经提出, 其在数据降维、不规则数据聚类、图像分割等方面获得了广泛的应用。

定义 1　高斯核函数。定义一个 d 维的欧式空间 $\boldsymbol{P}$, 其中有一点 $p \in \boldsymbol{X}$, p 到某一中心点 p_0 之间欧氏距离的单调函数记作 $k(\|p-p_0\|)$, 称之为核函数。高斯核函数是一种常用的核函数, 定义为

$$k_G(\|p-p_o\|) = \exp\left(-\frac{\|p-p_o\|^2}{2\sigma^2}\right) \tag{1}$$

式中: p 属于欧式空间 $\boldsymbol{P}$; p_o 为核函数中心; σ 为函数的宽度参数。

定义 2　相似矩阵。相似矩阵又称亲和矩阵, 其与原矩阵特征值和特征向量相同, 本文用 $\boldsymbol{W}$ 来表示, 矩阵定义为

$$\boldsymbol{W}_{ij} = \exp\left(-\frac{d(s_i, s_j)}{2\mu^2}\right) \tag{2}$$

式中: s_i 和 s_j 表示样本数据点; $d(s_i, s_j)$ 通常取 $\|s_i - s_j\|^2$; μ 是预先设定的尺度因子, 用来控制样本点 s_i 和 s_j 之间距离对 $\boldsymbol{W}_{ij}$ 的影响。

定义 3　度矩阵。将 $\boldsymbol{W}$ 的每行元素相加, 将

相加后的数值用作对角元素构建对角矩阵，称之为度矩阵，其中非对角元素取值为 0。本文用 $\boldsymbol{D}$ 来表示度矩阵，其定义为

$$\boldsymbol{D}_{ij}=\sum_{i,j=1}^{n}\boldsymbol{W}_{ij} \tag{3}$$

定义 4　**拉普拉斯矩阵**。拉普拉斯矩阵主要应用在图论中，是一种用来表示图的矩阵，用 $\boldsymbol{L}$ 来表示，本文选取不规范拉普拉斯矩阵，其定义为

$$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{4}$$

1.2　标记分布学习

标记分布学习是一种更加贴近真实世界的分布学习范式[21]。在传统的多标记学习中，每一个实例对应一个标记集合，其中每一个标记的描述度为“1”或“−1”。当将这种描述度用一种类似于概率分布的形式来表示，即为标记分布。一个实例对应一个标记分布的学习过程，称为标记分布学习[22]。真实世界中的实例所对应的标记多是有重要程度之分的，标记分布学习考虑这些重要程度，并且结合概率分布理论，将标记集合通过一种概率分布的形式来表达。

定义 5　在标记分布学习中，每一个用来描述实例 x 的标记 y 用 d_x^y 来标注重要程度，不失一般性，可知 $d_x^y\in[0,1]$，且 $\sum_y d_x^y=1$。

定义 6　给定标记分布学习中的训练集 $\boldsymbol{S}=\{(\boldsymbol{x}_1,\boldsymbol{D}_1),(\boldsymbol{x}_2,\boldsymbol{D}_2),(\boldsymbol{x}_3,\boldsymbol{D}_3),\cdots,(\boldsymbol{x}_n,\boldsymbol{D}_n)\}$，其中实例矩阵 $\boldsymbol{X}=\left[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n\right]$，$\boldsymbol{x}_i\in\mathbf{R}^d$ 表示 $\boldsymbol{X}$ 中第 i 个实例，$i=1,2,\cdots,n$。标记矩阵 $\boldsymbol{Y}=\left[\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_c\right]$，$y_j$ 是第 j 个标记，$j=1,2,\cdots,c$。训练集对应的标记分布为 $\boldsymbol{D}=\left[\boldsymbol{D}_1\ \boldsymbol{D}_2\ \cdots\ \boldsymbol{D}_n\right]$，实例 x_i 的标记分布为 $\boldsymbol{D}_i=\left[d_{x_i}^{y_1}\ d_{x_i}^{y_2}\ \cdots\ d_{x_i}^{y_c}\right]$。$d_{x_i}^{y_j}$ 为标记 y_j 对实例 x_i 的描述程度。

目前的标记分布学习通过训练集 $\boldsymbol{S}$ 得到条件概率 $P(y|x)$，进而得到预测标记分布。

2　结合谱聚类算法的标记分布学习算法的构造

2.1　谱聚类 NJW 算法

给定训练集 $\boldsymbol{S}$，将这些数据点看作图节点，相似度矩阵为 $W=\{W_{ij}|1\leqslant i\leqslant n,1\leqslant j\leqslant n\}$，相似性按照式 (2) 来计算。根据式 (3) 得到一个归一化的对角矩阵 $\boldsymbol{D}$，对 $\boldsymbol{D}$ 进行拉普拉斯变换，根据式 (4) 得到归一化拉普拉斯图矩阵 $\boldsymbol{L}$。计算 $\boldsymbol{L}$ 的特征向量，按照从小到大排列，将前 m 个特征向量按列排序得到新的矩阵 $\boldsymbol{X}'=[\boldsymbol{x}'_1\ \boldsymbol{x}'_2\ \cdots\ \boldsymbol{x}'_m]$，将该矩阵 $\boldsymbol{X}'$ 作为新的输入再进行 K-means 聚类，得到簇类中心 $\boldsymbol{C}$。

算法 1　NJW 算法

输入　预设参数 σ、k，其中 k 值选择样本数的 20%（k 为聚类数目），标记分布训练集 $\boldsymbol{S}=\{(\boldsymbol{x}_1,\boldsymbol{D}_1),(\boldsymbol{x}_2,\boldsymbol{D}_2),\cdots,(\boldsymbol{x}_n,\boldsymbol{D}_n)\}$；

输出　簇类中心 $\boldsymbol{C}$。

1) 利用式 (1) 和式 (2) 计算相似矩阵 $\boldsymbol{W}$；

2) 通过式 (3) 计算度矩阵 $\boldsymbol{D}$，度矩阵主对角线上的元素 D_{ij} 为相似矩阵 $\boldsymbol{W}$ 的第 i 行元素之和，其余元素均为 0；

3) 通过式 (4) 对度矩阵 $\boldsymbol{D}$ 进行拉普拉斯变换，得到变换后的矩阵 $\boldsymbol{L}$；

4) 对矩阵 $\boldsymbol{L}$ 进行特征值分解；

5) 升序排列；

6) 得出前 m 个特征值对应的特征向量，按列存储得到新的特征矩阵 $\boldsymbol{X}'$；

7) 对矩阵 $\boldsymbol{X}'$ 进行 K-means 聚类得到新的聚类中心；

8) 初始样本点 S_i 划分为 j 个聚类，当且仅当样本矩阵的第 i 行被划分为第 j 聚类；

9) 输出聚类中心 $\boldsymbol{C}$；

10) 结束。

通过 NJW 算法，将原有标记分布集合划分成 q 个不相关的簇类，其中簇类中心 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_q\}$，$q=1,2,\cdots,k$。

2.2　结合谱聚类算法的构造过程

当使用 NJW 算法对样本数据聚类得到簇类中心 C 后，对应的标记分布划分为相应簇类，对应新的分布矩阵，命名为 $\boldsymbol{S}'$。使用 KNN 分类器对 $\boldsymbol{S}'$ 中的样本 x' 进行分类，通过式 (5) 计算，得到最终预测标记分布。

$$P(y_j'|x')=\frac{1}{k}\sum_{i\in N_k(x)}d_{x_i}^{y_j} \tag{5}$$

式中：y_j' 为 $\boldsymbol{S}'$ 中标记；k 为示例 x' 的最近邻居数；$N_k(x)$ 表示 $\boldsymbol{S}'$ 中 x' 的索引集。

算法 2　结合谱聚类 NJW 的标记分布学习算法

输入　簇类中心 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_q\}$，$k$ 值；

输出　$P(y_j'|x')$。

1) 计算当前数据点和各个数据点的距离；

2) 将该距离升序排列；

3) 选择和当前数据点最近的 k 个数据点；

4) 判断这 k 个数据点对应类别的出现频率；

5) 将最高出现频率的这 k 个数据点对应类别作为当前数据点的预测分类；

6) 结束。

3　实验与结果分析

为了验证本文 SC-LDL 算法有效性，选取 5 种经典算法在 12 个标准数据集上进行对比实验。相关实验结果还使用统计分析中 Nemenyi 检验来进一步表明算法有效性。其中 Nemenyi 检验是统计学中一种针对成组数据的有效检验方法。

3.1　标记分布学习评价指标

标记分布学习输出的是一个标记分布，评价算法并不能简单地用标记准确度多少来建立。根据 Geng 等在文献 [4] 中提出的对标记分布学习评价算法的建议，本实验采用了 6 种代表性的标记分布评价指标，分别为 Chebyshev 距离、Clark 距离、Canberra 距离、Kullback-Leibler(KL-div) 散度、余弦相关系数 (Cosine) 和交叉相似度 (Intersection)。其中，前 4 个指标是衡量距离的指标，后两个是相似度指标。假设有 c 个标记的实例，真实标记分布为 $\boldsymbol{D}=[d_1\ d_2\ \cdots\ d_c]$，预测标记分布为 $\hat{\boldsymbol{D}}=[\hat{d}_1\ \hat{d}_2\ \cdots\ \hat{d}_c]$，各个指标的计算公式见表 1，其中↓表示该指标越低越好，而↑表示该指标越高越好。

表 1　标记分布评价指标

Table 1　Evaluation measures for label distribution learning

评价指标	计算公式
Chebyshev距离↓	$\mathrm{Dis}_1=\max_j\lvert d_j-\hat{d}_j\rvert$
Clark距离↓	$\mathrm{Dis}_2=\sqrt{\sum_{j=1}^{n}\frac{(d_j-\hat{d}_j)^2}{(d_j+\hat{d}_j)^2}}$
Canberra距离↓	$\mathrm{Dis}_3=\sum_{j=1}^{n}\frac{\lvert d_j-\hat{d}_j\rvert}{d_j+\hat{d}_j}$
KL-div散度↓	$\mathrm{Dis}_4=\sum_{j=1}^{n}d_j\ln\frac{d_j}{\hat{d}_j}$
Cosine相似度↑	$\mathrm{Dis}_5=\frac{\sum_{j=1}^{n}d_j\hat{d}_j}{\sqrt{\sum_{j=1}^{n}d_j^2}\sqrt{\sum_{j=1}^{n}\hat{d}_j^2}}$
Intersection相似度↑	$\mathrm{Dis}_6=\sum_{j=1}^{n}\min(d_j,\hat{d}_j)$

3.2　实验数据与环境描述

实验数据：本文 SC-LDL 算法与 5 种常用经典算法进行对比，使用的 12 个数据集相关信息如表 2。Yeast 类都是在酵母菌上做实验收集到的真实数据，每个数据集中含有 2 465 个酵母菌基因，每个基因由 24 个特征表达，标记对应不同的时间点，标记分布是不同的时间点上基因表达水平。s-JAFFE 和 SDU_3DFE 数据集是两个常用的表情数据集的拓展，其中 s-JAFFE 数据集中包含 213 张对 10 名日本女性模特进行人脸采集得到表情灰度图，特征向量是由 Local Binary Patterns 方法抽取图像特征获得的，实例中的标记分布包含了开心、难过、惊讶、害怕、生气和厌恶 6 种情感。SDU_3DFE 数据集与 s-JAFFE 不同的是，它包含了 2 500 张表情灰度图。Movie 是用户对电影评级的数据集。所有数据来自于 Netflix，电影评级分 5 级，作为标记，用户在各个级别上的评价比例作为标记分布存在，该数据集所有特征均来自于电影相关属性的提取。

表 2　数据集

Table 2　Data sets

数据集	样本	特征	标记
Yeast-alpha	2 465	24	18
Yeast-cdc	2 465	24	15
Yeast-diau	2 465	24	7
Yeast-heat	2 465	24	6
Yeast-spo	2 465	24	6
Yeast-cold	2 465	24	4
Yeast-dtt	2 465	24	4
Yeast-spo5	2 465	24	3
Yeast-elu	2 465	24	14
s-JAFFE	213	243	6
SDU_3DFE	2 500	243	6
Movie	7 700	1 869	5

本文所有实验均在 Matlab2016a 中运行，硬件环境 Intel® Core™i5-7500 3.40 GHz CPU，操作系统为 Windows 10。

本文 SC-LDL 为了加强说服力，使用十折交叉验证来进行实验，即每次进行实验时将数据集随机分为 10 个部分，其中 1 份数据集用来测试，剩余 9 份作为训练数据集，总计进行 10 次实验，综合 10 次实验得到的评价指标结果求出平均值 (mean) 和标准差 (std)。

3.3　实验结果与分析

表 3~8 给出了本文算法与 5 种对比算法在 12 个数据集上的实验对比结果，其中在每一个数据集上运行的最优结果用黑体表示；表格最后一行给出了 6 种算法的算法排位。算法排位越小，表示算法总体性能越好。

对以上 6 项评价指标结果进行统计检验，结果如图 1 所示。

表 3　Chebyshev(↓) 指标结果

Table 3　Results in Chebyshev (↓)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.017 0±0.002 0	0.099 6±0.005 1	0.014 6±0.000 4	0.036 7±0.001 3	0.017 0±0.000 5	**0.013 5±0.000 1**
Yeast-cdc	0.020 0±0.003 1	0.108 7±0.010 1	0.017 5±0.000 4	0.038 8±0.002 1	0.020 0±0.000 5	**0.016 2±0.000 1**
Yeast-elu	0.019 0±0.001 0	0.112 1±0.006 1	0.017 6±0.000 4	0.038 9±0.001 8	0.020 3±0.000 4	**0.016 3±0.000 1**
Yeast-diau	0.046 0±0.004 0	0.157 7±0.006 9	0.039 1±0.001 3	0.050 1±0.002 0	0.041 2±0.000 9	**0.037 5±0.001 2**
Yeast-heat	0.046 0±0.001 1	0.174 1±0.010 5	0.044 7±0.001 7	0.053 0±0.002 7	0.046 6±0.001 1	**0.042 7±0.001 2**
Yeast-spo	0.065 0±0.006 1	0.175 4±0.009 0	0.062 9±0.002 7	0.067 1±0.003 5	0.061 3±0.001 5	**0.058 1±0.001 7**
Yeast-cold	0.057 4±0.003 1	0.183 1±0.013 3	0.055 0±0.001 7	0.059 1±0.002 6	0.056 6±0.001 6	**0.050 8±0.001 4**
Yeast-dtt	0.040 1±0.001 1	0.181 8±0.013 4	0.039 7±0.001 7	0.043 3±0.001 5	0.043 6±0.001 3	**0.036 6±0.001 3**
Yeast-spo5	0.092 9±0.006 0	0.201 3±0.013 3	0.097 0±0.005 4	0.094 2±0.003 5	0.095 0±0.002 3	**0.092 0±0.002 7**
s-JAFFE	0.127 1±0.017 1	0.123 2±0.008 3	**0.098 8±0.013 9**	0.139 1±0.012 9	0.116 0±0.014 1	0.114 7±0.010 8
s-BU_3DFE	0.119 1±0.006 0	0.138 6±0.003 7	**0.103 0±0.003 5**	0.142 9±0.006 4	0.134 4±0.005 0	0.133 2±0.006 1
Movie	0.213 0±0.039 1	0.201 4±0.002 8	**0.124 0±0.002 6**	0.139 1±0.003 4	0.147 3±0.002 0	0.129 7±0.003 1
算法排位	3.750 0	5.666 7	2.166 7	4.666 7	3.416 7	1.333 3

表 4　Clark(↓) 指标结果

Table 4　Results in Clark (↓)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.277 1±0.031 0	1.172 9±0.039 5	0.230 5±0.004 1	0.734 9±0.031 1	0.261 4±0.007 1	0.211 1±0.005 1
Yeast-cdc	0.260 1±0.029 0	1.080 1±0.061 9	0.235 4±0.004 9	0.595 7±0.036 2	0.257 8±0.006 3	0.215 5±0.004 2
Yeast-elu	0.234 0±0.015 0	1.034 3±0.049 6	0.217 1±0.005 7	0.544 9±0.027 9	0.240 4±0.003 5	0.199 6±0.006 2
Yeast-diau	0.246 0±0.014 0	0.753 1±0.033 3	0.211 0±0.006 2	0.275 3±0.010 4	0.221 5±0.005 2	0.203 7±0.007 5
Yeast-heat	0.198 1±0.007 0	0.683 8±0.038 6	0.193 9±0.007 0	0.232 1±0.013 2	0.201 0±0.004 7	0.184 0±0.005 1
Yeast-spo	0.273 1±0.024 0	0.684 3±0.030 2	0.266 8±0.010 3	0.290 7±0.016 4	0.262 4±0.007 4	0.249 4±0.005 3
Yeast-cold	0.155 2±0.008 0	0.495 4±0.035 1	0.149 7±0.004 7	0.160 9±0.006 9	0.153 2±0.004 9	0.138 7±0.003 2
Yeast-dtt	0.108 0±0.005 0	0.498 6±0.035 8	0.107 5±0.005 1	0.118 3±0.004 2	0.117 2±0.003 9	0.099 7±0.004 0
Yeast-spo5	0.187 0±0.013 0	0.418 0±0.027 8	0.195 8±0.011 3	0.189 5±0.007 3	0.191 4±0.005 1	0.185 4±0.006 2
s-JAFFE	0.457 0±0.039 0	0.439 7±0.015 2	0.348 5±0.030 6	0.522 9±0.048 7	0.415 5±0.025 6	0.414 8±0.020 9
s-BU_3DFE	0.494 0±0.022 0	0.412 5±0.007 0	0.403 1±0.008 7	0.465 4±0.019 9	0.413 8±0.007 2	0.407 1±0.013 2
Movie	0.797 0±0.108 0	0.806 5±0.008 6	0.548 8±0.010 2	0.640 5±0.016 5	0.582 4±0.007 6	0.592 8±0.008 2
算法排位	4.000 0	5.083 3	2.166 7	4.833 3	3.166 7	1.333 3

表 5　Canberra (↓) 指标结果

Table 5　Results in Canberra (↓)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.921 0±0.107 0	4.193 7±0.149 2	0.753 2±0.014 0	2.420 8±0.095 6	0.861 8±0.023 2	**0.684 6±0.014 8**
Yeast-cdc	0.785 0±0.084 0	3.541 2±0.225 8	0.712 1±0.014 7	1.798 6±0.108 4	0.783 1±0.015 7	**0.646 3±0.010 4**
Yeast-elu	0.691 0±0.047 0	3.277 8±0.171 2	0.641 9±0.017 9	1.598 3±0.083 0	0.712 5±0.009 3	**0.585 7±0.016 5**
Yeast-diau	0.528 0±0.031 0	1.708 6±0.084 0	0.453 9±0.015 1	0.594 4±0.020 7	0.478 7±0.010 1	**0.437 7±0.015 5**
Yeast-heat	0.396 0±0.016 0	1.444 0±0.082 7	0.390 3±0.013 5	0.467 1±0.023 6	0.404 2±0.009 8	**0.367 2±0.009 6**
Yeast-spo	0.565 0±0.049 0	1.443 1±0.068 9	0.549 0±0.022 0	0.595 1±0.033 7	0.539 4±0.015 9	**0.513 0±0.010 7**
Yeast-cold	0.267 0±0.014 0	0.866 9±0.062 0	0.259 3±0.007 2	0.277 3±0.012 3	0.264 9±0.008 5	**0.239 0±0.005 3**
Yeast-dtt	0.186 0±0.008 0	0.871 9±0.064 9	0.184 7±0.008 1	0.204 1±0.007 6	0.202 7±0.007 2	**0.171 4±0.006 5**
Yeast-spo5	0.287 0±0.019 0	0.648 8±0.045 0	0.300 3±0.017 0	0.291 1±0.011 2	0.293 9±0.007 7	**0.284 7±0.009 2**
s-JAFFE	0.935 0±0.074 0	0.922 5±0.039 1	0.714 2±0.064 0	1.071 6±0.113 3	**0.862 0±0.058 0**	0.863 0±0.042 0
s-BU_3DFE	1.147 0±0.064 0	0.902 5±0.016 0	**0.831 5±0.019 3**	0.982 3±0.038 5	0.896 9±0.018 0	0.878 6±0.031 6
Movie	1.537 0±0.216 0	1.563 5±0.018 8	**1.055 3±0.021 0**	1.223 8±0.030 5	1.119 9±0.016 4	1.130 8±0.018 0
算法排位	3.416 7	5.666 7	2.166 7	4.833 3	3.000 0	1.416 7

表 6　Kullback-Leibler (↓) 指标结果

Table 6　Results in Kullback–Leibler (↓)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.009 0±0.002 0	0.277 6±0.023 8	0.006 5±0.000 2	0.087 2±0.947 5	0.008 5±0.000 4	**0.005 5±0.000 1**
Yeast-cdc	0.010 0±0.002 0	0.283 1±0.042 1	0.008 2±0.000 3	0.067 2±0.009 8	0.009 9±0.000 5	**0.007 0±0.000 1**
Yeast-elu	0.008 0±0.001 0	0.282 5±0.032 8	0.007 3±0.000 4	0.060 8±0.009 2	0.009 1±0.000 3	**0.006 2±0.000 1**
Yeast-diau	0.019 0±0.002 0	0.269 8±0.027 6	0.014 9±0.000 9	0.026 4±0.002 4	0.015 7±0.000 6	**0.013 5±0.000 1**
Yeast-heat	0.014 8±0.001 0	0.268 4±0.034 5	0.014 4±0.001 0	0.021 6±0.002 9	0.015 2±0.000 7	**0.013 0±0.000 1**
Yeast-spo	0.030 4±0.005 0	0.278 8±0.039 8	0.029 1±0.002 3	0.033 2±0.003 9	0.026 8±0.001 5	**0.024 5±0.001 3**
Yeast-cold	0.014 7±0.001 0	0.217 4±0.035 6	0.013 9±0.001 1	0.016 2±0.001 6	0.014 6±0.001 0	**0.012 1±0.000 1**
Yeast-dtt	0.007 3±0.001 0	0.226 4±0.039 3	0.007 4±0.000 8	0.008 9±0.000 6	0.008 0±0.000 6	**0.006 4±0.000 1**
Yeast-spo5	0.030 1±0.003 0	0.206 6±0.042 7	0.034 7±0.003 9	0.031 2±0.002 3	0.031 4±0.001 5	**0.029 7±0.001 9**
s-JAFFE	0.086 0±0.016 0	0.076 3±0.006 3	**0.053 8±0.010 7**	0.117 3±0.024 5	0.069 3±0.011 6	0.068 7±0.008 5
s-BU_3DFE	0.089 0±0.007 0	0.084 9±0.002 9	0.081 8±0.003 9	0.102 3±0.009 6	0.081 9±0.004 0	**0.080 6±0.005 1**
Movie	0.268 0±0.079 0	0.729 7±0.059 0	**0.117 7±0.005 1**	0.166 4±0.010 7	0.131 7±0.004 6	0.124 0±0.005 1
算法排位	3.916 7	4.916 7	2.333 3	4.916 7	3.250 0	1.166 7

表 7　Cosine (↑) 指标结果

Table 7　Results in Cosine (↑)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.991 0±0.002 0	0.848 5±0.007 0	0.993 6±0.000 2	0.947 5±0.003 2	0.991 4±0.000 4	**0.994 6±0.000 1**
Yeast-cdc	0.991 0±0.002 0	0.850 8±0.013 0	0.992 1±0.000 3	0.956 4±0.004 5	0.990 2±0.000 4	**0.993 3±0.000 1**
Yeast-elu	0.992 0±0.001 0	0.853 1±0.009 2	0.992 9±0.000 4	0.959 7±0.003 9	0.991 0±0.000 3	**0.994 0±0.000 1**
Yeast-diau	0.982 0±0.002 0	0.863 8±0.007 0	0.986 3±0.000 9	0.977 3±0.001 7	0.985 3±0.000 5	**0.987 6±0.000 1**
Yeast-heat	0.986 0±0.001 0	0.866 8±0.010 3	0.986 3±0.000 9	0.980 3±0.002 0	0.985 4±0.000 6	**0.987 7±0.000 1**
Yeast-spo	0.971 0±0.005 0	0.861 1±0.009 3	0.972 7±0.002 2	0.969 5±0.003 2	0.974 7±0.001 3	**0.977 0±0.001 1**
Yeast-cold	0.986 0±0.001 0	0.893 6±0.009 5	0.986 8±0.000 8	0.984 8±0.001 4	0.986 1±0.000 8	**0.988 6±0.000 1**
Yeast-dtt	0.993 0±0.000 5	0.895 0±0.009 8	0.992 9±0.000 6	0.991 6±0.000 5	0.991 5±0.000 5	**0.993 9±0.000 1**
Yeast-spo5	0.973 2±0.003 0	0.898 0±0.010 6	0.969 4±0.003 3	0.972 3±0.001 8	0.972 1±0.001 2	**0.973 8±0.001 4**
s-JAFFE	0.920 0±0.014 0	0.928 1±0.006 3	**0.948 4±0.010 9**	0.899 1±0.018 9	0.934 7±0.011 2	0.935 2±0.008 3
s-BU_3DFE	0.914 0±0.006 0	0.917 9±0.002 8	0.920 2±0.003 5	0.903 7±0.006 9	0.920 3±0.003 6	**0.924 1±0.004 7**
Movie	0.806 0±0.061 0	0.849 5±0.002 2	**0.922 4±0.003 2**	0.902 8±0.004 5	0.908 1±0.003 0	0.919 4±0.003 4
算法排位	4.000 0	5.583 3	2.250 0	4.833 3	3.166 7	1.166 7

表 8　Intersection (↑) 指标结果

Table 8　Results in Intersection (↑)

数据集	PT-SVM	PT-Bayes	AA-KNN	AA-BP	SA-IIS	SC-LDL
Yeast-alpha	0.949 0±0.006 0	0.772 5±0.007 8	0.958 4±0.000 8	0.874 0±0.004 3	0.951 8±0.001 3	**0.962 2±0.000 1**
Yeast-cdc	0.948 0±0.006 0	0.771 1±0.014 6	0.953 1±0.001 0	0.886 7±0.006 3	0.947 8±0.000 9	**0.957 5±0.000 1**
Yeast-elu	0.951 0±0.003 0	0.772 7±0.010 9	0.954 7±0.001 3	0.891 7±0.005 2	0.949 1±0.000 7	**0.958 7±0.001 2**
Yeast-diau	0.926 0±0.004 0	0.769 0±0.010 8	0.937 0±0.002 2	0.917 7±0.002 8	0.933 1±0.001 3	**0.939 3±0.002 0**
Yeast-heat	0.935 0±0.003 0	0.770 8±0.012 2	0.935 9±0.002 2	0.923 5±0.003 5	0.933 2±0.001 6	**0.939 4±0.001 5**
Yeast-spo	0.906 0±0.008 0	0.769 1±0.010 8	0.909 5±0.003 7	0.902 2±0.005 3	0.910 9±0.002 6	**0.915 6±0.001 7**
Yeast-cold	0.933 9±0.004 0	0.796 5±0.013 9	0.936 0±0.001 6	0.931 7±0.003 1	0.934 4±0.002 0	**0.941 1±0.001 3**
Yeast-dtt	0.954 0±0.002 0	0.796 4±0.014 7	0.954 4±0.001 8	0.949 6±0.001 9	0.949 6±0.001 8	**0.957 7±0.001 5**
Yeast-spo5	0.907 1±0.006 0	0.798 7±0.013 3	0.903 0±0.005 4	0.905 8±0.003 5	0.905 0±0.002 3	**0.908 0±0.002 7**
s-JAFFE	0.839 0±0.015 0	0.843 0±0.007 7	**0.876 3±0.013 3**	0.818 1±0.019 5	0.853 6±0.011 9	0.853 3±0.008 3
s-BU_3DFE	0.827 0±0.009 0	0.838 8±0.003 2	**0.847 9±0.003 8**	0.823 6±0.006 7	0.839 4±0.003 7	0.842 3±0.005 9
Movie	0.711 0±0.052 0	0.722 6±0.002 8	**0.822 4±0.003 8**	0.797 6±0.005 1	0.803 2±0.003 3	0.811 6±0.003 6
算法排位	3.333 3	5.583 3	2.083 3	4.916 7	3.250 0	1.333 3

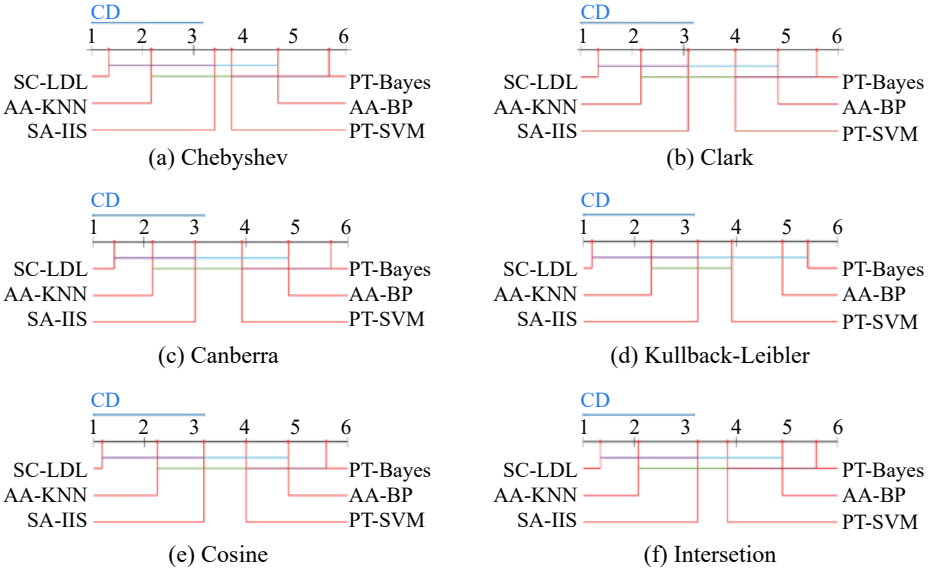

图 1　Nemenyi 检验在 6 个标记分布学习算法上的 CD 图

Fig. 1　CD diagrams of the nemenyi test on six label distribution learning algorithms

实验结果分析：表 2~8 给出了 6 种算法在 12 个标准数据集上的 6 项评价指标结果，图 1 是实验结果的统计分析图。由表 2~8 分析可以得到：

1) 在 6 项指标中，SC-LDL 至少 4 项占优，其它 2 项相差也很小；

2) 本实验选择的数据集中，9 个数据集为酵母菌基因，2 个表情数据集，1 个影评，其中在酵母菌基因数据集中，SC-LDL 效果均最优；

3) 在样本数少，特征数多的数据集 (例如 s-JAFFE) 中，少部分结果略优；

4) 在样本数多，特征数很多的数据集 (Movie) 中，算法效果偏低；

5) 通过综合算法排位分析，SC-LDL 总是总体最优。

图 1 统计分析 Nemenyi 检验结果均在显著性水平为 5% 下得出。Nemenyi 检验是当两个算法的评价指标结果在平均排序中差值不大于临界值 (critical difference, CD)，则说明二者无显著性差异，反之则有显著性差异。该统计分析中有 5 个对比算法，12 个数据集，其中 CD=2.176 7。在图 1 中，线段相连的算法表示二者的平均排序差值低于 CD 值，即无显著性差异，各子图中无显著性差异的算法用线段连接，可以看出，SC-LDL 与大部分算法有显著性差异。这进一步说明 SC-LDL 具有更好的效果。

综上所述，在考虑样本相似性的前提下，将原始样本转化为图模型进行降维，在大多数据集下提升了算法精度。这说明使用谱聚类是一种有效的方式。在大样本高维特征下，由于相似度计算的复杂性及 Movie 数据集的稀疏性，导致算法精度偏低，这也说明了进一步研究降低相似度矩阵构造复杂性的必要性。

4　结束语

谱聚类在图像分割领域取得了较大成就，将谱聚类引入标记分布学习是一个大胆的尝试。实验结果表明是有效的。本文提出的结合谱聚类的标记分布学习算法 SC-LDL，继承了标记分布学习和谱聚类的优点，考虑数据样本之间的联系，对原始数据降维处理，减少了原有标记分布学习计算复杂度。对比现有主流算法，实验结果表明，经过约简的数据建立概率分布模型预测未知样本的标记分布精度更高，假设统计检验也证明了算法的有效性。

虽然谱聚类效果理想，但是在高维样本下计算却变得复杂，如何解决谱聚类在高维下计算复杂问题，将是未来研究重点方向。

参考文献：

[1] ZHOU Zhihua, ZHANG Minling. Multi-label learning [M]//SAMMUT C, WEBB G I. Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer, 2017: 875–881.

[2] 王一宾, 程玉胜, 裴根生. 结合均值漂移的多示例多标记学习改进算法 [J]. 南京大学学报(自然科学版), 2018, 54(2): 422–435.
WANG Yibin, CHENG Yusheng, PEI Gensheng. Improved algorithm for multi-instance multi-label learning based on mean shift[J]. Journal of Nanjing University (Natural Science), 2018, 54(2): 422–435.

[3] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819–1837.

[4] GENG Xin. Label distribution learning[J]. IEEE transactions on knowledge and data engineering, 2016, 28(7): 1734–1748.

[5] 季荣姿. 标记分布学习及其应用 [D]. 南京: 东南大学, 2014.
JI Rongzi. Label distribution learning and its applications [D]. Nanjing: Southeast University, 2014.

[6] GENG Xin, HOU Peng. Pre-release prediction of crowd opinion on movies by label distribution learning[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3511−3517.

[7] GENG Xin, YIN Chao, ZHOU Zhihua. Facial age estimation by learning from label distributions[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(10): 2401–2412.

[8] GENG Xin, XIA Yu. Head pose estimation based on multivariate label distribution[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1837–1842.

[9] 伍育红. 聚类算法综述 [J]. 计算机科学, 2015, 42(6A): 491–499, 524.
WU Yuhong. General overview on clustering algorithms [J]. Computer science, 2015, 42(6A): 491–499, 524.

[10] GENOLINI C, FALISSARD B. KmL: k-means for longitudinal data[J]. Computational statistics, 2010, 25(2): 317–328.

[11] ZHOU Jin, CHEN Long, CHEN C L P, et al. Fuzzy clustering with the entropy of attribute weights[J]. Neurocomputing, 2016, 198: 125–134.

[12] ZHANG Minling. Lift: multi-label learning with label-specific features[C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1609−1614.

[13] 邵东恒, 杨文元, 赵红. 应用 k-means 算法实现标记分布学习 [J]. 智能系统学报, 2017, 12(3): 325–332.
SHAO Dongheng, YANG Wenyuan, ZHAO Hong. Label distribution learning based on k-means algorithm[J]. CAAI transactions on Intelligent systems, 2017, 12(3): 325–332.

[14] 管涛, 杨婷. 谱聚类广义模型与典型算法分析 [J]. 模式识别与人工智能, 2014, 27(11): 1015–1025.
GUAN Tao, YANG Ting. Analysis of general model and classical algorithms for spectral clustering[J]. Pattern recognition and artificial intelligence, 2014, 27(11): 1015–1025.

[15] ZELNIK-MANOR L, PERONA P. Self-tuning spectral clustering[C]//Proceedings of the 17th International Conference on Neural Information Processing Systems. Cambridge, USA, 2004: 1601−1608.

[16] CAI Deng, CHEN Xinlei. Large Scale spectral clustering via landmark-based sparse representation[J]. IEEE transactions on cybernetics, 2015, 45(8): 1669–1680.

[17] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, USA, 2001: 849–856.

[18] YANG Yifang, WANG Yuping, XUE Xingsi. A novel spectral clustering method with superpixels for image segmentation[J]. Optik, 2016, 127(1): 161–167.

[19] WANG Sheng, LU Jianfeng, GU Xingjian, et al. Unsupervised discriminant canonical correlation analysis based on spectral clustering[J]. Neurocomputing, 2016, 171: 425–433.

[20] LI Xinye, GUO Lijie. Constructing affinity matrix in spectral clustering based on neighbor propagation[J]. Neurocomputing, 2012, 97: 125–130.

[21] 赵权, 耿新. 标记分布学习中目标函数的选择 [J]. 计算机科学与探索, 2017, 11(5): 708–719.
ZHAO Quan, GENG Xin. Selection of target function in label distribution learning[J]. Journal of frontiers of computer science and technology, 2017, 11(5): 708–719.

[22] 耿新, 徐宁, 邵瑞枫. 面向标记分布学习的标记增强 [J]. 计算机研究与发展, 2017, 54(6): 1171–1184.
GENG Xin, XU Ning, SHAO Ruifeng. Label enhancement for label distribution learning[J]. Journal of computer research and development, 2017, 54(6): 1171–1184.

作者简介：

王一宾，男，1970 年生，教授，主要研究方向为多标记学习、机器学习和软件安全。发表学术论文 40 余篇。

李田力，男，1996 年生，硕士研究生，主要研究方向为标记分布学习。

程玉胜，男，1969 年生，教授，博士，主要研究方向为数据挖掘、粗糙集。发表学术论文 90 余篇。