DOI:10.11992/tis.201704024
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170704.0925.002.html

应用 k-means 算法实现标记分布学习

邵东恒,杨文元,赵红
(闽南师范大学 粒计算重点实验室,福建 漳州 363000)

摘 要:标记分布学习是近年来提出的一种新的机器学习范式,它能很好地解决某些标记多义性的问题。现有的标记分布学习算法均利用条件概率建立参数模型,但未能充分利用特征和标记间的联系。本文考虑到特征相似的样本所对应的标记分布也应当相似,利用原型聚类的 k 均值算法(k-means),将训练集的样本进行聚类,提出基于 k-means 算法的标记分布学习(label distribution learning based on k-means algorithm,LDLKM)。首先通过聚类算法 k-means 求得每一个簇的均值向量,然后分别求得对应标记分布的均值向量。最后将测试集和训练集的均值向量间的距离作为权重,应用到对测试集标记分布的预测上。在 6 个公开的数据集上进行实验,并与 3 种已有的标记分布学习算法在 5 种评价指标上进行比较,实验结果表明提出的 LDLKM 算法是有效的。

关键词:标记分布;聚类;k-means;闵可夫斯基距离;多标记;权重矩阵;均值向量;softmax 函数
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2017)03-0325-08

中文引用格式: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.
英文引用格式: 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.

Label distribution learning based on k-means algorithm

SHAO Dongheng, YANG Wenyuan, ZHAO Hong
(1. Lab of Granular Computing, Minnan Normal University, Zhangzhou 363000, China)

Abstract: Label distribution learning is a new type of machine learning paradigm that has emerged in recent years. It can solve the problem wherein different relevant labels have different importance. Existing label distribution learning algorithms adopt the parameter model with conditional probability, but they do not adequately exploit the relation between features and labels. In this study, the k-means clustering algorithm, a type of prototype-based clustering, was used to cluster the training set instance since samples having similar features have similar label distribution. Hence, a new algorithm known as label distribution learning based on k-means algorithm (LDLKM) was proposed. It firstly calculated each cluster's mean vector using the k-means algorithm. Then, it got the mean vector of the label distribution corresponding to the training set. Finally, the distance between the mean vectors of the test set and the training set was applied to predict label distribution of the test set as a weight. Experiments were conducted on six public data sets and then compared with three existing label distribution learning algorithms for five types of evaluation measures. The experimental results demonstrate the effectiveness of the proposed KM-LDL algorithm.

Keywords: label distribution; clustering; k-means; Minkowski distance; multi-label; weight matrix; mean vector; softmax function

近年来,标记多义性问题是机器学习和数据挖掘领域的热门问题。目前已有的两种比较成熟的学习范式是对每个实例分配单个标记的单标记学习(single-label learning)和对一个实例分配多个标记的多标记学习(multi-label learning)[1]。多标记学习是对单标记学习的拓展[2]。通常多标记学习能处理一个实例属于多个标记的分歧情况。通过大量的研究和实验[3-5]表明,多标记学习是一种有效且应用范围较广的学习范式。

多标记学习虽然对于一个实例允许标上多个

收稿日期:2017-04-19. 网络出版日期:2017-07-04.
基金项目:国家自然科学基金项目(61379049, 61379089).
通信作者:杨文元. E-mail:yangwy@xmu.edu.cn.

标记,拓展了单标记学习。但是仍有一些问题是不太适合用多标记学习解决的,例如,标记集中的每一个标记描述实例的准确度是多少。事实上,现实世界中有着比大多数人想象的多得多的关于每个标记的准确描述度的数据。在许多科学实验中[6],它们的输出结果不是单个值的,而是一系列的数值输出,例如,基因在不同时间点上的表达水平。这些输出中的单个数值可能不是那么重要,真正重要的是这一系列输出数值的分布情况。如果一个机器学习的任务是要预测一个数值分布,那么它很难放到多标记学习的框架中实现。因为在一个分布中每一个数值输出的准确度是至关重要的,而且这里也不再有相关标记与无关标记的区分了。因此,为了解决这类问题,Geng 等[7]拓展了多标记学习,提出了标记分布学习(label distribution learning, LDL)范式。对于一个特定的实例,标记集合中所有标记的描述度构建一个类似于概率分布的数据形式,称之为标记分布[8],即每个训练实例与一个标记分布相对应。与多标记学习输出一个标记集不同,标记分布学习输出的是一个标记分布,分布中的每个分量表示对应标记对实例的描述程度。事实上,标记分布学习是一种适用场景更广的学习范式,能够解决更多的标记多义性问题。单标记学习和多标记学习都可以看成标记分布学习的特例,相关的研究成果[7, 9-10]也说明了这一点。

目前,已有一些标记分布学习算法[7, 11]被提了出来。这些算法的设计策略主要可以分为以下 3 类。

1)问题转换,即将标记分布学习问题转换成单标记学习问题后,再利用相应范式中已有的算法进行求解,例如:PT-SVM 算法和 PT-Bayes 算法。

2)算法适应,即扩展现存的学习算法来处理标签分布学习问题,例如:LDSVR[12]算法和 AA-BP 算法。

3)专用化的算法,即根据 LDL 的特点设计特殊的算法,例如:SA-IIS 算法、CPNN[13]和 SA-BFGS 算法。

在这 3 种策略中,第 3 种直接针对标记分布学习设计专门算法的效果是最好的。事实上,专用化的算法是通过条件概率或逻辑回归方式训练模型,然后以这个模型预测想要的标记分布。但是在这个过程中算法并未充分考虑训练实例与对应标记分布之间的关系,例如:特征与标记间的函数关系,特征与标记间的分布关系和标记分布数据内部的分布关系。同时,现有的专门算法在处理较大数据集时花费的时间较多。

聚类[14]是研究分类问题的一种统计分析方法,同时也是数据挖掘的一个重要算法,在研究过程中也有许许多多的应用和改进[15]。聚类以相似性为基础,试图将数据集中的样本划分为若干个不相交的子集,每个子集称为一个簇,同一簇中样本之间的相似性比不在同一簇中的更高。在聚类算法中常用的 k-means 算法[16]及改进算法[17]是原型聚类的一种,它假设聚类结构能通过一组原型刻画。通常情况下,算法先对原型进行初始化,然后对原型进行迭代更新求解,直到均值向量不再改变或达到最大迭代次数,此时就能得到每一个簇的均值向量。

在同一个数据集中,特征相近的实例,它们的标记分布往往也相似,同时依据聚类的特性,本文提出一种新的标记分布算法:基于 k-means 的标记分布学习算法(label distribution learning algorithm based on k-means, LDLKM)。首先,利用 k-means 聚类算法求得训练样本集中每个簇的均值向量,此时与每一个训练样本对应的标记分布也相应被划分成簇。然后,求得标记分布的每个簇的均值向量。其次,测试集的样本到各个簇的均值向量的距离矩阵可通过常用的求距离方式,闵可夫斯基距离(Minkowski distance)[18]求得。最后,将距离矩阵通过一个 softmax 函数变换得到一个权重矩阵。权重矩阵和训练样本集的标记分布的均值向量的积就是测试集样本的标记分布,即需要预测的标记分布。本文提出的 LDLKM 算法与现有的专用化的算法相比并未采用条件概率的方式建立模型,而是充分考虑了特征间的分布关系和特征与对应的标记分布之间的联系,利用 k-means 聚类和权重矩阵将特征和标记分布联系到一起。事实上,特征之间的分布与对应标记之间的分布的关系是一种更加直接和强烈的联系。而直接利用这种关系预测得到的标记分布可以继续保持与对应特征的分布关系,从而得到一个较好的结果。LDLKM 和现有的 3 种 LDL 算法在 6 个公开数据集[7]上采用 5 种评价方式进行实验比较,实验的结果表明本文提出的标记分布学习算法在使用的所有数据集上均取得较好的效果,在其中的 5 个数据集上所有评价方式的结果均为最优。

1 标记分布学习的形式化

在标记分布中,标记一个实例 x 的方式是为它的每一个可能的标记 y 分配一个实数 d_x^y,用来表示标记 y 对实例 x 的描述程度。不失一般性,假设实

数 $d_x^y \in [0,1]$,进一步假设所有标记能够完整地描述实例,则 $\sum_y d_x^y = 1$。

标记分布学习是一种更为灵活复杂的标记学习范式,然而学习中更好的灵活性通常意味着更大的输出空间[19]。从单标记学习到多标记学习,再到标记分布学习,学习过程的输出空间的维度变得越来越大。更为详细地,对于标记集合中有 c 个不同标记的学习问题,单标记学习的分类器有 c 个可能的输出,多标记学习的分类器有 2^c-1 个可能的输出。对标记分布学习的分类器来讲,只要在满足 $d_x^y \in [0,1]$ 和 $\sum_y d_x^y = 1$ 这两个约束条件的前提下,它的输出空间的维度可能是无穷大的[19]。

标记分布学习定义的实例矩阵为 $\boldsymbol{X} = [x_1\ x_2 \cdots\ x_n]$,$x_i \in R^d$ 表示第 i 个实例,$i=1,2,\cdots,n$。给定对应标记矩阵 $\boldsymbol{Y} = [y_1\ y_2 \cdots\ y_c]$,$y_j$ 表示第 j 个标记,$j=1,2,\cdots,c$;给定样本的标记分布集是 $\boldsymbol{D} = [\boldsymbol{D}_1\ \boldsymbol{D}_2 \cdots\ \boldsymbol{D}_n]$,$\boldsymbol{D}_i = [d_{x_i}^{y_1}\ d_{x_i}^{y_2} \cdots\ d_{x_i}^{y_c}]$ 表示实例 x_i 的标记分布集,$d_{x_i}^{y_j}$ 表示标记 y_j 对实例 x_i 的描述度。基于此,标记分布学习的一个训练集能够表示为 $S=\{(x_1,D_1),(x_2,D_2),\cdots,(x_n,D_n)\}$。

目前的标记分布学习算法的输出模型是一个最大熵模型[7]:

$$p(y|x;\theta) = \frac{1}{Z}\exp\Big(\sum_j \theta_{y,j} f_j(x)\Big) \tag{1}$$

式中:$Z = \sum_y \exp\Big(\sum_j \theta_{y,j} f_j(x)\Big)$ 是归一化因数;$\theta_{y,j}$ 是模型参数;$f_j(x)$ 是 x 的第 j 个特征。随着标记分布的发展逐渐提出许多基于式(1)的标记分布学习方法[7,9-10]。这些算法先通过条件概率建立参数模型,再利用现有的模型求解参数 θ。

2　基于 k-means 的标记分布学习算法

k-means 算法是所有聚类算法中简单常用的一种算法[20]。它假设聚类结构能被一组原型刻画,先初始化一组原型,然后对原型进行迭代更新求解,最终得到每一个簇的均值向量。给定训练集 $\boldsymbol{H} = [x_1\ x_2 \cdots\ x_m]$,$m$ 为训练集样本数。k-means 算法针对聚类所得簇划分 $G=\{C_1,C_2,\cdots,C_k\}$,最小化平方误差为

$$E = \sum_{i=1}^{k}\sum_{x\in C_i} \| \boldsymbol{x} - \boldsymbol{\mu}_i \|_2^2 \tag{2}$$

式中:$\boldsymbol{\mu}_i = \frac{1}{|C_i|}\sum_{x\in C_i} \boldsymbol{x}$ 是簇 C_i 的均值向量。直观来看,式(2)在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度,E 值越小则簇内样本相似度越高。最小化式(2)并不容易,找到其最优解需考察训练集 $\boldsymbol{H}$ 所有可能的簇划分,这是一个 NP 难[20-21]的问题。因此,k-means 算法采用了贪心策略,通过迭代优化来近似求解式(2)。

首先,聚类过程中在 $\boldsymbol{H}$ 中随机选择 k 个样本作为初始均值向量$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}$,可以通过下式:

$$d_{ji} = \| \boldsymbol{x}_j - \boldsymbol{\mu}_i \|_2 \tag{3}$$

计算训练集样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i(i=1,2,\cdots,k)$ 的距离。根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记:

$$\lambda_j = \arg\min_{i\in[1,2,\cdots,k]} d_{ji} \tag{4}$$

式中 $\lambda_j \in (1,2,\cdots,k)$ 表示样本 $\boldsymbol{x}_j$ 的簇标记。将样本 $\boldsymbol{x}_j$ 划入相应的簇,即

$$C_{\lambda_j} = C_{\lambda_j} \cup [\boldsymbol{x}_j] \tag{5}$$

更新簇 C_{λ_j} 的均值向量。式(3)~式(5)这个过程不断迭代直到当前均值向量保持不变或迭代次数达到所规定的最大次数。

其次,当迭代结束求出所要划分的聚类和对应的均值向量后,便可以依据标记分布与训练样本集的对应关系得到标记分布的簇划分和标记分布每个簇的均值向量 $\boldsymbol{u}$。同时利用常用的距离计算公式“闵可夫斯基距离”公式,即

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \Big(\sum_{u=1}^{n} |\boldsymbol{x}_{iu} - \boldsymbol{x}_{ju}|^p\Big)^{\frac{1}{p}} \tag{6}$$

求得测试集每个样本与各个簇的均值向量的距离矩阵 $\boldsymbol{T}$。闵可夫斯基距离是欧式距离的推广,具有广泛的应用,当 $p=1$ 时是曼哈顿距离,$p=2$ 就是欧式距离,而当 p 趋于无穷大时就是切比雪夫距离。本文中将距离矩阵 $\boldsymbol{T}$ 的每个元素求倒数再通过一个 softmax 函数进行处理转换,从而得到从训练集样本的标记分布的均值向量转化为预测标记分布的权重矩阵。对矩阵 $\boldsymbol{T}$ 作以下处理,先对 $\boldsymbol{T}$ 中每个元素求导数:

$$\boldsymbol{T}'_{ab} = \frac{1}{\boldsymbol{T}_{ab}} \tag{7}$$

然后,为了尽量减小与测试样本实例距离较远的均值向量的影响和便于之后的计算,利用 softmax 函数 $Z_k = \mathrm{softmax}(x_k) = \frac{\exp(x_k)}{\sum_{i=1}^{k}\exp(x_i)}$ 再作以下处理:

$$\boldsymbol{W}_a = \frac{\exp(\boldsymbol{T}'_{ab})}{\sum_{b=1}^{k}\exp(\boldsymbol{T}'_{ab})},\quad a=1,2,\cdots,n \tag{8}$$

式中:n 是测试集样本实例数,$\boldsymbol{W}$ 为最后预测标记分布所使用的权重矩阵。

最后将 $\boldsymbol{W}$ 与训练集对应的标记分布的均值向量矩阵 $\boldsymbol{U}$ 相乘,即

$$\boldsymbol{P} = \boldsymbol{WU} \tag{9}$$

式中:$\boldsymbol{U}=[\boldsymbol{u}_1\ \boldsymbol{u}_2\cdots\ \boldsymbol{u}_b]$;$\boldsymbol{P}$ 就是所需要求的预测标记分布。

上述的算法过程可以通过图 1 的流程图来表示。

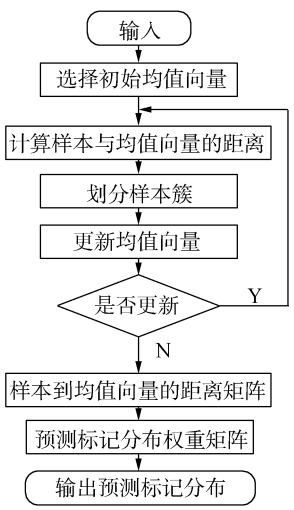


图 1 LDLKM 算法流程图

Fig.1 The flowchart of LDLKM

本文提出的 LDLKM 算法具体步骤如下:

算法 基于 k-means 算法的标记分布学习(LDLKM)。

输入 聚类的簇数 k,聚类迭代的最大次数 d,闵可夫斯基距离参数 p,训练集 $S=\{(\boldsymbol{x}_1,\boldsymbol{D}_1),(\boldsymbol{x}_2,\boldsymbol{D}_2),\cdots,(x_n,\boldsymbol{D}_n)\}$。

输出 测试集的预测标记分布 $\boldsymbol{P}$。

1)从训练集样本 $\boldsymbol{H}$ 中随机选择 k 个样本作为初始向量$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}$。

2)迭代开始,令 $C_i(1\leqslant i\leqslant k)$ 为空,利用式(3)计算样本 $\boldsymbol{x}_j$ 与各均值向量$\boldsymbol{\mu}_i$ 的距离。

3)依据式(4),根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记 $\boldsymbol{\lambda}_j$;将样本 $\boldsymbol{x}_j$ 划入相应的簇。

4)更新均值向量,分别计算划分完簇后的新的均值向量:$\boldsymbol{\mu}_i=\dfrac{1}{|C_i|}\sum\limits_{\boldsymbol{x}\in C_i}\boldsymbol{x}$。

5)若当前的均值向量均未更新或达到规定的最大迭代次数,继续下一步;否则,返回 2),重复 3)到 5)直到所有测法样本划分完毕。

6)依据式(6)求得测试集每个样本与各个均值向量的距离矩阵 $\boldsymbol{T}$。

7)利用式(7)和式(8)求得预测标记分布的权重矩阵 $\boldsymbol{W}$。

8)根据式(9)得出预测标记分布 $\boldsymbol{P}$。

3 实验与结果分析

在这部分,将通过实验来验证本文提出的基于 k-means 的标记分布学习算法。

标记分布学习算法的输出是一个标记分布,与单标记学习的单标记输出和多标记学习的标记集输出都不同。因此,标记分布学习算法的评价方式,应该与单标记学习和多标记学习算法不同。这种方式不是通过预测标记的准确度来评价算法优劣,而是通过测量预测结果和真实标记分布之间的距离或相似度来衡量算法效果。有很多测量概率分布之间的距离或相似度的方法[7]可以用来很好地测量标记分布之间的距离或相似度。例如,表 1 中根据文献[7]和[22]选出的 5 种测量方式就能很好地用来评价标记分布算法。评价标准距离名称之后的"↓"代表距离值越小越好,相似度名称之后的"↑"表示相似值越大越好。这 5 种评价方法分别是切比雪夫距离(Chebyshev)、克拉克距离(Clark)、堪培拉量度(Canberra)、弦系数(Cosine)以及交叉相似性(Intersection),前 3 种以距离作为评价,即越小越好,后两种以相似度作为评价,即越大越好。

表 1 评价指标

Table 1 Evaluation index

评价标准	计算形式
Chebyshev ↓	$\mathrm{Dis}_1=\max\limits_l\lvert d_l-\hat{d}_l\rvert$
Clark ↓	$\mathrm{Dis}_2=\sqrt{\sum_{l=1}^{c}\dfrac{(d_l-\hat{d}_l)^2}{(d_l,\hat{d}_l)^2}}$
Canberra ↓	$\mathrm{Dis}_3=\sum_{l=1}^{c}\dfrac{\lvert d_l-\hat{d}_l\rvert}{d_l+\hat{d}_l}$
Cosine ↑	$\mathrm{Dis}_4=\dfrac{\sum_{l=1}^{c}d_l\hat{d}_l}{\sqrt{\sum_{l=1}^{c}d_l^2}\sqrt{\sum_{l=1}^{c}\hat{d}_l^2}}$
Intersection ↑	$\mathrm{Dis}_5=\sum_{l=1}^{c}\min(d_l,\hat{d}_l)$

3.1 实验设置

通过上述 5 种评价方式,本次实验在 6 个公开的数据集上进行,它们分别是 Yeast-alpha、Yeast-cdc、Yeast-elu、SJAFFE、Human Gene 和 Movie,详细的信息如表 2 所示。

表 2　实验数据集描述

Table 2　Describe experimental data set

数据集	样本	特征	标记
Yeast-alpha	2 465	24	18
Yeast-cdc	2 465	24	15
Yeast-elu	2 465	24	14
SJAFFE	213	243	6
Human Gene	30 542	36	68
Movie	7 755	1 869	5

第 1 个到第 3 个数据集(从 Yeast-alpha 到 Yeast-elu)是从酿酒酵母[6]的 4 个生物实验上收集的真实数据集。每个数据集总共包括2 465个酵母基因,每个基因通过 24 个特征表示。标记对应于离散的时间点,标记分布是每个时间点的基因表达水平。第四个数据集拓展来自一个脸部表情图像数据集 JAFFE,它包括来自 10 个日本女性的 213 张灰度图,并利用局部二值模式[23]从每张图像中采集 243 个特征,每张图像由 60 个人在 6 种感情上打分。第 5 个数据集 Human Gene 是一个大规模的真实数据集,来自于人类基因和疾病的关系生物实验[24]。在数据集中总共包括30 542个人类基因,每一个都被一个基因序列的 36 个特征数值表示。68 个标记对应于 68 种疾病,标记分布是基因在 68 种疾病上的表达水平。第 6 个数据集 Movie 是关于电影的用户评级。评级数据来源于 Netflix,范围是 15 级(5 个标记)。标记分布描述了每个评级所占的比例。特征则提取自电影的元数据,一共有1 869个特征。

为了能使实验结果更具说服力,采用了十折交叉的方式进行实验。聚类划分的簇的数目为 5,最大迭代次数设置为 20,闵可夫斯基距离参数 p 设置成 5。在表 2 的数据集上进行实验,并采用表 1 中的五种评价方式,分别与现有的 3 种标记分布学习算法进行比较。这 3 种比较算法分别是 PT-Bayes、AA-BP 和 SA-IIS。

3.2　实验结果分析

表 3~8 分别列出在 6 个不同的数据集上,4 种算法对应不同评价标准的测量值。前 3 个评价指标(Cheby、Clark 和 Canbe)值越小表示算法效果越好,后两个评价指标(Cosine 和 Interse)值越大表示算法效果越好。在每个表中最后一列是本文算法的结果。从表中可以看出本文提出的算法在 5 种评价标准下都有很好的效果。前 3 个酵母基因数据集和第 5 个人类基因数据集完全优于和它对比的算法,第 4 个和第 6 个数据集也优于其他两个对比算法,并在总体上优于第 3 个对比算法。整体上来看,LDLKM 在基因数据集上可以取得比在其他类型数据集上更好的效果,在非基因数据集 SJAFFE 和 Movie 上的效果略微差于在基因数据集上的效果,而在 Human Gene 数据集上 LDLKM 的效果与 SA-IIS 较为接近,提升效果不大。这说明不同类型的数据集对我们的算法有着一定的影响。同时,可以进一步看到,专用化的算法 SA-IIS 比算法 PT-Bayes 和 AA-BP 的效果更好,处于第二的位置。

表 3　数据集 Yeast-alpha 的实验结果

Table 3　The experimental results of Yeast-alpha

评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.101 0	0.036 1	0.020 2	0.0135
Clark	1.172 8	0.722 1	0.303 4	0.210 1
Canbe	4.198 9	2.383 1	1.014 0	0.681 2
Cosine	0.848 6	0.948 5	0.988 1	0.9946
Interse	0.772 7	0.875 9	0.942 8	0.9624

表 4　数据集 Yeast-cdc 的实验结果

Table 4　The experimental results of Yeast-cdc

评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.1124	0.0393	0.0233	0.016 2
Clark	1.075 8	0.607 3	0.292 6	0.215 8
Canbe	3.526 3	1.829 8	0.897 6	0.6467
Cosine	0.849 1	0.955 0	0.987 0	0.993 3
Interse	0.771 2	0.885 1	0.939 7	0.957 5

表 5　数据集 Yeast-elu 的实验结果

Table 5　The experimental results of Yeast-elu

评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.109 8	0.038 8	0.024 0	0.016 3
Clark	1.014 9	0.543 8	0.275 6	0.199 5
Canbe	3.211 9	1.584 1	0.823 9	0.585 5
Cosine	0.857 2	0.960 0	0.987 6	0.994 0
Interse	0.777 6	0.893 0	0.940 6	0.958 7

表 6 数据集 JAFFE 的实验结果

Table 6 The experimental results of JAFFE

评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.1205	0.1403	0.1191	0.117 9
Clark	0.439 4	0.535 0	0.424 6	0.416 4
Canbe	0.900 9	1.102 2	0.886 3	0.866 0
Cosine	0.930 3	0.895 0	0.931 7	0.9288
Interse	0.846 5	0.842 6	0.849 0	0.850 5

表 7 数据集 Human Gene 的实验结果

Table 7 The experimental results of Human Gene

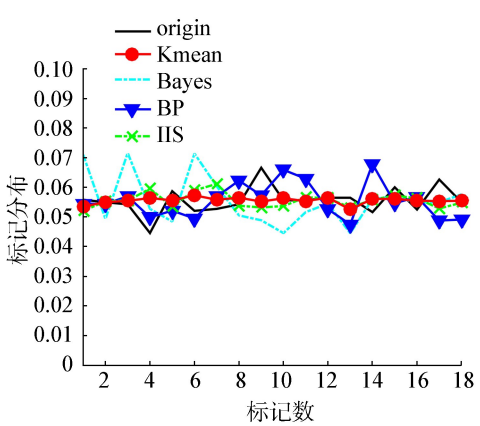
评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.183 4	0.063 0	0.053 4	0.053 3
Clark	4.680 8	3.678 9	2.127 7	2.113 3
Canbe	34.217 8	25.229 7	14.577 8	14.464 2
Cosine	0.452 8	0.690 0	0.832 9	0.834 8
Interse	0.469 6	0.637 7	0.782 4	0.784 4

表 8 数据集 Movie 的实验结果

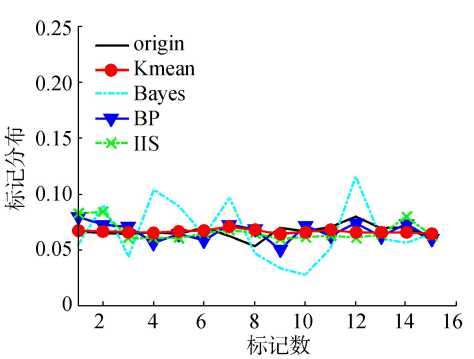
Table 8 The experimental results of Movie

评价标准	PT-Bayes	AA-BP	SA-IIS	LDLKM
Cheby	0.199 2	0.138 5	0.146 7	0.129 3
Clark	0.800 3	0.638 6	0.580 6	0.587 2
Canbe	1.549 0	1.219 1	1.115 8	1.123 8
Cosine	0.850 6	0.903 5	0.908 8	0.918 3
Interse	0.724 5	0.798 9	0.804 1	0.812 4

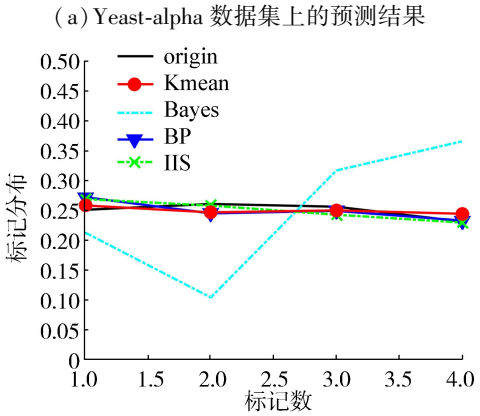
4 种标记分布算法在 6 个数据集上的预测结果如图 2 所示,内容是标记分布算法对数据集中某个实例的标记分布预测结果和实际标记分布的比较。从图 2 中可以看出,LDLKM 的预测结果与实际标记分布最为接近,曲线的形状最为相似,即预测效果最好。在实验过程中,由于 LDLKM 直接利用了特征与标记之间的分布关系,训练模型的时间比现有的专用化的算法还要少。



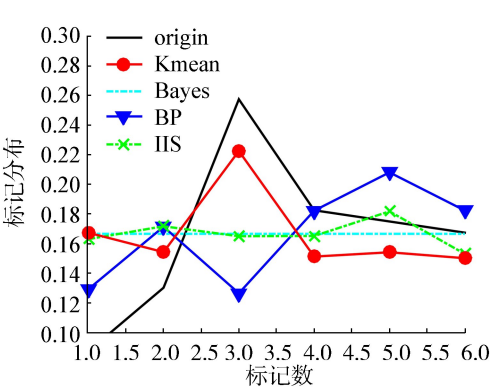
(a) Yeast-alpha 数据集上的预测结果



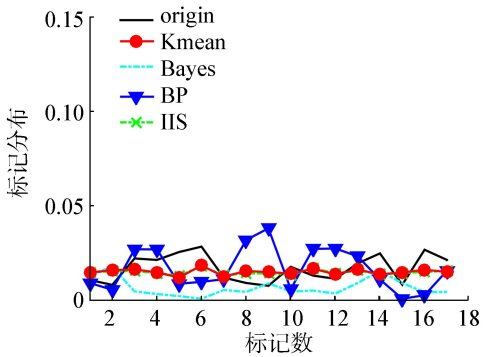
(b) Yeast-cdc 数据集上的预测结果



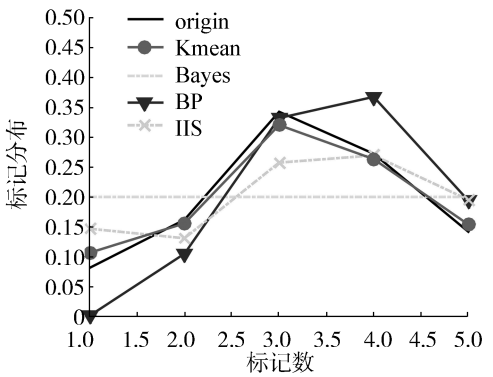
(c) Yeast-elu 数据集上的预测结果



(d) JAFFE 数据集上的预测结果



(e) Human-Gene 数据集上的预测结果



(f) Movie 数据集上的预测结果

图 2　4 种标记分布算法在 6 个公开数据集上预测结果

Fig.2　The prediction distribution of four LDL algorithms on six datasets

4　结束语

本文提出的基于 k-means 标记分布学习算法，是在标记分布框架下，利用标记分布和样本集所具有的联系，通过求得一个权重矩阵来得到预测标记分布，而不是与现有的算法一样，通过求每一个标记的条件概率来得到预测标记分布。LDLKM 主要通过将训练集的样本作为 k-means 聚类的样本，获得每个簇的均值向量。然后将求得的测试集样本与均值向量的距离矩阵，作为预测标记分布与训练集对应的标记分布间的关系，直接求得所需的预测标记分布。本算法充分利用了特征和标记之间的分布关系，又通过 softmax 函数减小了与测试样本距离较远的均值向量的影响，同时本算法相对于现有的专门化的算法在较大的数据集上花费的时间更少。在公开的 6 个数据集上进行的实验所得的结果说明，本文提出的基于 k-means 的标记分布学习算法是有效的。在以后的工作中，我们将对算法进一步优化，还可以引入集成学习来强化聚类效果，或采用一种改进的聚类算法[25]，或针对标记分布学习的特性来专门设计一个聚类算法。

参考文献：

[1] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014, 26(8): 1819-1837.

[2] WEI Yunchao, XIA Wei, HUANG Junshi, et al. CNN: Single-label to multi-label[J]. Computer science, 2014,11: 26-56.

[3] TSOUMAKAS G, KATAKIS I, TANIAR D. Multi-label classification: an overview[J]. International journal of data warehousing and mining, 2007, 3(3): 1-13.

[4] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine learning, 2011, 85(3): 333-359.

[5] READ J, PFAHRINGER B, HOLMES G. Multi-label classification using ensembles of pruned sets [C]// Proceedings of Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008. Washington, USA: IEEE Computer Society, 2008: 995-1000.

[6] EISEN M B, SPELLMAN P T, BROWN P O, et al. Cluster analysis and display of genome-wide expression patterns[J]. Proceedings of the national academy of sciences of the united states of America, 1998, 95(25): 14863-14868.

[7] Geng X. Label distribution learning[J]. IEEE transactions on knowledge and data engineering, 2014, 28(7): 1734-1748.

[8] 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 application [D].Nanjing: Southeast University, 2014.

[9] ZHANG Z, WANG M, GENG X. Crowd counting in public video surveillance by label distribution learning [J]. Neurocomputing, 2015, 166(C): 151-163.

[10] GENG X, WANG Q, XIA Y. Facial age estimation by adaptive label distribution learning[C]// Proceedings of IEEE International Conference on Pattern Recognition, Stockholm, Sweden, 2014. Washington, USA: IEEE Computer Society, 2014: 4465-4470.

[11] GENG X, XIA Y. Head pose estimation based on multivariate label distribution[C]//Proceedings of IEEE International Conference on Computer Vision and Pattern

Recognition, Columbus, USA, 2014. Washington, USA: IEEE Computer Society, 2014:1837-1842.

[12] GENG X, HOU P. Pre-release prediction of crowd opinion on movies by label distribution learning[C]// Proceedings of the International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015. San Francisco, USA:Morgan Kaufmann, 2015: 3511-3517.

[13] GENG X, YIN C, ZHOU Z H. Facial age estimation by learning from label distributions.[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(10): 2401-2412.

[14] JAIN A K. Data clustering: a review[J]. ACM computing surveys, 1999, 31(3): 264-323.

[15] 程旸, 王士同. 基于局部保留投影的多可选聚类发掘算法[J]. 智能系统学报, 2016, 11(5): 600-607.
CHENG Yang, WANG Shitong. A multiple alternative clusterings mining algorithm using locality preserving projections[J]. CAAI transactions on intelligent systems, 2016, 11(5): 600-607.

[16] HARTIGAN J A, WONG M A. A k-means clustering algorithm[J]. Applied statistics, 2013, 28(1): 100-108.

[17] 申彦, 朱玉全. CMP 上基于数据集划分的 k-means 多核优化算法[J]. 智能系统学报, 2015(4):607-614.
SHEN Yan, ZHU Yuquan. An optimized algorithm of k-means based on data set partition on CMP systems[J]. CAAI transactions on intelligent systems, 2015, 10(4): 607-614.

[18] GROENEN P J F, KAYMAK U, VAN Rosmalen J. Fuzzy clustering with minkowski distance functions[J]. Fuzzy sets and systems, 2001, 120(2): 227-237.

[19] 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): 1-12.
ZHAO Quan, GENG Xin. Selection of target function in label distribution learning[J]. Journal of frontiers of computer science and technology, 2017,11(5): 1-12.

[20] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.

[21] ALOISE D, DESHPANDE A, HANSEN P, et al. NP-hardness of euclidean sum-of-squares clustering[J]. Machine learning, 2009, 75(2): 245-248.

[22] CHA S H. Comprehensive survey on distance/similarity measures between probability density functions[J]. International journal of mathematical models and methods in applied sciences, 2007, 1(4): 300-307.

[23] AHONEN T, HADID A, PIETIKÄINEN M. Face description with local binary patterns: application to face recognition[J]. IEEE trans pattern anal mach intell, 2006, 28(12): 2037-2041.

[24] YU J F, JIANG D K, XIAO K, et al. Discriminate the falsely predicted protein-coding genes in Aeropyrum Pernix K1 genome based on graphical representation[J]. Match communications in mathematical and in computer chemistry, 2012, 67(3): 845-866.

[25] 周治平, 王杰锋, 朱书伟,等. 一种改进的自适应快速 AF-DBSCAN 聚类算法[J]. 智能系统学报, 2016, 11(1):93-98.
ZHOU Zhiping, WANG Jiefeng, ZHU Shuwei, et al. An improved adaptive and fast AF-DBSCAN clustering algorithm[J]. CAAI transaction on intelligent systems, 2016, 11(1): 93-98.

作者简介:

邵东恒,男,1992 年生,硕士研究生,主要研究方向为标记分布学习。

杨文元,男,1967 年生,副教授,博士,主要研究方向为机器学习、标记分布学习。发表学术论文 20 余篇。

赵红,女,1979 年生,副教授,主要研究方向为粒计算、分层分类学习。发表学术论文 40 余篇。