第 41 卷 第 1 期 计 算 机 工 程 2015 年 1 月
Vol. 41 No. 1 Computer Engineering January 2015

·人工智能及识别技术· 文章编号:1000-3428(2015)01-0207-04 文献标识码:A 中图分类号:TP18

基于距离度量学习的集成谱聚类

牛 科,张小琴,贾郭军

(山西师范大学数学与计算机科学学院,山西 临汾 041004)

摘 要:无监督学习聚类算法的性能依赖于用户在输入数据集上指定的距离度量,该距离度量直接影响数据样本之间的相似性计算,因此,不同的距离度量往往对数据集的聚类结果具有重要的影响。针对谱聚类算法中距离度量的选取问题,提出一种基于边信息距离度量学习的谱聚类算法。该算法利用数据集本身蕴涵的边信息,即在数据集中抽样产生的若干数据样本之间是否具有相似性的信息,进行距离度量学习,将学习所得的距离度量准则应用于谱聚类算法的相似度计算函数,并据此构造相似度矩阵。通过在 UCI 标准数据集上的实验进行分析,结果表明,与标准谱聚类算法相比,该算法的预测精度得到明显提高。

关键词:数据挖掘;边信息;相似度矩阵;距离度量学习;谱聚类;UCI 数据集

中文引用格式:牛 科,张小琴,贾郭军. 基于距离度量学习的集成谱聚类[J]. 计算机工程,2015,41(1):207-210.
英文引用格式:Niu Ke,Zhang Xiaoqin,Jia Guojun. Integrated Spectral Clustering Based on Distance Metric Learning[J]. Computer Engineering,2015,41(1):207-210.

Integrated Spectral Clustering Based on Distance Metric Learning

NIU Ke,ZHANG Xiaoqin,JIA Guojun

(School of Mathematics and Computer Science,Shanxi Normal University,Linfen 041004,China)

【Abstract】 The performance of the unsupervised learning clustering algorithm is critically dependent on the distance metric being given by a user over the inputs of the data set. The calculation of the similarity between the data samples lies on the specified metric,therefore,the distance metric has a significant influence to the results of the clustering algorithm. Aiming at the problem of the selection of the distance metric for the spectral clustering algorithm,a spectral clustering algorithm based on distance metric learning with side-information is presented. The algorithm learns a distance metric with the side-information. The similarity between the data samples is chosen randomly from the data set,and is applied to the similarity function of spectral clustering algorithm. It structures the similarity matrix of the algorithm. The effectiveness of the algorithm is verified on real standard data sets on UCI,and experimental results show that compared with the standard spectral clustering algorithms,the prediction accuracy of the proposed algorithm is improved significantly.

【Key words】 data mining;side-information;similarity matrix;distance metric learning;spectral clustering;UCI data set
DOI:10.3969/j.issn.1000-3428.2015.01.038

1 概述

聚类是数据挖掘技术的一种重要手段,旨在将集合中的数据对象分组成不同的类别,使不同的类别之间呈现出高内聚低耦合的特点。

事实上,聚类是一种无监督分类,它没有任何先验知识可用[1],因此,数据对象之间的距离度量对聚类算法的表现性能具有重要影响。合理的距离度量能够准确地反映出数据集中数据对象之间的相互关系,从而提升聚类算法的表现性能,使聚类结果更合理客观的呈现出数据集中数据对象的所属类别及其之间的相互关系。

谱聚类是建立在谱图理论基础上的一种聚类算法。与传统的聚类算法相比,谱聚类能够在任意形状的样本空间上进行聚类且收敛于全局最优解[2],并且使用线性代数基本知识即可实现。

文献[3]提出利用抽样数据样本的相似性作为边信息进行距离度量学习的方法,从而使距离度量能够更加合理地刻画出该数据集中数据样本之间的相互关系,并将其应用 k-means 和 Comstrained k-

基金项目:山西省软科学基金资助项目(2009041052-03)。
作者简介:牛 科(1987-),男,硕士研究生,主研方向:智能计算,软件工程;张小琴,硕士研究生;贾郭军,副教授。
收稿日期:2013-10-31 **修回日期**:2014-03-03 **E-mail**:niuke870505@163.com

means 聚类算法中,使算法的预测精度得到了显著提高。

本文利用数据集抽样样本的相似性作为边信息进行距离度量学习,并且将学习所得的距离度量应用于谱聚类算法。

2 基于边信息的距离度量学习

聚类是一种无监督的机器学习算法,所谓的相似性边信息也就是从数据集中随机抽取出来的一些数据样本以及数据样本之间所属类别的相互关系。

设现有数据集 $X=\{x_i\}\subseteq R^n, i=1,2,\cdots,n$,从 X 中随机取出一些数据样本,并确认数据样本 x_i 与 x_j 是否具有某种相似性,即数据样本 x_i 与 x_j 否属于同一类别。若 x_i 与 x_j 属于同一类别,则 $(x_i,x_j)\in S$,否则 $(x_i,x_j)\in DS$。对于采集到的边信息,本文使用矩阵 $\boldsymbol{SC},\boldsymbol{DC}$ 进行描述如下:

$$\boldsymbol{SC}=\boldsymbol{SC}(i,j)=\begin{cases}1 & \text{if } (x_i,x_j)\in S\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

$$\boldsymbol{DC}=\boldsymbol{DC}(i,j)=\begin{cases}0 & \text{if } (x_i,x_j)\in DS\\ 1 & \text{otherwise}\end{cases} \tag{2}$$

其中,$i,j=1,2,\cdots,N$,N 为随机抽取的数据样本个数。

假设利用抽样样本相似性边信息进行学习所得距离度量具有如下所示的马氏距离[4-5]形式:

$$d(x,y)=d_{\boldsymbol{A}}(x,y)=\|x-y\|_{\boldsymbol{A}}=\sqrt{(x-y)^{\mathrm{T}}\boldsymbol{A}(x-y)} \tag{3}$$

其中,为了保证 $d(x,y)$ 能够满足非负性和三角不等式,$\boldsymbol{A}$ 必须是一个半正定的矩阵。

若 $\boldsymbol{A}=\boldsymbol{I}$,则 $d(x,y)$ 即为标准欧氏距离;若 $\boldsymbol{A}$ 是一个对角阵,则相当于对数据样本的每一个维度赋予了不同的权值。更广泛地说,$d(x,y)$ 是 R^n 上的一个使用矩阵 $\boldsymbol{A}$ 参数化了的马氏距离。

文献[3]提出,求解矩阵 $\boldsymbol{A}$ 的一种可行方法就是规定集合 S 中的数据点对 (x_i,x_j) 之间具有最小的距离和。而集合 DS 中的数据点对 (x_i,x_j) 之间的距离大于等于 1 且矩阵 $\boldsymbol{A}$ 满足半正定。因此,可以通过求解如下凸优化问题,得到矩阵 $\boldsymbol{A}$:

$$\min_{\boldsymbol{A}}\sum_{(x_i,x_j)\in S}\|x_i-x_j\|_{\boldsymbol{A}}^2 \tag{4}$$

$$\text{s.t.}\sum_{(x_i,x_j)\in DS}\|x_i-x_j\|_{\boldsymbol{A}}\geqslant 1 \tag{5}$$

$$\boldsymbol{A}\geqslant 0 \tag{6}$$

此时,可以使用 Newton-Raphson 方法,定义:

$$g(\boldsymbol{A})=g(A_{11},A_{22},\cdots,A_{nn})=\sum_{(x_i,x_j)\in S}\|x_i-x_j\|_{\boldsymbol{A}}^2-\lg\left(\sum_{(x_i,x_j)\in DS}\|x_i-x_j\|_{\boldsymbol{A}}\right) \tag{7}$$

在约束条件 $\boldsymbol{A}\geqslant 0$ 下最小化 g 对式(7)进行求解,即可以得到一个对角矩阵 $\boldsymbol{A}$。

若使学习所得的矩阵 $\boldsymbol{A}$ 为全矩阵,可以使用梯度下降和迭代预测算法,求解如下与式(4)~式(6)等价的凸优化问题,即可得到全矩阵 $\boldsymbol{A}$:

$$\max_{\boldsymbol{A}}g(\boldsymbol{A})=\sum_{(x_i,x_j)\in DS}\|x_i-x_j\|_{\boldsymbol{A}} \tag{8}$$

$$\text{s.t. } f(\boldsymbol{A})=\sum_{(x_i,x_j)\in S}\|x_i-x_j\|_{\boldsymbol{A}}^2\leqslant 1 \tag{9}$$

$$\boldsymbol{A}\geqslant 0 \tag{10}$$

梯度下降和迭代预测算法[3]如下:

```
Iterate
  Iterate
    A: = arg min_A' {||A' - A||_F : A' ∈ C_1}
    A: = arg min_A' {||A' - A||_F : A' ∈ C_2}
    Until A converges
    A: = A + α(∇_A g(A)) ⊥ ∇_f f
    Until  convergenc
```

其中,$\|M\|_F$ 表示 M 的 F 范数($\|M\|_F=(\sum_i\sum_j M_{ij}^2)^{1/2}$)。

本文实验的距离度量学习采用 Newton-Raphson 方法,所得的矩阵 $\boldsymbol{A}$ 为对角矩阵。

3 谱聚类算法

谱聚类是近年来机器学习领域的一个新的研究热点,基于谱图理论的谱聚类已逐渐成为最为广泛使用的聚类算法之一。在计算机科学、统计学、物理学等领域越来越受到人们的关注[5]。与传统的 k-means 等聚类算法相比,谱聚类在复杂形状样本空间聚类中表现出了良好的性能。

标准谱聚类算法的实现主要依赖于数据集中数据样本的相似度矩阵 $\boldsymbol{S}=\boldsymbol{R}^{n\times n}$、连接矩阵 $\boldsymbol{W}=\boldsymbol{R}^{n\times n}$和度矩阵 $\boldsymbol{D}=\boldsymbol{R}^{n\times n}$。

在谱聚类中通常采用高斯函数计算数据点之间的相似度[6]:

$$\boldsymbol{S}=(s_{ij})=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)$$

$$i,j=1,2,\cdots,n \tag{11}$$

通常将相似度矩阵 $\boldsymbol{S}$ 采用 ξ-近邻、k-近邻、全连通 3 种方式进行稀疏化处理即可得到连接矩阵 $\boldsymbol{W}$,$w_{ij}\geqslant 0, i,j=1,2,\cdots,n$ 且 $w_{ij}=w_{ji}$。

度矩阵 $\boldsymbol{D}$ 则由下式计算所得:

$$\boldsymbol{D}=(d_i)=\sum_{j=1}^{n}w_{ij}, i=1,2,\cdots,n \tag{12}$$

由连接矩阵 $\boldsymbol{W}$ 和度矩阵 $\boldsymbol{D}$ 可以得到顶点集的拉普拉斯矩阵[7]。拉普拉斯矩阵分为非归一化和归一化 2 种。非归一化的拉普拉斯矩阵计算公式如下:

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W} \tag{13}$$

归一化的拉普拉斯矩阵计算公式如下:

$$\boldsymbol{L}_{\text{sym}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \quad (14)$$

$$\boldsymbol{L}_{\text{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W} \quad (15)$$

其中,式(14)和式(15)中的 $\boldsymbol{L}$ 即为式(13)中的非归一化拉普拉斯矩阵。$\boldsymbol{L}_{\text{sym}}$ 是对称矩阵,$\boldsymbol{L}_{\text{rw}}$ 是一个随机游走矩阵,通常是非对称的[6]。

传统谱聚类算法就是在构建的拉普拉斯矩阵中,根据聚类个数 k,求解其前 k 个特征值以及与其对应的特征向量并构建特征向量空间。然后采用 k-means 算法对特征向量空间中的特征向量进行聚类[8]。算法步骤描述如下:

输入　数据集 $X=\{x_i\}\subseteq R^n, i=1,2,\cdots,n$

输出　聚类结果 $C_1, C_2, \cdots, C_k$

Step1　构造基于样本间相似度的相似度图,并计算加权连接矩阵 $\boldsymbol{W}$,度矩阵 $\boldsymbol{D}$。

Step2　计算拉普拉斯矩阵 $\boldsymbol{L}$(依据需要解决的实际应用问题采用非归一化的拉普拉斯矩阵或者归一化的拉普拉斯矩阵 $\boldsymbol{L}_{\text{sym}}$ 或者 $\boldsymbol{L}_{\text{rw}}$)。

Step3　计算拉普拉斯矩阵 $\boldsymbol{L}$ 的前 k 个特征值及其对应的特征向量 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$($k$ 为需要将数据集进行聚类的个数)。

Step4　采用经典的 k-means[9-10] 聚类算法对特征向量空间的特征向量进行聚类,得到聚类结果 $C_1, C_2, \cdots, C_k$。

4　基于距离度量学习的谱聚类算法

在基于相似性边信息进行距离度量学习的谱聚类算法中,将学习所得的距离度量 $d(x,y)$ 应用于谱聚类算法的相似度计算函数,构造数据集的相似度矩阵 $\boldsymbol{S}$。数据样本 x_i 和 x_j 的相似度采用如下公式进行计算:

$$\boldsymbol{S} = (s_{ij}) = \exp\left(-\frac{d(x_i,x_j)}{2\sigma^2}\right), \quad i,j = 1,2,\cdots,n \quad (16)$$

文献[3]提出学习这样的一个距离度量等价于寻找数据样本的一种缩放尺度,即使用 $\boldsymbol{A}^{1/2}x$ 替代数据样本 x,然后再在缩放后的数据集上使用标准的欧氏距离作为距离度量。

基于边信息进行距离度量学习的谱聚类算法步骤如下:

输入　数据集 $X=\{x_i\}\subseteq R^n, i=1,2,\cdots,n$

输出　聚类结果 $C_1, C_2, \cdots, C_k$

Step1　从数据集 $X=\{x_i\}\subseteq R^n, i=1,2,\cdots,n$ 中随机抽取 $N(N<n)$ 个样本,记为样本 $Y=\{y_i\}\subseteq R^n$, $i=1,2,\cdots,N$。

Step2　通过识别判断集合 Y 中的数据点(y_i, y_j), $i,j=1,2,\cdots,N$ 是否属于同一类别,从而得到相似性边信息矩阵 $\boldsymbol{SC}$,$\boldsymbol{DC}$。

Step3　利用相似性边信息进行距离度量学习,得到距离度量公式 $d(x,y)$。

Step4　利用学习所得的距离度量公式,计算数据点之间的相似度矩阵 $\boldsymbol{S}$。

Step5　计算连接权矩阵 $\boldsymbol{W}$ 和度矩阵 $\boldsymbol{D}$。

Step6　计算拉普拉斯矩阵 $\boldsymbol{L}$(依据需要解决的实际应用问题采用非归一化的拉普拉斯矩阵或者归一化的拉普拉斯矩阵 $\boldsymbol{L}_{\text{sym}}$ 或者 $\boldsymbol{L}_{\text{rw}}$)。

Step7　计算拉普拉斯矩阵 $\boldsymbol{L}$ 的前 k 个特征值及其对应的特征向量 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$($k$ 为需要将数据集进行聚类的个数)。

Step8　采用经典的 k-means[9-10] 聚类算法对特征向量空间的特征向量进行聚类,得到聚类结果 C_1, $C_2, \cdots, C_k$。

5　实验结果与分析

5.1　实验平台

本文所有实验都是在 1 台 PC 机(Intel Core i3 CPU,主频 2.40 GHz,内存 2.0 GB)上进行。采用 Windows XP SP3 操作系统,Matlab R2009a 开发平台。

5.2　结果分析

为了验证本文提出的使用数据集抽样样本相似性边信息进行距离度量学习的谱聚类算法的有效性,分别在 5 个 UCI[11] 标准数据集上进行了实验,并将聚类结果与标准的谱聚类算法进行了比较。

各个数据集的属性如表 1 所示。

表 1　各个 UCI 标准数据集的属性

数据集	类别数	样本维数	样本数
soy bean	4	35	47
Iris	3	4	150
wine	3	12	178
heart	2	13	256
boston housing	3	13	506

在实验中采用聚类精度作为评价指标。分别在各个数据集中随机抽取 5 个、10 个、15 个、20 个、25 个、30 个数据样本,并判断各样本的所属类别的相互关系作为相似性边信息进行距离度量学习。实验使用 Hungarian 算法[12] 进行聚类精度计算。

各数据集实验结果如表 2 所示。

表 2 UCI 数据集实验结果对比

数据集	标准谱聚类结果	基于距离度量学习的谱聚类结果					
		5 个样本	10 个样本	15 个样本	20 个样本	25 个样本	30 个样本
soy bean	0.617 0	0.744 7	0.766 0	0.787 2	0.767 8	0.784 5	0.795 7
Iris	0.692 1	0.815 2	0.923 6	0.954 5	0.927 3	0.900 0	0.926 1
wine	0.623 1	0.634 3	0.656 3	0.677 7	0.705 3	0.654 2	0.694 1
heart	0.551 9	0.639 7	0.703 0	0.720 9	0.697 3	0.752 9	0.743 8
boston housing	0.721 2	0.752 4	0.773 6	0.769 0	0.786 4	0.796 1	0.790 5

图 1 ~ 图 5 为各个 UCI 标准数据集实验对比结果。

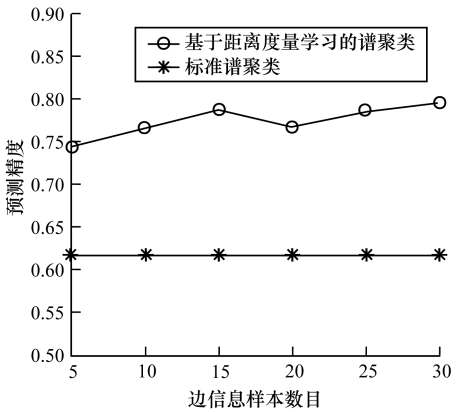

图 1 数据集 soy bean 实验结果对比

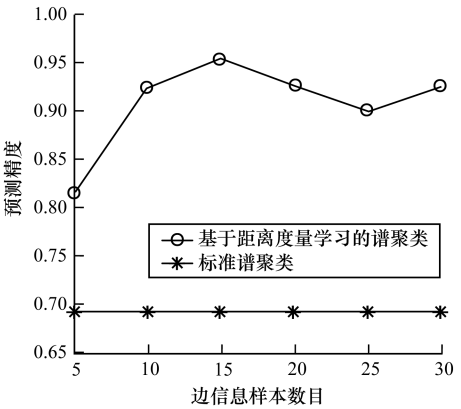

图 2 数据集 Iris 实验结果对比

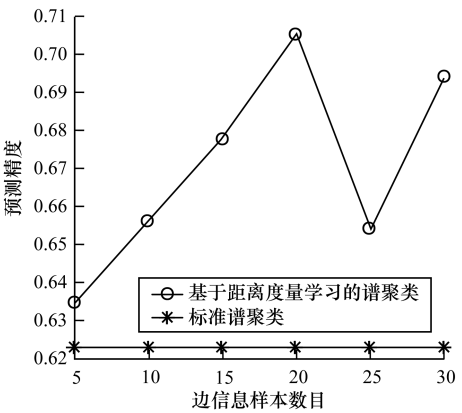

图 3 数据集 wine 实验结果对比

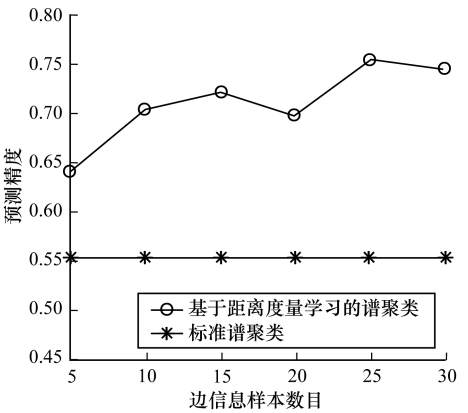

图 4 数据集 heart 实验结果对比

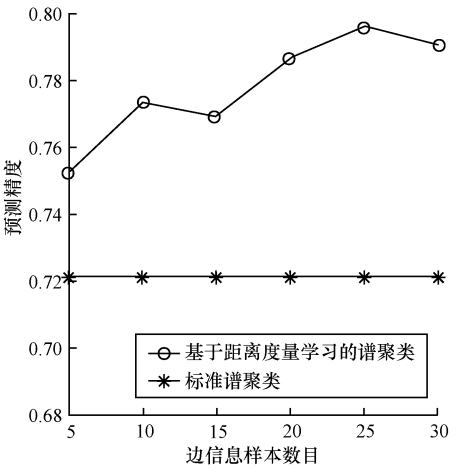

图 5 数据集 boston housing 实验结果对比

通过对比分析可得,与标准谱聚类算法相比,利用数据集抽样样本相似性边信息进行距离度量学习的谱聚类算法的聚类性能得到了显著提高。然而,随着数据集相似性边信息的增多,聚类的性能并不会成正比例上升。

6 结束语

本文提出一种利用数据集抽样样本相似性边信息进行距离度量学习的谱聚类算法。实验结果表明,与标准谱聚类算法相比,该算法的聚类性能得到了明显改善。在执行谱聚类算法时,高斯相似度函数中的参数 σ 选取的是否合理,也会对聚类结果产生重要的影响。因此,参数 σ 的选取是下一步需要研究的工作。

(下转第 244 页)