DOI:10.11992/tis.201508022
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160824.0928.004.html

基于局部保留投影的多可选聚类发掘算法

程旸,王士同
(江南大学 数字媒体学院,江苏 无锡 214122)

摘　要:绝大多数的聚类分析算法仅能得到单一的聚类结果,考虑到数据的复杂程度普遍较高,以及看待数据的视角不同,所得到的聚类结果在保证其合理性的基础上应当是不唯一的,针对此问题,提出了一个新的算法 RLPP,用于发掘多种可供选择的聚类结果。RLPP 的目标函数兼顾了聚类质量和相异性两大要素,采用子空间流形学习技术,通过新的子空间不断生成多种互不相同的聚类结果。RLPP 同时适用于线性以及非线性的数据集。实验表明,RLPP 成功地发掘了多种可供选择的聚类结果,其性能相当或优于现有的算法。

关键词:可供选择的聚类结果;无监督学习;流形学习;多聚类;特征分解
中图分类号:TP18　　**文献标志码**:A　　**文章编号**:1673-4785(2016)05-0600-08

中文引用格式: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.
英文引用格式: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.

A multiple alternative clusterings mining algorithm using locality preserving projections

CHENG Yang, WANG Shitong
(School of Digit Media, Jiangnan University, Wuxi 214122, China)

Abstract: Most clustering algorithms typically find just one single result for the data inputted. Considering that the complexity of the data is generally high, combined with the need to allow the data to be viewed from different perspectives (on the basis of ensuring reasonableness), means that clustering results are often not unique. We present a new algorithm RLPP for an alternative clustering generation method. The objective of RLPP is to find a balance between clustering quality and dissimilarity using a subspace manifold learning technique in a new subspace so that a variety of clustering results can be generated. Experimental results using both linear and nonlinear datasets show that RLPP successfully provides a variety of alternative clustering results, and is able to outperform or at least match a range of existing methods.

Keywords: alternative clustering; unsupervised learning; manifold learning; multiple clusterings; eigendecomposition

大多数传统的聚类算法仅仅能得到单个结果,但是当对复杂数据进行聚类分析时,很可能存在多个具有合理性的聚类结果。这一特点在高维数据上表现得尤为明显,例如文本、图像、基因数据等,这些数据具有多种特征,而不同的特征子空间往往会得到完全不同的聚类结果,同时每一种结果都能体现数据不同的结构信息。

本文根据文献[1]所述原理,提出了一种能够发掘多个可供选择的聚类结果的算法 RLPP。算法结合了希尔伯特施密特独立性度量准则(hilbert-schmidt independence criterion, HSIC)[2]以及局部保持投影(locality preserving projections, LPP)[3],改进了 LPP 算法学习子空间的过程。由于 HSIC 可以高效地评估不同随机变量之间的依赖性,而 LPP 算法具有流形学习能力,因此 RLPP 同时兼顾了聚类结果的相异性和聚类质量这两大要素。并且由于其目标函数最终在特征分解问题的框架内求解,因此能

收稿日期:2015-08-26.　网络出版日期:2016-08-24.
基金项目:国家自然科学基金项目(61272210).
通信作者:程旸. E-mail:szhchengyang@163.com.

够确保求出的新的子空间一定存在,并且解是全局最优的。

总的来说,本文所做的工作为:1)提出了一种新的算法 RLPP,用于发掘多种可供选择的聚类结果;2)RLPP 根据同时满足质量和相异性要求的目标函数,生成一个新的特征子空间,该特征子空间能够确保存在,并且是全局最优的;3)通过实验,验证了 RLPP 的效果,并与其他现有的算法进行了性能比较。

1　当前典型的可选聚类发掘方法

当前,有关发掘可选聚类结果的算法大体上可以分为两类:一类直接利用原始数据空间寻找,另一类则是基于投影(变形)子空间寻找。

1.1　基于全部原始数据空间

这类研究利用的是整个原始特征空间,大多数研究的不同之处在于优化聚类质量和相异性的目标函数不同。文献[4-9]中的研究可以归类为此类。文献[4]提出了一种分层聚类(hierarchical clustering)算法 COALA,该算法把从提供的聚类结果中生成的 cannot-link 约束项合并入它的每一个凝聚步骤中,即尽可能多地满足这些 cannot-link 约束项。在文献[7]中,提出了 CAMI 算法,用于同时寻找两个可供选择的聚类结果。CAMI 算法在混合模型下构造聚类问题,优化了一个双重目标函数(dual-objective function),使得当两个混合模型之间的互信息(反映了两种聚类结果之间的不同)最小时,对数似然(反映了聚类质量)最大。文献[6]提出的两种算法 Dec-kmeans 和 Conv-EM 也属于此类,这两种算法分别改进了 k-means 和 EM 的目标函数,结合了一个修正项,用于表示两种聚类结果之间的去相关信息。文献[8]中的工作采用了不同的方式,其原理来源于信息论,它的目标函数最大化全部数据实例和可选聚类结果类标之间的互信息(MI),同时最小化可选聚类和所提供的聚类结果之间的互信息。文献[8]中并没有基于传统的香农熵[10],而是采用了 Renyi 熵,以及相对应的二次互信息[11-12],这种方法在结合了非参数 Parzen 窗[13]后使得 MI 基本近似。这种双重优化聚类目标同样被用于文献[9]中,区别在于文献[9]使用的是迭代法,而不是文献[8]中所使用的分层技术。

1.2　基于投影子空间

如果原数据空间的子空间与原数据空间是相互独立的(比如是正交的),那么根据该子空间得到的聚类结果也与原聚类结果不同。文献[14-18]就是根据这样的理论基础提出了各自的算法。文献[14]由正交投影方法提出了两种寻找可供选择的聚类结果的算法。已知一个向量 $\boldsymbol{b}$,投影到矩阵的列空间中,可以用 $\boldsymbol{P}\times\boldsymbol{b}$ 计算,其中 $\boldsymbol{P}$ 被称为投影矩阵,$\boldsymbol{P}=\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$。而$(\mathrm{I}-\boldsymbol{P})$同样也是一个投影矩阵,表示把投影到了$\boldsymbol{A}^{\mathrm{T}}$的零空间中。文献[14]中提出的 2 种算法把每个数据实例看作向量,利用了上述投影等式。文献[15]中的研究也与此相关,投影矩阵被应用于从所提供的聚类结果导出的距离矩阵上。相比于文献[14]中的 2 种算法,这种方法的优势在于能够解决数据维数比类别数小的情况。文献[16]提出的算法采用了不同的方法,通过对数据的投影,使得在参考聚类结果中属于相同类别的数据点经过映射后在新的空间中拉开距离。这一方法与其他方法之间的不同之处在于它并不寻找一个全新的可选聚类,而是通过设定 2 个聚类结果之间的相异度阈值,允许已知的聚类结果中的部分在可选聚类结果中保留下来。文献[17]和文献[18]中所提出的算法基于谱聚类实现,前者表明可选聚类结果可以通过拉普拉斯矩阵不同的特征向量找到,后者所提出的多重谱聚类(multiple spectral clustering, MSC)把子空间学习技术融入了谱聚类的过程中,也就是说,MSC 的目标函数是一个对偶函数(dual-function),通过最优化一项来修正另一项。另外,文献[1]提出了正则化 PCA(regularized PCA, RPCA)和正则化的图方法(regularized graph-based method, RegGB)算法,其中 RPCA 与 MSC 一样,都采用了 HSIC,用于评估相关性,而 RegGB 算法则是基于图论构造。总的来说,RPCA 和 RegGB 算法在寻找可选聚类的能力上要优于之前所提到的算法,但是 RPCA 算法只适用于线性结构的数据集,并且其寻找可选聚类结果的能力有限,往往只能找到一个可选聚类,这些都极大地影响了它在使用上的灵活性。因此,本文在文献[1]所提出的思路上,探索了一种新的算法,通过引入流形学习大大提高了其发掘低维流形结构的能力和子空间学习能力,并通过核化扩大了其适用范围,使得其既适用于线性,同时又适用于非线性的数据集。

2　问题描述

假设数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\boldsymbol{x}_n\}$,$\boldsymbol{x}_i\in\mathbf{R}^d$,即 $\boldsymbol{X}$ 是 $d\times n$的矩阵,并提供一个使用任意聚类算法得到的参考聚类结果 $C^{(1)}$。则本文研究的目标为:发掘数据集 $\boldsymbol{X}$ 上的可供选择的聚类结果 $C^{(2)}$,并且 $C^{(2)}$ 中的所有类别 $C_i^{(2)}$ 必须满足两个条件,$U_iC_i^{(2)}=\boldsymbol{X}$ 和 $C_i^{(2)}\cap C_j^{(2)}=\varnothing(\forall i\neq j)$。除了与 $C^{(1)}$不同外,还要求 $C^{(2)}$的聚类质量较高。同理,若提供一组参考聚类结果$\{C^{(1)},C^{(2)},\cdots\}$,必须生成高质量的可供选择的聚类结果 $C^{(k)}$,且与之前所有的聚类结果$\{C^{(1)},$

$C^{(2)},\cdots\}$不同。

为了发掘另一个可供选择的聚类结果,使用子空间流形学习方法,将原始数据空间 $\boldsymbol{X}$ 映射到一个新的子空间中。该空间保留了 $\boldsymbol{X}$ 的特征,并且完全独立于其他的参考聚类结果。任何聚类算法都可以使用这个新的子空间进行聚类分析。

3 局部保持投影

局部保持投影(locality preserving projections, LPP)[3]是一种非监督降维方法,是流形学习算法 Laplacian Eigenmap 的线性逼近。给定 R^d 中的 n 个数据点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$,LPP 通过寻找转换矩阵 $\boldsymbol{A}$,将这 n 个数据点映射为 $R^l(l\ll d)$ 上的数据点 $\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$,即:

$$\boldsymbol{y}_i=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i,i=1,2,\cdots,n \tag{1}$$

式中所需的转换矩阵 $\boldsymbol{A}$ 可以通过最小化式(2)目标函数得到:

$$\boldsymbol{A}=\operatorname{argmin}\sum_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij} \tag{2}$$

式中:$\boldsymbol{W}_{ij}$是权值矩阵,可采用 k 最近邻算法得到邻接图,再求出权值矩阵。

如果 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的 k 近邻点,则 $\boldsymbol{W}_{ij}=\exp-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{t}$ $(t\in R)$;否则 $\boldsymbol{W}_{ij}=0$。显然,$\boldsymbol{W}$ 是一个$n\times n$的稀疏对称矩阵。

从目标函数式(2)可以看出,降维后的特征空间可以保持原始高维空间的局部结构。结合式(1)和式(2),做简单的代数变换:

$$\frac{1}{2}\sum_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij}=$$

$$\frac{1}{2}\sum_{ij}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_j)^2\boldsymbol{W}_{ij}=$$

$$\sum_i\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{D}_{ii}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}-\sum_{ij}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{W}_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{A}=$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A} \tag{3}$$

式中:$\boldsymbol{X}=[\boldsymbol{x}_1\,\boldsymbol{x}_2\cdots\boldsymbol{x}_n]$,$\boldsymbol{D}$ 是一个 $n\times n$ 的对角矩阵,对角线元素 $\boldsymbol{D}_{ii}=\sum_j\boldsymbol{W}_{ij}$,$\boldsymbol{L}$ 是拉普拉斯矩阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$。

能够使得式(3)取最小值的变换矩阵 A 的求解可以转换为如下的广义特征值问题:

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A} \tag{4}$$

将式(4)求解出的特征值按从小到大排列,即 $\lambda_0<\cdots<\lambda_{l-1}$,取前 k 个最小的特征值对应的特征向量 $\boldsymbol{a}_0,\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{k-1}$组成 $\boldsymbol{A}$,即 $\boldsymbol{A}=[\boldsymbol{a}_0\,\boldsymbol{a}_1\cdots\boldsymbol{a}_{k-1}]$,由于 $\boldsymbol{a}_i$ 是列向量,所以 $\boldsymbol{A}$ 是 $d\times k$ 的矩阵。

此外,LPP 不仅适用于原始数据空间,还适用于再生核希尔伯特空间(reproducing kernel hilbert space, RKHS),这样就可以引出核 LPP 算法。

假设欧式空间 R^n 中的数据矩阵通过非线性映射函数 φ 映射到希尔伯特空间 K,即 $\varphi:R^n\to K$。使用 $\boldsymbol{\varphi}(\boldsymbol{X})$ 表示希尔伯特空间中的数据矩阵,即 $\boldsymbol{\varphi}(\boldsymbol{X})=[\boldsymbol{\varphi}(\boldsymbol{x}_1)\boldsymbol{\varphi}(\boldsymbol{x}_2)\cdots\boldsymbol{\varphi}(\boldsymbol{x}_n)]$。那么,在希尔伯特空间中的特征向量问题就可以表示为

$$[\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}]\boldsymbol{v}=\lambda[\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{D}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}]\boldsymbol{v} \tag{5}$$

考虑如下的核函数:

$$\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{\varphi}(\boldsymbol{x}_i)\cdot\boldsymbol{\varphi}(\boldsymbol{x}_j))=\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_j)$$

式(5)中的特征向量是 $\boldsymbol{\varphi}(\boldsymbol{x}_1),\boldsymbol{\varphi}(\boldsymbol{x}_2),\cdots,\boldsymbol{\varphi}(\boldsymbol{x}_n)$ 的线性组合,每一项的系数分别为 $\boldsymbol{a}_i$,$i=1,2,\cdots,m$,即 $\boldsymbol{v}=\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{\varphi}(\boldsymbol{x}_i)=\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{a}$。其中,$\boldsymbol{a}=[\boldsymbol{a}_1\,\boldsymbol{a}_2\cdots\boldsymbol{a}_n]^{\mathrm{T}}$。经过简单的代数变换,可以得到如下特征向量问题:$\boldsymbol{KLKa}=\lambda\boldsymbol{KDKa}$。

4 希尔伯特-施密特独立性度量准则

已知一个参考聚类结果 $C^{(1)}$,使用 RLPP 算法学习相对于 $C^{(1)}$ 独立的子空间 $\boldsymbol{A}$,这样就确保了使用 $\boldsymbol{A}$ 得到的聚类结果 $C^{(2)}$ 与 $C^{(1)}$ 不同。为了计算不同子空间之间的相异性,采用了 HSIC(hilbert-schmidt independence criterion)[1],更重要的是,LPP 与 HSIC 结合后可以导出一个特征分解问题,这样就一定可以计算出全局最佳解。

HSIC 是一种基于核的独立性度量方法,采用 Hilbert-Schmidt 互协方差算子,通过对该算子范数的经验估计得到独立性判断准则。具体来说,已知 X 和 Y 两个随机变量,$\mathrm{HSIC}_{(X,Y)}$ 的值越大说明 X 和 Y 的关联性越强,值等于 0 时说明 X 和 Y 相互之间完全独立。

数学上,令 F 表示再生核希尔伯特空间,$\boldsymbol{\varphi}(\boldsymbol{x})$ 表示数据 $\boldsymbol{x}$ 从原空间映射到 F 中的映射函数,则核函数可以写为$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}})=\langle\boldsymbol{\varphi}(\boldsymbol{x}),\boldsymbol{\varphi}(\boldsymbol{x}^{\mathrm{T}})\rangle$。同样的,定义 $\boldsymbol{\psi}(\boldsymbol{y})$ 为原空间中的数据 y 映射到再生希尔伯特空间 G 的映射函数,核函数可以写为 $\boldsymbol{L}(\boldsymbol{y},\boldsymbol{y}^{\mathrm{T}})=\langle\boldsymbol{\psi}(\boldsymbol{y}),\boldsymbol{\psi}(\boldsymbol{y}^{\mathrm{T}})\rangle$。则互协方差算子 $C_{xy}:G\to F$ 可以被定义为 $C_{xy}=\boldsymbol{E}_{xy}[(\boldsymbol{\varphi}(\boldsymbol{x})-\mu_x)\otimes(\boldsymbol{\psi}(\boldsymbol{y})-\mu_y)]$,$\otimes$ 表示张量积。$\boldsymbol{C}_{xy}$即为 Hilbert-Schmidt 算子,而 HSIC 定义为 $\boldsymbol{C}_{xy}$ 的 Hilbert-Schmidt 算子范数,即 $\mathrm{HSIC}_{(P_{xy},F,G)}=\|\boldsymbol{C}_{xy}\|_{HS}^2$,其中 P_{xy} 表示 X 和 Y 的联合分布。实际上,不需要知道联合分布 P_{xy},已知 n 个观测值 $Z=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$,可以直接给出 HSIC 的经验估计值为 $\mathrm{HSIC}_{(Z,F,G)}=(n-1)^{-2}\mathrm{tr}(\boldsymbol{KHL}_y\boldsymbol{H})$。其中 $\boldsymbol{K},\boldsymbol{L}_y\in\mathbf{R}^{n\times n}$,且 $\boldsymbol{K},\boldsymbol{L}_y$ 分别是核 $\boldsymbol{K}$ 和 $\boldsymbol{L}$ 关于 $\boldsymbol{Z}$ 观测值的 Gram 矩阵,即 $\boldsymbol{K}_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$,$\boldsymbol{L}_{yij}=l(\boldsymbol{y}_i,\boldsymbol{y}_j)=\langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle$,其中 $\boldsymbol{y}_i$ 是一个二元向量,表示对 $\boldsymbol{x}_i$ 的类标签所做的编码(稍后将举例说明)。$\boldsymbol{H}=\boldsymbol{I}-\dfrac{1}{n}\boldsymbol{e}_n\boldsymbol{e}_n^{\mathrm{T}}$,$\mathbf{e}_n$

表示元素值全为 1 的列向量。tr(·)表示矩阵的迹。

为了表示简单,使用 $\mathrm{HSIC}_{(X,Y)}$ 代替 $\mathrm{HSIC}_{(Z,F,G)}$,表示随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{\varphi}(\boldsymbol{x})=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}$,也就是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的依赖性。

假设有 8 个数据$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_8,\}$,其中 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$,$\boldsymbol{x}_3$ 和 $\boldsymbol{x}_4$,$\boldsymbol{x}_5$ 和 $\boldsymbol{x}_6$,$\boldsymbol{x}_7$ 和 $\boldsymbol{x}_8$ 分别为一类。则向量 $\boldsymbol{y}_1=\boldsymbol{y}_2=(1\ 0\ 0\ 0)^{\mathrm{T}}$,$\boldsymbol{y}_3=\boldsymbol{y}_4=(0\ 1\ 0\ 0)^{\mathrm{T}}$,$\boldsymbol{y}_5=\boldsymbol{y}_6=(0\ 0\ 1\ 0)^{\mathrm{T}}$,$\boldsymbol{y}_7=\boldsymbol{y}_8=(0\ 0\ 0\ 1)^{\mathrm{T}}$。矩阵 $\boldsymbol{Y}$ 的每一行对应一个 $\boldsymbol{y}_i$。$\boldsymbol{L}_y$ 是一个 8×8 的矩阵,由 $\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$ 的点积构成。$\boldsymbol{K}$ 是一个 8×8 的矩阵,表示 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$和$\boldsymbol{\varphi}(\boldsymbol{x}_j)$之间的相似度。同时注意,根据定义,$\boldsymbol{H}$ 是一个 $n\times n$(在本例中是 8×8)的常数矩阵,每行每列的和都等于 0。因此,在上述示例中,每一行(列)都包含 7 个 $(-\frac{1}{8})$和 1 个$\frac{7}{8}$。

5　基于局部保留投影的多可选聚类发掘算法

由于通过 $\mathrm{HSIC}_{(X,Y)}$ 可以自然地评估结构很复杂的样本 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关性,因此结合 $\mathrm{HSIC}_{(X,Y)}$ 对 LPP 的目标函数进行修改。要求是转换矩阵 $\boldsymbol{A}$ 必须能够发掘嵌入在高维数据中的低维流形结构,并且与已知的聚类结果 $C^{(1)}$ 完全独立。换句话说,在所有与已经存在的聚类结果 $C^{(1)}$ 不同的子空间中,要选出能够最好地保持高维数据流形结构的子空间。因此,改进 LPP 的目标函数如下:

$$\begin{aligned}\boldsymbol{A}_{\mathrm{opt}}&=\operatorname{argmin}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}+\mathrm{HSIC}_{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X},C^{(1)})}=\\&\operatorname{argmin}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}+\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}_y)\end{aligned}\tag{6}$$

式中:$\boldsymbol{A}_{\mathrm{opt}}$ 表示 $\boldsymbol{A}$ 的最佳解,且由迹的性质可知 $\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}_y)=\mathrm{tr}(\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H})$。不同的核函数在计算变量之间的独立性时结果不同,这里采用线性核函数,映射函数定义为:$\boldsymbol{\varphi}(\boldsymbol{x})=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}$,因此,$\boldsymbol{K}=\langle\boldsymbol{\varphi}(\boldsymbol{X}),\boldsymbol{\varphi}(\boldsymbol{X})\rangle=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$。即

$$\begin{aligned}&\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}+\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}_y)=\\&\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\\&\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}+\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{A}\end{aligned}\tag{7}$$

将数据集合 $\boldsymbol{X}$ 映射到高维特征空间中后,就可以最终得到 $\boldsymbol{\varphi}(\boldsymbol{X})=[\boldsymbol{\varphi}(\boldsymbol{x}_1)\ \boldsymbol{\varphi}(\boldsymbol{x}_2)\ \cdots\ \boldsymbol{\varphi}(\boldsymbol{x}_n)]$。其中,核矩阵 $\boldsymbol{K}$ 的元素为 $K_{ij}=\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_j)$。即:

$$\boldsymbol{A}_{\mathrm{opt}}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}+\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}})\boldsymbol{A}\tag{8}$$

因为 $\boldsymbol{H}$ 和 $\boldsymbol{L}_y$ 都是对称矩阵,所以 $\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$也是对称矩阵,同样,因为 $\boldsymbol{L}$ 是对称矩阵,所以 $\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$ 也是对称矩阵。因此,$\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}+\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$ 是实对称矩阵。作为一个特征分解问题,$\boldsymbol{A}_{\mathrm{opt}}$的最优解由前 k 个最小非零特征值对应的特征向量构成,即 $\boldsymbol{A}=[\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\cdots\ \boldsymbol{\alpha}_k]$。下一步,可以使用 k-means[19] 算法对子空间 $\boldsymbol{A}$ 进行聚类,得到可供选择的聚类结果 $C^{(2)}$。

可以看到,$\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$ 直接影响了 LPP 算法中 $\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$ 项,也就是说,可以把两个聚类结果之间的独立性看作添加的约束项。同时,通过添加更多的 HSIC 项,将算法推广可以找到更多可供选择的聚类结果。

举例来说,在寻找第 3 个可供选择的聚类结果 $C^{(3)}$时,只要提供之前找到的两个聚类结果 $C^{(1)}$ 和 $C^{(2)}$,并把式(6)中的 $\mathrm{HSIC}_{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X},C^{(1)})}$ 一项替换为 $\mathrm{HSIC}_{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X},C_{(1)})}+\mathrm{HSIC}_{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X},C^{(2)})}$ 即可。因此只要在式(8)中使用 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}_{y1}\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}_{y2}\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}$,即直接使用 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}(\boldsymbol{L}_{y1}+\boldsymbol{L}_{y2})\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}$ 代替 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}$。也就是说,使用$(\boldsymbol{L}_{y1}+\boldsymbol{L}_{y2})$代替了 $\boldsymbol{L}_y$,其他矩阵保持不变即可。

RLPP 算法描述如下:

1)输入　数据集 $\boldsymbol{X}$;一个 $\boldsymbol{X}$ 上的参考聚类结果 $C^{(1)}$。

2)输出　一个数据集 $\boldsymbol{X}$ 上可供选择的参考聚类结果 $C^{(2)}$。

3)算法流程:

①计算 $\boldsymbol{L}_y$,$\boldsymbol{L}_y=\langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle$,其中 $\boldsymbol{y}_i$ 是一个二元向量,表示 $C^{(1)}$ 中 $\boldsymbol{x}_i$ 的类标签的编码。

②计算 $\boldsymbol{H}=\boldsymbol{I}-\frac{1}{n}\boldsymbol{e}_n\boldsymbol{e}_n^{\mathrm{T}}$。

③计算权值矩阵 $\boldsymbol{W}$,如果 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的 k 近邻点,那么 $\boldsymbol{W}_{ij}=\exp-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{t}(t\in\mathrm{R})$,否则 $\boldsymbol{W}_{ij}=0$。

④计算矩阵 $\boldsymbol{D}$,$\boldsymbol{D}_{ii}=\sum_j\boldsymbol{W}_{ij}$,计算拉普拉斯矩阵 $\boldsymbol{L}$,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$。

⑤使用高斯核计算核矩阵 $\boldsymbol{K}$,$\boldsymbol{K}_{ij}=\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_j)$。

⑥分解核矩阵 $\boldsymbol{K}$,$\boldsymbol{K}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{P}$,根据 $\boldsymbol{\varphi}(\boldsymbol{X})=\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}$ 得到 $\boldsymbol{\varphi}(\boldsymbol{X})$。

⑦计算 $\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{L}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}+\boldsymbol{\varphi}(\boldsymbol{X})\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}\boldsymbol{\varphi}(\boldsymbol{X})^{\mathrm{T}}$ 的特征值和特征向量。

⑧按特征值从小到大的顺序对特征向量排序。

⑨选择前 k 个最小的特征值对应的特征向量,即 $\boldsymbol{A}=[a_0\ a_1\cdots\ a_{k-1}]$。

⑩$C^{(2)}=\mathrm{k\text{-}means}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{X}))$。

RLPP 算法的时间复杂度完全由计算最近邻矩

阵以及核矩阵决定,因为它们的时间复杂度均为 $O(n^2d)$,因此整体的时间复杂度也为 $O(n^2d)$。

6 实验与分析

6.1 聚类结果评估

聚类结果根据聚类质量和相异性两方面进行评估。聚类质量分为两种情况:如果已知正确的类标,则可选聚类结果和正确的类标之间通过 F-measure 计算,计算公式为 $F=2P\times R/(P+R)$,其中 P 和 R 分别表示准确率(precision)和召回率(recall);否则,使用 Dunn Index 计算,表示为 $DI_{(C)}$。数学上,Dunn Index 定义为 $DI_{(C)}=\dfrac{\min_{i\neq j}\{\delta(c_i,c_j)\}}{x_1\leqslant l\leqslant k\{\Delta(c_l)\}}$,其中 $\delta:C\times C\rightarrow \mathbf{R}_0^+$,表示类与类之间的距离,$\Delta:C\rightarrow\mathbf{R}_0^+$ 表示类内直径。对于评估聚类结果的相异性,使用了两种不同的方法。第 1 种是最为常用的标准化互信息(normalized mutual information, NMI),第 2 种是杰卡德指数(Jaccard index, JI)。

对于 NMI 和 JI 指标,值越小意味着不同聚类结果之间的相似度越高;对于 F-measure 和 Dunn Index 指标,值越大意味着更高的聚类质量。

6.2 人工数据集

使用两种流行的人工数据集评估 RLPP 的效果,并与其他算法进行比较。第 1 组人工数据集 Syn1 分布在二维空间内,分为 4 部分,每部分由 200 个数据点组成,共 800 个数据点。使用数据集 Syn1 的目的是检验算法是否能够尽可能多的发现可供选择的聚类结果,且所有结果均满足与初始聚类结果正交的条件。第 2 组人工数据集 Syn2 的结构较为复杂,每部分的形状都是非凸的。使用数据集 Syn2 的目的是检验算法是否能够处理非线性的数据结构,并且发掘出嵌入在高维数据中的低维流形结构。

图 1 中的第 1 行表示的是 RLPP 使用数据集 Syn1 得到的运行结果。其中,第 1 列表示的是所提供的参考聚类结果 $C^{(1)}$,第 2 列表示的是由 RLPP 得到的可供选择的聚类结果 $C^{(2)}$。从图中可以直观地看出,RLPP 成功地找到了与所提供的参考聚类结果完全不相同,但是聚类质量很高的可选聚类结果。另外,如果我们把该结果 $C^{(2)}$ 看作除 $C^{(1)}$ 外新增的参考聚类结果,并且寻找第 2 个可选的参考聚类结果 $C^{(3)}$,RLPP 会得到第 3 列所显示的聚类结果。$C^{(3)}$ 在欧氏距离下与前两个聚类结果相比不是特别得自然,但是 $C^{(3)}$ 仍然很有启发性,并且它完全独立于前 2 个参考聚类结果 $C^{(1)}$ 和 $C^{(2)}$。同时注意到,RPCA 算法无法寻找出合适的 $C^{(3)}$。在表 1 中,提供了这些算法的表现。

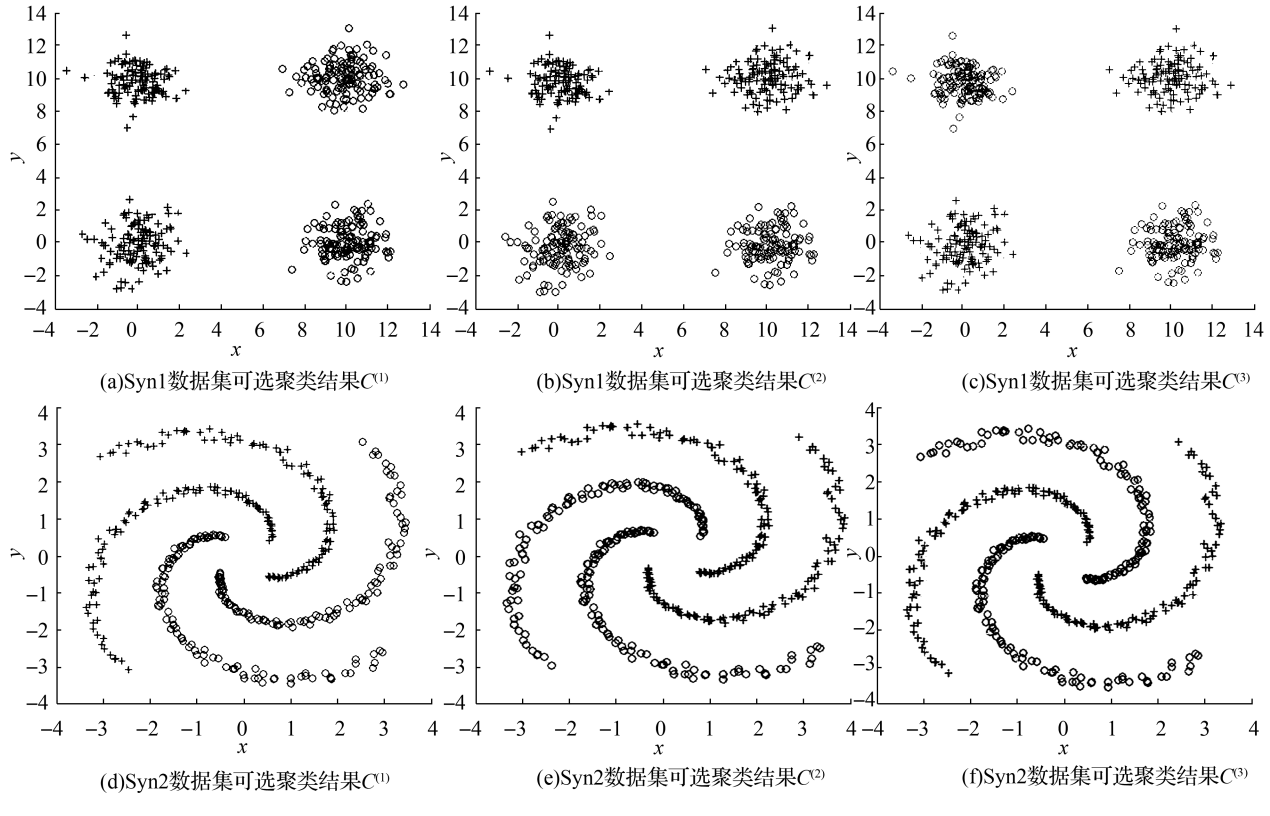

图 1 由数据集 Syn1(第 1 行)和 Syn2(第 2 行)得到的可选聚类结果

Fig.1 Alternative clusterings uncovered from Syn1(1st row) and Syn2(2nd row) datasets

表 1　人工数据集 Syn1 上 3 种算法的表现

Table 1　Clustering performance of all algorithms for synthetic dataset Syn1

算法	NMI_{12}	NMI_{13}	NMI_{23}	JI_{12}	JI_{13}	JI_{23}	F_{12}	F_{13}	F_{23}
RPCA	0.00	\	\	0.33	\	\	1.00	\	\
RegGB	0.00	0.00	0.00	0.33	0.33	0.33	1.00	1.00	1.00
RLPP	0.00	0.00	0.00	0.33	0.33	0.33	1.00	1.00	1.00

表 2　人工数据集 Syn2 上 3 种算法的表现

Table 2　Clustering performance of all algorithms for synthetic dataset Syn2

算法	NMI_{12}	NMI_{13}	NMI_{23}	JI_{12}	JI_{13}	JI_{23}	F_{12}	F_{13}	F_{23}
RegGB	0.00	0.00	0.00	0.33	0.33	0.33	1.00	1.00	1.00
RLPP	0.00	0.00	0.00	0.33	0.33	0.33	1.00	1.00	1.00

6.3　舍尔图数据集

选择文献[11]中所介绍的埃舍尔图(escher image)作为另一个用于寻找多个可选聚类结果实验的数据集。对于人眼来说,埃舍尔图有多种分割结果(即聚类结果)。图 2(a)显示的图片为原始图片,可以看到图中有多只爬行动物,并且聚类时明显可以有多种聚类结果。在分割过程中,图中的每个像素点都表示一个反映了 RGB 信息的数据点。我们使用 k-means 对图 2(a)进行聚类。图 2(b)为 k-means 得到的聚类结果,作为其他算法所需要的参考聚类结果。图 2(c)和图 2(d)分别为 RLPP 得到的可选聚类结果 $C^{(2)}$ 和 $C^{(3)}$,可以看出图 2(c)中的爬行动物为水平姿势,图 2(d)中的爬行动物为垂直姿势。为了对比,提供了由 RegGB 算法得到的结果(RPCA 算法得到的 $C^{(2)}$ 与 RegGB 算法近似,$C^{(3)}$ 则效果很差,因此不加入对比)。图 2(e)和图 2(f)为 RegGB 得到的可选聚类结果 $C^{(2)}$ 和 $C^{(3)}$。从肉眼观察的角度可以发现,图 2(c)与图 2(e)相比轮廓更加清晰,聚类的效果更好。图 2(d)与图 2(f)相比,虽然图 2(f)的结果看似更佳,但是图 2(d)保留了原图中更多的信息,每只爬行动物的轮廓都能够得到保留,这是由于 RLPP 采用了流形子空间学习技术,能够最大程度地保留原始数据的结构。对每种算法重复运行了 10 次,表 3 给出了这些算法的平均表现。

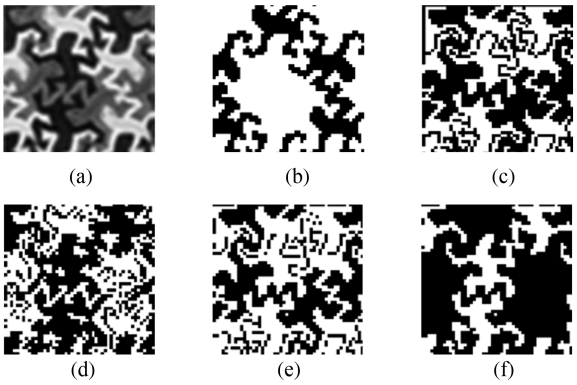

图 2　埃舍尔图数据集上的图像分割结果

Fig.2　Image segmentation results on Escher image data

表 3 埃舍尔图数据集上两种算法的表现

Table 3　Clustering performance of two algorithms on the Escher image data

算法	NMI_{12}	NMI_{13}	NMI_{23}	JI_{12}	JI_{13}	JI_{23}	DI_{12}	DI_{13}	DI_{23}
RegGB	0.05	0.27	0.26	0.39	0.33	0.28	3.81	0.05	2.38
RLPP	0.03	0.06	0.01	0.19	0.39	0.34	3.81	0.02	1.60

6.4　CMUFace 数据集

使用 UCI 数据库中的 CMUFace 数据集检验算法。CMUFace 数据集包含 20 个人的图像,每个人又分为不同的面部表情(正常、高兴、悲伤、生气),不同的头部朝向(向左、向右、向前、向上),不同眼部状况(睁开、墨镜)。每个人有 32 张图片,包含了上述特征的组合。由于图片中的人的身份是已知的,因此身份信息可以作为参考聚类结果直接使用。随机选取了 3 个人的全部图像进行试验。

图 3 显示的是聚类结果的平均值的图像。其中第 1 行是原始图像经由 k-means 算法得到的平均值图像,第 2 行由 RLPP 算法得到,第 3 行和第 4 行由 RPCA 与 RegGB 算法得到。

从图像上看,第 1 行聚类的依据是不同的人,其余 3 行聚类的依据是人不同的头部朝向。很明显,3 种算法都从数据集中得到了另一组完全不同,但是

同等重要的聚类结果。从图中可以看出,RLPP 和 RPCA 的聚类效果最好,RegGB 稍差。表 4 是这 3 种算法的对比。

表 4 CMUFace 数据集上 3 种算法的表现

Table 4 Clustering performance of all algorithms for CMUFace data

算法	NMI	JI	*F*
RPCA	0.012 4	0.294 1	0.713 9
RegGB	0.669 0	0.444 4	0.751 2
RLPP	0.076 6	0.215 1	0.702 1

(a)k-means算法基于不同人的聚类结果

(b)RLPP算法基于头部朝向的聚类结果

(c)RPCA算法基于头部朝向的聚类结果

(d)RegGB算法基于头部朝向的聚类结果

图 3 CMUFace 数据集上的运行结果

Fig.3 Results on CMUFace data

6.5 算法运行时间

实验均在 MARTLAB 8.1.0.604(R2013a)平台下完成,操作系统为 64 位 Windows7,CPU 为 Intel(R) Core(TM) i3-3240 3.40G Hz,内存为 4 GB。

对于人工数据集 Syn1 和 Syn2,RLPP 算法发掘出一个可供选择的聚类结果分别耗时 3.4 s 和2.9 s。对于 Esher 图,由于聚类之前需要进行图像一维化处理,因此数据集的维数很大,共耗时 136 s。对于 CMUFace 数据集,RLPP 算法找到一个可供选择的聚类结果共耗时 2.7 s。以上运行时间均为运行 10 次试验的平均时间。

上述运行时间表明本文算法在合适的数据集上是完全适用的,但是在数据集规模很大的情况下,仍存有改进的空间。

7 结束语

本文提出了一种新的算法 RLPP,采用子空间流形学习技术,寻找可供选择的聚类结果。RLPP 算法的优势在于最终能够转化为特征分解问题,也就是说可以找到封闭解,并且子空间一定是全局最优的,这也是本文区别于其他相关研究的重要特点之一。在文章中对 RLPP 算法进行了论证和实验,并对比了目前效果最好的著名算法。实验结果表明 RLPP 算法的性能不输于甚至优于其他算法。对于如何更好地选取最小特征值的个数 k,以及如何降低算法在处理维数较大数据时的时间复杂度,都是继续研究的方向。

参考文献:

[1] DANG Xuanhong, BAILEY J. Generating multiple alternative clusterings via globally optimal subspaces[J]. Data mining and knowledge discovery, 2014, 28(3): 569-592.

[2] GRETTON A, BOUSQUET O, SMOLA A, et al. Measuring statistical dependence with Hilbert-Schmidt norms[M]// JAIN S, SIMON H U, TOMITA E. Algorithmic Learning Theory. Berlin Heidelberg: Springer, 2005: 63-77.

[3] HE Xiaofei, NIYOGI X. Locality preserving projections [C]//Advances in Neural Information Processing Systems. Vancouver, Canada, 2003, 16: 153-160.

[4] BAE E, BAILEY J. COALA: a novel approach for the extraction of an alternate clustering of high quality and high dissimilarity[C]//Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 53-62.

[5] GONDEK D, HOFMANN T. Non-redundant data clustering [J]. Knowledge and information systems, 2007, 12(1): 1-24.

[6] JAIN P, MEKA R, DHILLON I S. Simultaneous unsupervised learning of disparate clusterings[J]. Statistical analysis and data mining: the ASA data science journal, 2008, 1(3): 195-210.

[7] DANG Xuanhong, BAILEY J. Generation of alternative clusterings using the CAMI approach[C]//Proceedings of the SIAM International Conference on Data Mining, SDM 2010. Columbus, Ohio, USA, 2010, 10: 118-129.

[8] DANG Xuanhong, BAILEY J. A hierarchical information theoretic technique for the discovery of non linear alternative clusterings[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 573-582.

[9] VINH N X, EPPS J. MinCEntropy: a novel information theoretic approach for the generation of alternative clusterings [C]//Proceedings of the IEEE International Conference on Data Mining. Sydney, Australia, 2010: 521-530.

[10] COVER T M, THOMAS J A. Elements of information theory[M]. Chichester: John Wiley & Sons, 2012.

[11] KAPUR J N. Measures of information and their applications[M]. New York: Wiley-Interscience, 1994.

[12]PRINCIPE J C, XU D, FISHER J. Information theoretic learning[M]//HAYKIN S. Unsupervised Adaptive Filtering. New York: Wiley, 2000, 1: 265-319.
[13]PARZEN E. On estimation of a probability density function and mode[J]. The annals of mathematical statistics, 1962, 33(3): 1065-1076.
[14]CUI Ying, FERN X Z, DY J G. Non-redundant multi-view clustering via orthogonalization[C]//Proceedings of the 7th IEEE International Conference on Data Mining. Omaha, Nebraska, USA, 2007: 133-142.
[15]DAVIDSON I, QI Zijie. Finding alternative clusterings using constraints[C]//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 773-778.
[16]QI Zijie, DAVIDSON I. A principled and flexible framework for finding alternative clusterings[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 717-726.
[17]DASGUPTA S, NG V. Mining clustering dimensions [C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 263-270.
[18]NIU Donglin, DY J G, JORDAN M I. Multiple non-redundant spectral clustering views[C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 831-838.

作者简介:

程旸,男,1991 年生,硕士研究生,主要研究方向为人工智能与模式识别、数据挖掘。

王士同,男,1964 年生,教授,博士生导师,中国离散数学学会常务理事,中国机器学习学会常务理事。主要研究方向为人工智能、模式识别和图像处理。发表学术论文近百篇,其中被 SCI、EI 检索 50 余篇。

第 2 届物联网,大数据和安全国际会议
2nd International Conference on Internet of Things, Big Data and Security

24-26 April, 2017, Porto, Portugal

Internet of Things (IoT) is a platform and a phenomenon that allows everything to process information, communicate data, analyze context collaboratively and in the service or individuals, organizations and businesses. In the process of doing so, a large amount of data with different formats and content has to be processed efficiently, quickly and intelligently through advanced algorithms, techniques, models and tools. This new paradigm is enabled by the maturity of several different technologies, including the internet, wireless communication, cloud computing, sensors, big data analytics and machine learning algorithms.

Big Data is another paradigm to describe processing of data to make it 'make sense' to people using IoT. Big Data has five characteristics: volume, velocity, variety, veracity and value. There are reports that businesses and research communities equipped with Big Data skills can provide additional incentives, opportunities, funding and innovation to their long-term strategies. The new knowledge, tools, practices, and infrastructures produced will enable breakthrough discoveries and innovation in physical science, engineering, mobile services, medicine, business, education, earth science, security and risk analysis. For organizations that adopt Big Data, the boundary between the use of private clouds, public clouds, IoT is sometimes very thin to allow better access, performance and efficiency of analyzing the data and understanding the data analysis. A common approach is to develop Big Data in the IoT to deliver "Everything as a Service". In the process of doing so, innovative services known as "Emerging Services and Analytics" can be the highlight and strategic solutions to organizations adopting IoT and Big Data.

Website: http://www.iotbd.org/? y=2017