ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2791−2802
DOI: 10.3724/SP.J.1001.2008.02791

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于成对约束的判别型半监督聚类分析*

尹学松[1,2], 胡恩良[1], 陈松灿[1+]

[1](南京航空航天大学 信息科学与技术学院,江苏 南京 210016)

[2](浙江广播电视大学 计算机科学与技术系,浙江 杭州 310012)

Discriminative Semi-Supervised Clustering Analysis with Pairwise Constraints

YIN Xue-Song[1,2], HU En-Liang[1], CHEN Song-Can[1+]

[1](College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)

[2](Department of Computer Science and Technology, Zhejiang Radio & TV University, Hangzhou 310012, China)

+ Corresponding author: E-mail: s.chen@nuaa.edu.cn

Yin XS, Hu EL, Chen SC. Discriminative semi-supervised clustering analysis with pairwise constraints. *Journal of Software*, 2008,19(11):2791−2802. http://www.jos.org.cn/1000-9825/19/2791.htm

Abstract: Most existing semi-supervised clustering algorithms with pairwise constraints neither solve the problem of violation of pairwise constraints effectively, nor handle the high-dimensional data simultaneously. This paper presents a discriminative semi-supervised clustering analysis algorithm with pairwise constraints, called DSCA, which effectively utilizes supervised information to integrate dimensionality reduction and clustering. The proposed algorithm projects the data onto a low-dimensional manifold, where pairwise constraints based *K*-means algorithm is simultaneously used to cluster the data. Meanwhile, pairwise constraints based *K*-means algorithm presented in this paper reduces the computational complexity of constraints based semi-supervised algorithm and resolve the problem of violating pairwise constraints in the existing semi-supervised clustering algorithms. Experimental results on real-world datasets demonstrate that the proposed algorithm can effectively deal with high-dimensional data and provide an appealing clustering performance compared with the state-of-the-art semi-supervised algorithm.

Key words: semi-supervised clustering; pairwise constraints; closure centroid; projection matrix; clustering analysis

摘 要: 现有一些典型的半监督聚类方法一方面难以有效地解决成对约束的违反问题,另一方面未能同时处理高维数据.通过提出一种基于成对约束的判别型半监督聚类分析方法来同时解决上述问题.该方法有效地利用了监督信息集成数据降维和聚类,即在投影空间中使用基于成对约束的 K 均值算法对数据聚类,再利用聚类结果选择投影空间.同时,该算法降低了基于约束的半监督聚类算法的计算复杂度,并解决了聚类过程中成对约束的违反问题.在一组真实数据集上的实验结果表明,与现有相关半监督聚类算法相比,新方法不仅能够处理高维数据,还有效地提高了聚类性能.

关键词: 半监督聚类;成对约束;闭包中心;投影矩阵;聚类分析

* Supported by the National Natural Science Foundation of China under Grant Nos.60505004, 60773061 (国家自然科学基金)
Received 2008-01-08; Accepted 2008-08-26

中图法分类号: TP181 文献标识码: A

在机器学习和数据挖掘领域中,人们经常遇到大量的无类标号数据.对这些无标号数据进行标号时,不仅费时、费力,有时甚至要付出相当大的代价,如会谈中说话人语音的分割与识别[1]、GPS 数据中的道路检测[2]以及电影片段中不同男演员或女演员的分组[3]等问题.因此,利用样本的先验知识来解决这一问题已成为机器学习领域的研究热点[2−10].半监督聚类正是利用样本的先验信息或背景知识,通过充分利用无标号数据来完成对样本数据的聚类.它也能自然地应用于无监督聚类算法,以达到提高无监督聚类性能的目的,故已开始成为机器学习和数据挖掘中的重要研究内容之一.

现有的半监督聚类算法大致可分为 3 类.第 1 类是基于约束的半监督聚类算法(constraint-based semi-supervised clustering method,简称 CBSSC)[2,4,9,10].这类算法一般使用 must-link 和 cannot-link 成对约束来引导聚类过程.must-link 约束规定:如果两个样本属于 must-link 约束,那么这两个样本在聚类时必须被分配到同一个聚类中.cannot-link 约束则相应地规定:如果两个样本属于 cannot-link 约束,那么这两个样本在聚类时必须被分配到不同聚类之中.第 2 类是基于距离的半监督聚类算法(distance-based semi-supervised clustering method,简称 DBSSC)[3,5,11].这类算法利用成对约束来学习距离度量,从而改变各样本之间的距离,使其有利于聚类.第 3 类是集成了约束与距离的半监督聚类算法(constraint and distance based semi-supervised clustering method,简称 CDBSSC)[6,7].它实际上是上述两类方法的组合.

以上 3 类算法尽管利用成对约束来指导聚类,但在求解过程中常常遇到成对约束的违反问题,因而聚类结果并不十分令人满意.如 DBSSC 通过利用 must-link 和 cannot-link 成对约束来学习一个距离函数,从而改变各样本之间的距离,达到提升聚类性能的效果.但这类算法不能保证在改变样本间的距离后,must-link 点对总能被分组到同一聚类之中,而 cannot-link 点对则常有部分被分配到同一聚类之中,致使约束被违反.CBSSC 是在 K 均值算法的目标函数中通过添加惩罚项来试图解决成对约束的违反问题,但选择合适的惩罚因子是这类算法面临的难题,因此,这类算法仍未能有效地解决约束违反问题.作为前两类方法组合的 CBSSC 继承了它们的不足,因而同样未能实现对约束违反问题的有效解决.此外,这 3 类算法只适用于较低维数的样本数据.在遇到较高维数的数据时,这些算法会显得力不从心,因为在高维数据空间中,不同数据分布和不同距离函数的样本点对之间的距离几乎是相同的[8,12].

针对高维数据聚类问题,Tang[8]等人提出了一种基于特征投影的半监督聚类算法,部分地解决了该问题,但其不足是仅采用了 must-link 和 cannot-link 成对约束得到投影矩阵,而没有采用大量无标号样本数据,因此限制了聚类性能的提高.

针对这些问题,本文提出一种基于成对约束的判别型半监督聚类分析方法(discriminative semi- supervised clustering analysis with pairwise constraints,简称 DSCA).该方法首先利用 must-link 和 cannot-link 成对约束得到投影矩阵,在投影空间中对数据聚类得到聚类标号;其次,利用线性判别分析(linear discriminant analysis,简称 LDA)选择子空间;最后,使用基于成对约束的 K 均值算法对子空间中的数据聚类.该方法有效地利用了监督信息集成数据降维和聚类,即在投影空间中使用基于成对约束的 K 均值算法对数据聚类,再利用聚类结果选择投影空间.同时,新方法提出的基于成对约束的 K 均值算法降低了基于约束的半监督聚类算法的计算复杂度,并解决了聚类过程中成对约束的违反问题.

事实上,对数据聚类来说,得到聚类标号后,LDA 选择的线性子空间是最好的子空间,因为在 LDA 子空间里,各聚类之间能够被有效地分开[13].因此,本文提出的算法利用 LDA 来选择子空间,在子空间里使用新的基于成对约束的 K 均值算法对数据聚类.新算法的贡献表现为以下 3 个方面:

(1) 新算法将动态聚类方法引入半监督聚类之中,即聚类和降维同时进行.现有的半监督聚类方法要么只关注监督信息对聚类的帮助[4−7,9,10]而忽略了对数据的降维,要么分离了聚类与降维[8].新算法利用聚类结果进行子空间选择,然后在子空间中完成数据聚类,两者交替迭代进行,有效地提高了聚类性能.

(2) 新算法解决了成对约束的违反问题.CBSSC 在 K 均值算法的目标函数中加入惩罚项来限制违反

must-link 和 cannot-link 约束的样本对[6,7,9,10];DBSSC 用监督信息去学习度量,进行聚类[3,5,11].通常,这样处理并不能有效地解决 must-link 和 cannot-link 约束的违反问题.新算法借助 must-link 成对约束的等价关系,简化 must-link 成对约束,构成新的 cannot-link 约束,并将其应用到 K 均值聚类之中,基本上解决了 must-link 和 cannot-link 约束的违反问题.

(3) 基于成对约束的 K 均值算法改进了 CBSSC.CBSSC 在 K 均值聚类的目标函数中添加惩罚项,构成新的目标函数来解决约束的违反问题,但选择合适的惩罚因子是该算法面临的难题.因此,该方法不仅难以有效地解决成对约束的违反问题,还增加了算法的计算复杂度.基于成对约束的 K 均值算法只是将 cannot-link 成对约束运用到 K 均值聚类中,在保持 K 均值聚类计算复杂度的情况下,提高了聚类性能.

本文提出的算法不仅集成了投影空间选择与数据聚类,还努力架起一座连接原空间中样本和子空间中样本的桥梁.通过该桥梁,可以在全局最优的子空间中对数据聚类,避免了维数灾难的发生.

1 基于成对约束的判别型半监督聚类分析算法

基于成对约束的判别型半监督聚类分析算法的框图如图 1 所示.对于给定的样本集合 $X=[x_1,x_2,\ldots,x_n]$,其中 $x_i \in \Re_m$, must-link 成对约束集合为 $M=\{(x_i,x_j)\}$,cannot-link 成对约束集合为 $C=\{(x_k,x_l)\}$.基于成对约束的判别型半监督聚类分析算法由 3 步组成.首先是算法初始化,利用给定的 must-link 和 cannot-link 成对约束集合得到一个投影矩阵,在投影空间中对数据聚类得到聚类标号;其次,使用 LDA 选择子空间;最后,利用基于成对约束的 K 均值算法对数据聚类.

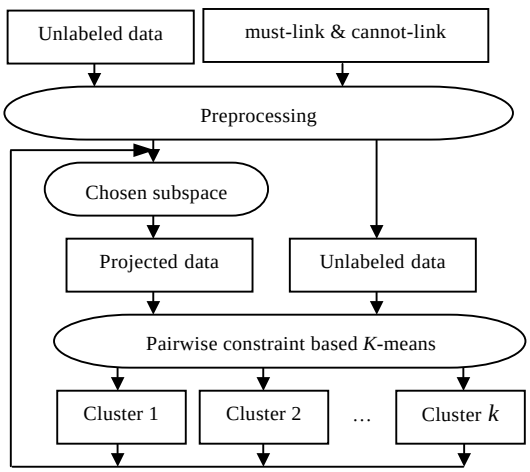


Fig.1 Framework of the DSCA method

图 1 DSCA 算法的框图

1.1 初始化

在高维空间中,不同数据分布和不同距离函数的样本点对之间的距离几乎是相同的,因此,如果对样本聚类,就有必要将高维数据投影到低维空间.设投影矩阵为 $W_{m\times l}=[W_1,\ldots,W_l]$,它包含 l 个 m 维正交单位向量,将原始数据投影到一个低维空间:

$$y_i = W^T x_i \in \Re^l, \quad l<m \tag{1}$$

投影矩阵不仅要在投影空间中尽可能地保持原始数据的结构,还要使 cannot-link 集合中的点对之间距离最大化、must-link 集合中的点对之间距离最小化.因此,定义一个目标函数 $J(W)$,并相对 W 最大化其值来求取投影矩阵[14]:

$$J(W)=\frac{1}{2|C|}\sum_{(x_i,x_j)\in C}\left\|W_{x_i}^T-W_{x_j}^T\right\|^2-\frac{1}{2|M|}\sum_{(x_i',x_j')\in M}\left\|W_{x_i'}^T-W_{x_j'}^T\right\|^2$$
$$=\sum_i W_i^T\left(\frac{1}{2|C|}\sum_{(x_i,x_j)\in C}(x_i-x_j)(x_i-x_j)^T-\frac{1}{2|M|}\sum_{(x_i',x_j')\in M}(x_i'-x_j')(x_i'-x_j')^T\right)W_i \quad (2)$$
$$=\sum_i W_i^T DW_i$$

其中,$|C|$ 和 $|M|$ 分别表示 cannot-link 和 must-link 成对约束中的点对数.

$$D=\frac{1}{2|C|}\sum_{(x_i,x_j)\in C}(x_i-x_j)(x_i-x_j)^T-\frac{1}{2|M|}\sum_{(x_i',x_j')\in M}(x_i'-x_j')(x_i'-x_j')^T$$

$$W_i^TW_j=\begin{cases}1, & \text{if } i=j\\ 0, & \text{otherwise}\end{cases}.$$

通过最大化式(2)可以得到最优的投影矩阵 W.不难发现,由于目标和约束所具有的凸性,我们能够获得 W 的解析解.因此,利用 KT 定理,定义如下 Lagrange 函数:

$$L(W)=J(W_1,W_2,...,W_l)-\sum_{i=1}^{l}\lambda_i(W_i^TW_i-1) \quad (3)$$

相对 W_i 求 $L(W)$的偏导,得到:

$$\frac{\partial L}{\partial W_i}=2DW_i-2\lambda_iW_i=0,\ \forall i=1,...,l$$
$$DW_i=\lambda_iW_i,\ \forall i=1,...,l \quad (4)$$

显然,由方程式(4)可以解出最优的投影矩阵 $W_{m\times l}=[W_1,\ldots,W_l]$就是由矩阵 D 的 l 个最大特征值所对应的特征向量组成.进而可以利用式(1)将原始数据投影到低维空间并使用如下 K 均值算法(K-means)实现对低维数据的聚类:

$$\min J_K=\sum_{j=1}^{K}\sum_{i=1}^{n_j}\|x_i-u_j\|^2 \quad (5)$$

得到聚类标号.

对高维原始数据的聚类,一般地,首先降低它们的维数以避免维数灾难的发生.利用 must-link 和 cannot-link 成对约束,由式(2)得到投影矩阵,在对数据投影时,尽可能地保持原始数据结构,并最大化 cannot-link 点对间的距离、最小化 must-link 点对间的距离.因此,能够得到令人较为满意聚类结果,便于使用 LDA 选择子空间.

1.2 基于成对约束的K均值聚类

作为 CBSSC 和 DBSSC 组合的 CDBSSC,在目标函数中添加惩罚项[6,7],试图解决约束的违反问题.

$$J=\sum_{x_i\in X}\left(\left\|x_i-u_{l_i}\right\|_{A_{l_i}}^2-\log\left(\det(A_{l_i})\right)\right)+\sum_{(x_i,x_j)\in M}\omega_{ij}1[l_i\neq l_j]+\sum_{(x_i,x_j)\in C}\varpi_{ij}1[l_i=l_j] \quad (6)$$

在 CDBSSC 中,违反 must-link 约束的惩罚因子 ω_{ij} 和违反 cannot-link 约束的惩罚因子 ϖ_{ij} 取值分别如下:

$$\omega_{ij}=\frac{1}{2}\left\|x_i-x_j\right\|_{A_{l_i}}^2+\frac{1}{2}\left\|x_i-x_j\right\|_{A_{l_j}}^2 \quad (7)$$

$$\varpi_{ij}=\left\|x_{l_i}'-x_{l_i}''\right\|_{A_{l_i}}^2-\left\|x_i-x_j\right\|_{A_{l_i}}^2 \quad (8)$$

CDBSSC 指出:如果两个样本违反了 must-link 约束,则 ω_{ij} 的值是使用不同聚类中的度量得到的距离之和;如果两个样本违反了 cannot-link 约束,则 ϖ_{ij} 的值是该聚类里选择最远的两个样本之间的距离与这两个样本之间的距离之差.显然,惩罚因子 ω_{ij} 和 ϖ_{ij} 并不能够有效地解决约束的违反问题.

针对该问题,并受文献[8]的启发,本节引入一种基于成对约束的 K 均值聚类算法(pairwise constraint based K-means clustering method,简称 PCBKM),该算法的目的就是寻找 K 个互不相交的样本分割,且不违反

cannot-link 成对约束和 must-link 成对约束.为了解决成对约束的违反问题,首先合并 must-link 约束构建新的 cannot-link 约束作为预处理,具体描述如下:

定义 1(同类闭包). 如果有 3 个样本 a_1,a_2 和 $a_3,(a_1,a_2)\in M,(a_2,a_2)\in M$,则$(a_1,a_3)\in M$,并称 a_1,a_2 和 a_3 构成的集合为同类闭包,简称闭包.

定义 2(闭包中心). 如果$\{a_1,a_2,\ldots,a_p\}$构成一个同类闭包, $a=\frac{1}{p}\sum_{i=1}^{p}a_i$, 则称 a 为闭包中心.

定义 3(异类闭包). 如果存在两个闭包$A=\{a_1,a_2,\ldots,a_p\}$和$B=\{b_1,b_2,\ldots,b_p\}$,以及$(a_i,a_j)\in C$,其中 $a_j\in A,b_j\in B$,则称 A,B 互为异类闭包.

为了简化约束,使用闭包中心之间的 cannot-link 约束代替样本之间的 cannot-link 约束.如图 2 所示,实线表示 must-link 约束,虚线表示 cannot-link 约束,白点表示原始样本,黑点代表闭包中心.其中,$\{a_1,a_2,a_3\}$和$\{b_1,b_2,b_3,b_4,b_5\}$分别表示两个闭包,a,b 分别代表它们的闭包中心.样本之间的cannot-link约束就可以简化为闭包中心 a,b 之间的 cannot-link 约束.特别地,如果一个样本不属于任何约束,则可以构造一个闭包,在该闭包里只有这一个样本元素.

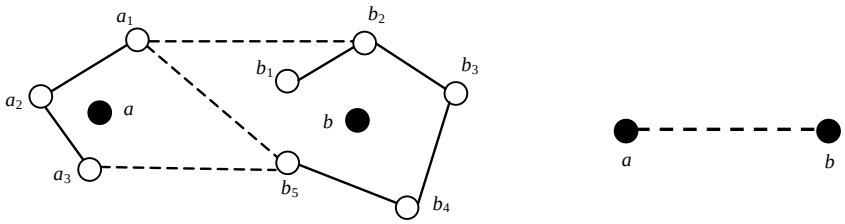


Fig.2 An illustration of combining must-link constraints to form new cannot-link constraints

图 2 合并 must-link 约束构成新的 cannot-link 约束示意图

用闭包中心代替闭包,用两个闭包中心的 cannot-link 约束代替相应的异类闭包中样本间的 cannot-link 约束,则合并 must-link 成对约束以后,原样本数据 X 可简化为 $X_{ml}=[x'_1,x'_2,...,x'_{n_{ml}}](n_{ml}<n)$,cannot-link 成对约束的集合改变为 $C_{ml}=\{(x'_i,x'_j)\}$.

命题 1. 若以闭包中心取代闭包,两个异类闭包的闭包中心分别代替相应的 cannot-link 约束的异类闭包,PCBKM 对闭包中心聚类,则 PCBKM 能够解决 must-link 和 cannot-link 成对约束的违反问题.

证明:分两种情况来证明,一是证明 PCBKM 不违反 must-link 成对约束,二是证明 PCBKM 不违反 cannot-link 成对约束.

(1) PCBKM 不违反 must-link 成对约束

证明:设 $A_i=\{x_1,x_2,\ldots,x_p\}$为第 i 个同类闭包, $\bar{x}_i$ 是 A_i 的闭包中心,$C=\{C_1,C_2,\ldots,C_K\}$是 K 个聚类集合,$U=\{u_1,u_2,\ldots,u_K\}$是 K 个聚类分别对应的聚类中心集合.

$\exists u_i\in U, u_i\in U$, 使得 $u_i=\arg\min_{u_s\in U}\left(\|\bar{x}_i-u_s\|^2\right)$,得到 $\bar{x}_i\in C_i$.

因为闭包中心 $\bar{x}_i$ 代替了 A_i,所以 $\bar{x}_i\in C_i\Rightarrow A_i\subset C_i$.

于是,$\forall x_i\in A_i,A_i\in C_i,i=1,2,\ldots,p$.证毕. □

上述结果避免了同类闭包 A_i 中的元素聚类到其他聚类中,解决了 A_i 中 must-link 成对约束的违反问题.类似地,可以证明其他同类闭包不违反 must-link 成对约束.

(2) PCBKM 不违反 cannot-link 成对约束

证明:设 $A_i=\{x_1,x_2,\ldots,x_p\}$为第 i 个同类闭包, $\bar{x}_i$ 是 A_i 的闭包中心;$A_j=\{x_1,x_2,\ldots,x_q\}$为第 j 个同类闭包, $\bar{x}_j$ 是 A_j 的闭包中心,且 A_i 和 A_j 属于异类闭包.

由于 cannot-link 成对约束要求 $\bar{x}_i$ 和 $\bar{x}_j$ 被聚类到不同的聚类中,故存在 u_i,u_j,且 $u_i\in U,u_j\in U,u_i\in\neq u_j$,使得

$$(u_i,u_j)=\arg\min_{u_s,u_t\in U}\left(\left\|\overline{x}_i-u_s\right\|^2+\left\|\overline{x}_j-u_t\right\|^2\right). \tag{9}$$

因此, $\overline{x}_i\in C_i$, $\overline{x}_j\in C_j$.

因为 $\overline{x}_i$ 是 A_i 的闭包中心, $\overline{x}_j$ 是 A_j 的闭包中心,

所以,可得 $A_i\subset C_i$,$A_j\subset C_j$.

即$\forall x_i\in A_i,x_i\in C_i,i=1,2,\ldots,p;\forall x_j\in A_j,x_j\in C_j,j=1,2,\ldots,p$.证毕. □

由以上证明不难发现,互为异类闭包的两个闭包中的样本分别聚类到两个不同的聚类中,符合 cannot-link 的要求,解决了 cannot-link 成对约束违反问题.

PCBKM 的伪代码见算法 1.PCBKM 的计算复杂度是 $O(tn_{ml}l^2)$ ($n_{ml}<n$),其中,l 是样本维数,n_{ml} 是合并 must-link 成对约束后的样本数,n 是原样本数,t 是迭代次数.不难发现,PCBKM 的计算复杂度小于或者等于 *K*-means 的计算复杂度($O(tnl^2)$).因此,PCBKM 是一种简单而有效的算法.

算法 1. 基于成对约束的 *K* 均值算法(PCBKM).

输入:样本数据 X_{ml}、cannot-link 约束集合 C_{ml} 和聚类数目 *K*.

输出:*K* 个不相交的样本分割.

Step 1. 初始化聚类中心.

Step 2. 重复执行下面的步骤,直到收敛为止.

for i=1 to n_{ml}

(a) 对一个闭包不与任一个闭包构成异类闭包, 找到一个聚类中心 u_k 使该闭包中心 x_i' 满足 $c_k=\arg\min_k\left\|x_i'-u_k'\right\|^2$;

(b) 对于互为异类闭包的两个闭包, x_i' 和 x_j' 是它们的闭包中心,(x_i', x_j')$\in C_{ml}$,找到两个聚类中心 u_i 和 u_j,使 $\min\left(\left\|x_i'-u_i\right\|^2+\left\|x_j'-u_j\right\|^2\right)$;

(c) 对某个聚类 c_i,更新其聚类中心 $u_i=\frac{1}{n_i}\sum_{j=1}^{n_i}x_j^i$.

Step 3. 返回 *K* 个不相交的样本分割.

CDBSSC 在 *K*-means 的目标函数中添加惩罚项,试图解决约束的违反问题.但选择合适的惩罚因子是该类算法面临的难题.因此,这类算法不仅难以解决约束的违反问题,而且还增加了算法的计算复杂度($O(tnl^4)$).本节提出的 PCBKM,其思想是用 must-link,cannot-link 约束分别构成同类闭包和异类闭包,并应用于 *K*-means 中,在保持其计算复杂度的情况下,解决成对约束的违反问题,有效地提高聚类性能,并通过合并 must-link 约束,可以简化原始样本,有助于算法实现大规模数据集的处理.

1.3 基于成对约束的判别型半监督聚类算法

基于上面的描述,本文提出一种同时执行子空间选择和聚类的判别型半监督聚类分析算法,见算法 2.其步骤如下:

算法 2. 基于成对约束的判别型半监督聚类分析算法(DSCA).

输入:样本数据 *X*、must-link 和 cannot-link 成对约束、聚类数目 *K*.

输出:*K* 个样本分割.

Step 1. 根据第 1.1 节,得到 *K* 个样本分割,t=1.

Step 2. 执行以下步骤:

(a) 利用 *K* 个样本分割,在原样本空间中计算 LDA,得到子空间;

(b) 使用 PCBKM 对子空间中的样本聚类,得到 *K* 个样本分割;

(c) $t=t+1$,返回 Step 2,直到算法收敛为止.

Step 3. 返回 K 个样本分割.

类似于文献[13,15,16]的结果,可以证明 DSCA 在有限步内收敛.在实验中观察到,不到 10 次迭代即收敛.

从上述算法可以发现,DSCA 的复杂度主要是在求解 PCBKM 上.PCBKM 的计算复杂度是 $O(tn_{ml}l^2)$,因此,DSCA 的复杂度为 $O(ptn_{ml}l^2)$,p 是 DSCA 迭代次数.

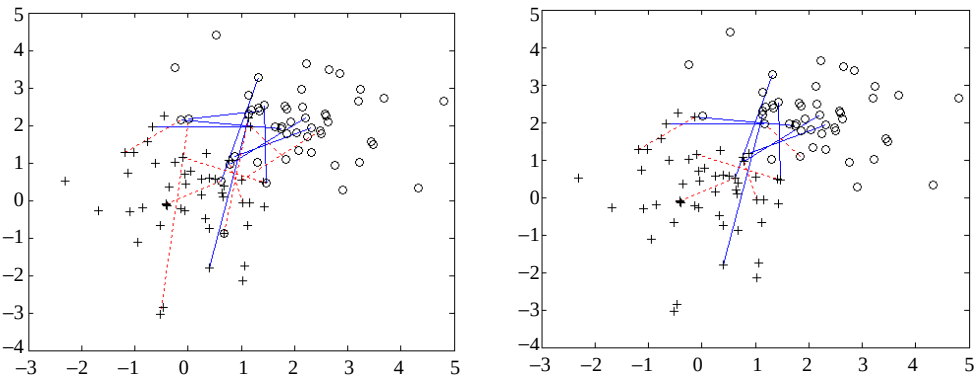
2 实验及其分析

首先比较本文的算法 PCBKM 与 CDBSSC,以验证它们解决成对约束的违反问题.其次,将本文算法 DSCA 与现有相关的半监督算法进行比较.

2.1 PCBKM与CDBSSC的比较

2.1.1 人工数据集上的实验

考虑两类人工数据集,如图 3(a)所示,实线连接的两个样本属于同一聚类,是 must-link 约束;虚线连接的两个样本属于不同聚类,是 cannot-link 约束.在实验中,违反 must-link 约束用实线表示,违反 cannot-link 约束用虚线表示.如果解决了约束违反问题,则不用实线和虚线表示.



(a) Original must-link and cannot-link constraints
(a) 原 must-link 和 cannot-link 约束

(b) Clustered results of CBSSC
(b) CBSSC 聚类结果

(c) Clustered results of CDBSSC
(c) CDBSSC 聚类结果

(d) Clustered results of PCBKM
(d) PCBKM 聚类结果

Fig.3 Comparison of the violation issue of the constraints solved by PCBKM and CDBSSC

图 3 PCBKM 与 CDBSSC 解决约束违反问题的比较

从图 3(b)中不难发现,实线连接的 must-link 点对已经被 CBSSC 分组到不同的聚类中,虚线连接的 cannot-link 点对被聚类成相同的类.因此,该算法未能有效地解决约束的违反问题.同样的问题可以由图 3(c)中

发现,即 CDBSSC 仍未能有效地解决约束的违反问题.本文提出的 PCBKM 解决了成对约束的违反问题,如图 3(d)所示.

2.1.2 真实数据集上的实验

通过两个 UCI 数据集,分别测试 CDBSSC 和 PCBKM 解决约束的违反问题.在表 1 中,对 Ionosphere 数据集而言,选择 must-link 和 cannot-link 的点对分别为 79 对和 66 对,CDBSSC 聚类时,违反 must-link 约束的点对数为 14 对,违反 cannot-link 约束的点对数为 36 对.本文提出的算法 PCBKM 解决了成对约束的违反问题.

Table 1 Comparison of the numbers of constraints violated by PCBKM and CDBSSC

表 1 PCBKM 和 CDBSSC 违反成对约束数的比较

Dataset	must-link	cannot-link	CDBSSC		PCBKM	
			must-link	cannot-link	must-link	cannot-link
Ionosphere	79	66	14	36	0	0
Iris	16	16	6	7	0	0

2.2 DSCA与其他相关半监督算法的比较

2.2.1 实验设置

首先,我们从 UCI 数据集上选择了 7 个较低维数的数据集,它们分别是 Balance,Ionosphere,Iris,Letter,Soybean,Vehicle 和 Wine.同时选择了 5 个较高维数的数据集,它们是 YaleB 人脸数据集、ORL 人脸数据集和 3 个文本数据集 News-Different,News-Same,News-Similar.

然后,我们选择 4 种聚类算法作为对比,来验证 DSCA 的性能.这 4 种算法是:

(1) *K*-means.它是一种常用的、适合于较低维数的样本数据的聚类算法.

(2) 相关成分分析算法(RCA)[3].该算法是最近被提出来的一种半监督度量学习算法,实验结果已经表明其性能优于由 Xing 等人[5]提出的基于约束的半监督度量学习算法.在使用 RCA 对样本数据进行度量变换后,使用 *K*-means 对变换后的数据聚类.

(3) 局部线性嵌入(LLE)[17].它是无监督降维方法.使用该方法对较高维数的样本数据降维,然后使用 *K*-means 对数据聚类.

(4) 特征投影的半监督聚类算法(SCREEN)[8].该方法借助于 must-link 和 cannot-link 成对约束得到投影矩阵,在子空间中用基于约束的球形 *K* 均值算法对数据聚类.它既可用于对低维数据聚类,也可用于对高维数据聚类.

本文采用规范化互信息(NMI)作为聚类的评价方法[6].如果 *C* 是样本聚类后的类标号,*Y* 是样本原有类标号,则 NMI 表示为

$$NMI(C,Y)=\frac{I(C;Y)}{(H(C)+H(Y))/2} \tag{10}$$

其中 $I(C;Y)=H(Y)-H(Y|C)$是 *C* 和 *Y* 之间的互信息,$H(Y)$是 *Y* 的香农熵,$H(Y|C)$是在给定 *C* 的条件下,*Y* 的条件熵.NMI 值的范围在 0,1 之间,NMI 值越大,聚类的性能就越好.一般而言,NMI 评价方法要优于其他评价方法[8].

最后,算法运行在 Intel Pentium 3.00GHz CUP,1G 内存 Windows 环境下的机器上.5 种算法分别在每个数据集上重复实验 15 次,在每次实验中选择 100 对成对约束,并在整个数据集上来测试算法的性能,取 15 次实验 NMI 值的均值作为最终的聚类结果.在对数据降维时,一般设置维数降到 *K*–1 维(其中 *K* 是聚类数).

2.2.2 实验结果

表 2 和表 3 是在 12 个数据集上分别执行 5 种算法得到的 NMI 值.表 2 中的数据是低维数据,在对低维数据聚类时,DSCA 的 NMI 值在 Iris,Letter,Soybean 和 Wine 四个数据集上占优,SCREEN 的 NMI 值在 Balance 数据集上效果好,RCA 的 NMI 值在 Ionosphere 和 Vehicle 两个数据集上占优.总体而言,DSCA 的性能在低维数据集上优于其他 3 种算法.表 3 中的数据是高维数据.不难发现,在对高维数据聚类时,DSCA 的 NMI 值要高于其他 3 种算法的 NMI 值.其次,NMI 值较好的是 SCREEN,而 RCA 对高维数据进行处理时 NMI 值较差.所以,对高维数

据聚类,DSCA 的性能要优于其他几种算法.DSCA 利用了监督信息集成数据降维和聚类,且聚类时考虑了成对约束的违反问题,因此,DSCA 的性能要优于其他几种算法.

在 DSCA 初始化得到投影矩阵后,使用 PCBKM 代替 *K*-means,这样有利于 LDA 选择子空间.

Table 2 Comparison of NMI achieved by DSCA and three methods on seven low-dimensional datasets

表 2 DSCA 与 3 种算法在 7 个低维数据集上的 NMI 值的比较

Dataset	Instance	Dimension	Class	*K*-means	RCA	SCREEN	DSCA
Balance	625	4	3	0.172 6	0.431 8	**0.565 7**	0.458 7
Ionosphere	351	34	2	0.167 7	**0.578 9**	0.445 5	0.411 8
Iris	150	4	3	0.732 7	0.855 8	0.921 1	**0.929 3**
Letter (a~d)	3 096	16	4	0.206 7	0.473 4	0.437 4	**0.502 1**
Soybean	47	35	4	0.617 8	0.882 4	0.803 2	**0.895 2**
Vehicle	846	18	4	0.214 0	**0.470 8**	0.246 6	0.251 5
Wine	178	13	3	0.351 1	0.467 1	0.485 6	**0.522 7**

Table 3 Comparison of NMI achieved by DSCA and three methods on five high-dimensional datasets

表 3 DSCA 与 3 种算法在 5 个高维数据集上的 NMI 值的比较

Dataset	Instance	Dimension	Class	PCA+RCA	LLE	SCREEN	DSCA
Different	300	16 090	3	0.247 3	0.582 2	0.650 8	**0.690 0**
Same	295	16 090	3	0.100 7	0.396 0	0.434 1	**0.544 1**
Similar	288	16 090	3	0.190 9	0.385 1	0.520 3	**0.527 4**
YaleB	110	2 500	10	0.599 2	0.790 2	0.858 6	**0.886 8**
ORL	100	1 024	10	0.582 7	0.665 4	0.828 5	**0.869 1**

2.3 成对约束对DSCA性能的影响

Must-Link 和 cannot-link 成对约束由用户提供,半监督聚类算法一般认为这两个约束是正确的.本文提出的 DSCA 和其他半监督聚类算法借助于 must-link 和 cannot-link 成对约束来提高聚类性能.

由表 2 和表 3 不难发现,DSCA 的性能优于其他几种算法.为了更好地理解成对约束对半监督聚类算法性能的影响,实验过程中选择了不同数量的成对约束.从图 4 中可以观察到,尽管几种算法的性能都随着成对约束的数量增多而逐渐提高,但不同数量的成对约束对几种算法性能的影响是不同的.在成对约束数量较少(如 10)时,DSCA 的 NMI 值高于其他几种算法的 NMI 值.在成对约束数量逐渐增多时,DSCA 的 NMI 值平稳上升,其性能优于其他几种算法.

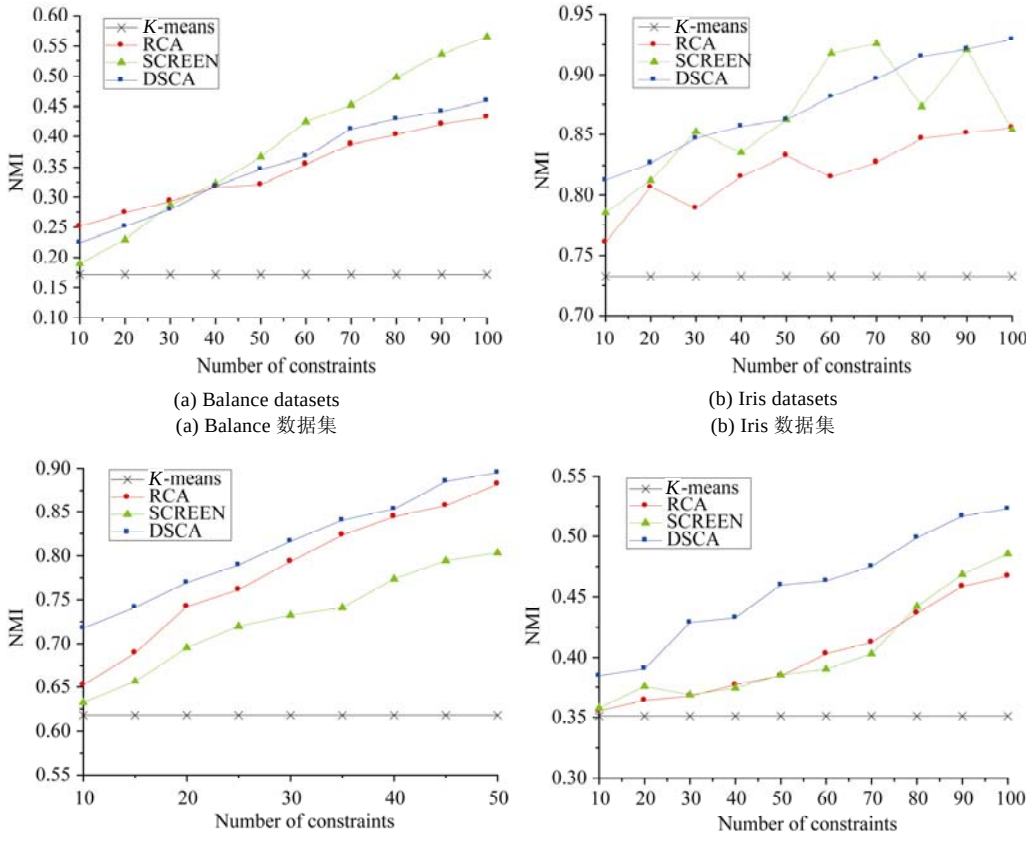
SCREEN 借助于 must-link 和 cannot-link 成对约束得到投影矩阵,在子空间中用基于约束的球形 *K* 均值算法对数据聚类.因此,选择的 must-link 和 cannot-link 成对约束得到好的投影矩阵对该算法至关重要.本文提出的 DSCA 是在初始化步骤里借助于这两种约束得到投影矩阵,进而在子空间中得到初始聚类.所以,当选择的 must-link 和 cannot-link 成对约束得不到好的投影矩阵时,对 DSCA 的聚类会产生一定的影响,但由于 DSCA 利用无标号样本来重新得到投影矩阵,所以在固定成对约束数量的情况下,选择相同的成对约束对 DSCA 性能的影响比对 SCREEN 性能影响要小.

要解决上述问题,一个自然的想法就是需要尽可能多地选择 must-link 和 cannot-link 成对约束,但提供更多的成对约束需要付出很大的代价.因此,在有限的 must-link 和 cannot-link 成对约束中,选择有利于聚类算法的成对约束将是一个挑战.

3 相关工作

CBSSC 首先是由 Wagstaff[2]等人提出来的,他们将 must-link 和 cannot-link 成对约束运用到无监督聚类中.由于该算法在聚类过程中严格限制成对约束的使用,因此,该算法的聚类性能受到了影响.Bual[4]等人提出,CBSSC 是从带有类标号的样本中设法找到更好的初始聚类中心,并将这些带有类标号的样本用于聚类过程.该方法使用的有监督信息是带有类标号的样本,而不是成对约束.相对于约束的半监督聚类算法,基于距离的半监督聚类算法是通过有监督信息去学习一个距离函数,以此来提高聚类性能.Xing[5]等人利用成对约束和牛顿

迭代法来学习一个马氏距离并应用到聚类中.Bar-Hillel[3]等人提出的方法仅仅利用 must-link 成对约束得到块(chunklet)协方差矩阵,然后对块协方差矩阵进行白化变换来学习一个马氏距离.Yeung[11]等人提出的方法是Bar-Hillel 所提出方法的改进,用 must-link 和 cannot-link 成对约束得到块协方差矩阵.Schultz[18]等人提出的方法是从相对约束关系中学习一个带有权值的欧氏距离.Bual 和 Bilenko[6,7]提出的方法是集成 must-link,cannot-link 成对约束和度量学习,应用到聚类中.



(a) Balance datasets
(a) Balance 数据集

(b) Iris datasets
(b) Iris 数据集

(c) Soybean datasets
(c) Soybean 数据集

(d) Wine datasets
(d) Wine 数据集

Fig.4 Relative impact on the performance of choosing the different numbers of pairwis constraints

图 4 选择不同数量的成对约束对算法性能的影响

上述算法只适用于低维空间.对高维数据聚类,一种方法是利用无监督降维方法对高维数据降维,然后使用上述方法对数据聚类;另一种方法是寻找新的算法,借助于有监督信息,对高维数据降维并聚类.前一种聚类方法已经得到验证,结果不能令人满意[8].因此,只能寻找新的半监督聚类方法来解决高维数据聚类问题.Wei[8]等人提出基于特征投影的半监督聚类算法,该算法利用 must-link 和 cannot-link 成对约束得到投影矩阵,在投影空间中应用基于约束的球形 *K* 均值聚类方法对数据聚类.该算法可以解决高维数据聚类问题,但其缺点是只使用 must-link 和 cannot-link 成对约束得到投影矩阵,既没有考虑到大量无标号样本数据,又忽略了降维和聚类的相互促进,因此限制了聚类性能的提高.

最近,一些学者针对高维数据的无监督聚类提出了新的方法.De la Torre[19]等人提出判别聚类分析算法.该算法集成了降维和聚类,即首先对高维数据降维,然后在低维空间中对数据聚类.Chris Ding[13]等人提出了一种自适应无监督降维迭代算法.该算法使用 *K*-means 来产生数据的类标号,然后用线性判别分析方法对高维数据

降维.在降维空间中,再使用 *K*-means 方法对数据聚类.Ye[15]提出了自适应距离学习聚类算法.该算法同样是集成降维和聚类,但不同的是,算法是最优化同一个目标函数来得到聚类和降维.以上 3 种方法的目的是通过降维来帮助聚类,再通过聚类来指导降维,但它们都没有解决好一个难题,即在高维空间中,算法是先执行降维还是先执行聚类.本文提出的基于成对约束的判别型半监督聚类分析算法先借助于成对约束求解投影矩阵,然后在子空间中聚类,再利用聚类结果指导降维.显然,新算法不仅在很大程度上解决了上述无监督聚类算法所面临的问题,而且充分利用有监督信息,有效地提高了聚类性能.

借助于一部分有监督信息,半监督聚类在很大程度上提高了无监督聚类的性能.因此,一些研究人员着力去研究监督信息在聚类分析中的作用,并将其运用到半监督分类和半监督回归等方面.目前,半监督学习已经受到越来越多的研究者的重视.

4 结束语

本文提出了一种基于成对约束的判别型半监督聚类分析方法.新方法首先利用 must-link 和 cannot-link 成对约束得到初始投影矩阵,在投影空间中对数据聚类;然后,利用 LDA 选择子空间;最后,使用基于成对约束的 *K* 均值算法对子空间中的数据聚类.该方法有效地利用了监督信息集成数据降维和聚类,即在投影空间中使用基于成对约束的 *K* 均值算法对数据聚类,再利用聚类结果选择投影空间.同时,新方法提出的基于成对约束的 *K* 均值算法降低了基于约束的半监督聚类算法的计算复杂度,并解决了聚类过程中成对约束的违反问题.

在半监督聚类算法中,监督信息越多,越有助于算法性能的提高.但由于监督信息由用户提供,代价较大.因此,在有限的 must-link 和 cannot-link 成对约束中,选择有利于提高算法性能的成对约束将是一个让人感兴趣的课题,也是我们下一阶段工作的方向.

致谢 张道强教授和蔡维玲博士对本文的工作提出了有益的建议,我们在此表示感谢.

References:

[1] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2005,6(5):937−965.

[2] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained *K*-means clustering with background knowledge. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th Int'l Conf. on Machine Learning. Williamstown: Morgan Kaufmann Publishers, 2001. 577−584.

[3] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning distance functions using equivalence relations. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning. Washington: Morgan Kaufmann Publishers, 2003. 11−18.

[4] Basu S, Banerjee A, Mooney RJ. Semi-Supervised clustering by seeding. In: Sammut C, Hoffmann AG, eds. Proc. of the 19th Int'l Conf. on Machine Learning. Sydney: Morgan Kaufmann Publishers, 2002. 19−26.

[5] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning with application to clustering with side-information. In: Becher S, Thrun S, Obermayer K, eds. Proc. of the 16th Annual Conf. on Neural Information Processing System. Cambridge: MIT Press, 2003. 505−512.

[6] Basu S, Banerjee A, Mooney RJ. A probabilistic framework for semi-supervised clustering. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 59−68.

[7] Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning. New York: ACM Press, 2004. 81−88.

[8] Tang W, Xiong H, Zhong S, Wu J. Enhancing semi-supervised clustering: a feature projection perspective. In: Berkhin P, Caruana R, Wu XD, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 707−716.

[9] Basu S, Banerjee A, Mooney RJ. Active semi-supervision for pairwise constrained clustering. In: Jonker W, Petkovic M, eds. Proc. of the SIAM Int'l Conf. on Data Mining. Cambridge: MIT Press, 2004. 333−344.

[10] Yan B, Domeniconi C. An adaptive kernel method for semi-supervised clustering. In: Fürnkranz J, Scheffer T, Spiliopoulou M, eds. Proc. of the 17th European Conf. on Machine Learning. Berlin: Sigma Press, 2006. 18−22.

[11] Yeung DY, Chang H. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints. Pattern Recognition, 2006,39(5):1007−1010.

[12] Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “Nearest Neighbors Meaningful”? In: Beeri C, Buneman P, eds. Proc. of the Int’l Conf. on Database Theory. New York: ACM Press, 1999. 217−235.

[13] Ding CH, Li T. Adaptive dimension reduction using discriminant analysis and *K*-means clustering. In: Ghahramani Z, ed. Proc. of the 19th Int’l Conf. on Machine Learning. New York: ACM Press, 2007. 521−528.

[14] Zhang DQ, Zhou ZH, Chen SC. Semi-Supervised dimensionality reduction. In: Mandoiu I, Zelikovsky A, eds. Proc. of the 7th SIAM Int’l Conf. on Data Mining. Cambridge: MIT Press, 2007. 629−634.

[15] Ye JP, Zhao Z, Liu H. Adaptive distance metric learning for clustering. In: Bishop CM, Frey B, eds. Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Madison: IEEE Computer Society Press, 2007. 1−7.

[16] Chen JH, Zhao Z, Ye JP, Liu H. Nonlinear adaptive distance metric learning for clustering. In: Berkhin P, Caruana R, Wu XD, eds. Proc. of the 13th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 123−132.

[17] Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2003,4(3):119−155.

[18] Schultz M, Joachims T. Learning a distance metric from relative comparisons. In: Thrun S, Saul LK, Schölkopf B, eds. Proc. of the 17th Annual Conf. on Neural Information Processing System. Cambridge: MIT Press, 2004. 41−48.

[19] De la Torre F, Kanade T. Discriminative cluster analysis. In: William WC, Andrew M, eds. Proc. of the 19th Int’l Conf. on Machine Learning. New York: ACM Press, 2006. 241−248.

尹学松(1975－),男,安徽长丰人,博士生,主要研究领域为模式识别,神经计算.

胡恩良(1975－),男,博士生,主要研究领域为模式识别,神经计算.

陈松灿(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为模式识别,神经计算,机器学习,图像处理.