计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700092

基于安全性的成对约束扩充算法

杨 帆[1] 王俊斌[1] 白 亮[1,2]
1 山西大学计算机与信息技术学院 太原 030006
2 山西大学计算机智能与中文信息处理教育部重点实验室 太原 030006
(18335149237@163.com)

摘 要 基于成对约束的聚类分析是半监督学习的一个重要研究方向。成对约束的数量已成为影响该类算法有效性的重要因素。然而,在现实应用中,成对约束的获取需要耗费大量的成本。因此,文中提出了一种基于安全性的成对约束扩充方法(Extended Algorithm of Pairwise Constraints Based on Security,PCES)。该算法将传递闭包中最大局部连通距离作为安全值,并根据安全值来修改传递闭包之间的相似性,减少合并传递闭包带来的风险,最后利用图聚类方法合并相似的传递闭包达到扩充成对约束的目的。该算法不仅可以安全有效地扩充成对约束,同时可以将扩充后的成对约束应用到不同半监督聚类算法中。文中在8个基准数据集上进行了成对约束扩充算法的比较。实验结果表明,该算法可以安全有效地扩充成对约束。

关键词:成对约束;半监督聚类;监督信息的有效性;监督信息的扩展
中图法分类号 TP391

Extended Algorithm of Pairwise Constraints Based on Security

YANG Fan[1],WANG Jun-bin[1] and BAI Liang[1,2]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory Computation Intelligence and Chinese Information Processing of Ministry of Education,Shanxi University,Taiyuan 030006,China

Abstract Cluster analysis based on pairwise constraints is an important research direction of semi-supervised learning. The number of pairwise constraints has become an important factor affecting the effectiveness of this type of the algorithm. However,in practical applications,the acquisition of pairwise constraints requires a lot of costs. Therefore,the extended algorithm of pairwise constraints based on security (PCES) is proposed. This algorithm takes the maximum local connected distance in the transitive closures as the safe value. According to the safe value,the similarity between the different transitive closures is modified to reduce the risk of merging transitive closures. Finally,the method of graph clustering is used to merge similar transitive closures to extend the pairwise constraints. This algorithm can not only safely and effectively expand pairwise constraints,but also apply the extended pairwise constraints to different semi-supervised clustering algorithms. This paper compares the extended algorithm of pairwise constraints on eight benchmark data sets. The experimental results show that the proposed algorithm can extend pairwise constraints safely and effectively.

Keywords Pairwise constraints,Semi-supervised clustering,Effectiveness of supervision information,Expansion of supervision information

1 引言

在现实世界中,各个领域内都存在大量的数据。为了从数据中找出有价值的信息,机器学习和数据挖掘等技术应运而生。其中,聚类分析是机器学习和数据挖掘研究中的一项重要技术[1-3]。聚类的目的是根据一组对象的特征将它们划分为不同的类,使得同一类中的对象高度相似,而不同类中的对象明显不同。为了达到这个目的,已经有大量的不同类型的聚类算法被提出,如划分聚类算法[4]、层次聚类算法[5]以及基于密度的聚类算法[6]。

由于聚类是通过计算对象间相似性获得聚类结果的无监督方法,聚类算法得到的聚类结构可能与用户所期望的结果

到稿日期:2020-04-30 返修日期:2020-07-14 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61773247,61876103);山西省基础研究计划(201901D211192)
This work was supported by the National Natural Science Foundation of China (61773247,61876103) and Technology Research Development Projects of Shanxi (201901D211192).
通信作者:白亮(bailiang@sxu.edu.cn)

不同[7]。半监督聚类算法通过利用预先给定的少量关于类结构的先验信息来指导聚类过程，使得聚类结果尽可能地满足用户期望的类划分结果[8-10]。目前，已经有大量的半监督聚类算法被提出，一些方法已经被成功地应用于图像分割、自然语言处理和社会网络分析等不同领域。然而，在实际应用的许多任务中，由于数据和标注数据方式的不同，往往会得到许多不同类型的监督信息[11]。

2　相关工作

基于成对约束的半监督聚类算法利用成对约束（两个数据对象之间的关系）来指导聚类过程。Wagstaff 等[12]提出了一种名为 COP-K-Means 的半监督聚类算法。该算法将每次 K-Means 聚类结果中出现违反成对约束的次数作为惩罚项加入目标函数中以改进聚类结果。Yang 等[13]提出了一种约束自组织映射集成框架来改进 COP-K-Means 算法。Wei 等[14]提出了一种基于成对约束和度量学习的半监督聚类框架。Li 等[15]提出了一种基于成对约束的非负矩阵分解算法。Ren 等[16]将成对约束引入聚类过程中，解决了基于密度的聚类方法难以找到合适聚类数量的问题。Miyamoto 等[17]将成对约束作为合并不同类的惩罚项，提出了一种半监督层次聚类算法。Kamvar 等[18]根据成对约束修改相似性矩阵，并使用谱聚类方法得到最终聚类结果。类似地，Xu 等[19]以相同的方法修改相似性矩阵，然后通过随机游走算法获得聚类结果。Ji 等[20]和 Wang 等[21]将约束矩阵看作正则化因子来修改相似性矩阵。Kulis 等[22]通过核方法提升半监督图聚类算法的性能。Ding 等[23]将成对约束引入图分割目标函数，得到了基于隐式马尔可夫随机场的半监督近似谱聚类结果。

成对约束的数量是影响基于成对约束的半监督聚类算法有效性的重要因素。对预先给定的少量成对约束，进行安全有效的扩充，是在较少成对约束的情况下提升半监督聚类算法精度的一种方法。为此，一部分学者在算法过程中扩充成对约束，使得聚类结果尽可能地满足扩充后的约束。如 Klein 等[24]利用初始的成对约束修改数据对象之间的距离，并通过在修改后的距离空间上执行最短路径算法和层次聚类算法扩充成对约束。由于该算法是在修改后的样本空间中执行层次聚类算法来隐式地扩充勿连约束，可能导致错误扩充。Lu 等[25]提出了基于成对约束的标签传播算法，该算法可以在相似性矩阵的行和列两个方向上扩充成对约束。然而，算法的性能很容易受成对约束稀疏性的影响。

另一部分学者利用成对约束的性质来扩充成对约束，并将扩充后的成对约束应用到半监督聚类算法中。如 Wang 等[26]利用成对约束的传递性将扩充后的成对约束应用到一种密度敏感的半监督谱聚类中。Wei 等[27]利用成对约束的传递性和对称性进行扩充，并提出了一种约束与度量相结合的半监督集成算法来提升聚类精度。同样，Yu 等[28]不仅利用成对约束的传递性扩充了成对约束，而且使用文献[25]中的算法，在相似性矩阵的行列方向传播成对约束，将传播后的结果应用到一种自适应性的聚类集成框架中以提升聚类性能。

虽然许多文献都使用了扩充后的成对约束，但是大多是利用成对约束的传递性来进行扩充的。当成对约束数量过少时，相对于初始的成对约束，扩充后的成对约束对聚类结果性能的影响可能微乎其微。而且，在文献[24]和文献[25]所提出算法的过程中扩充成对约束，不能将扩充后的成对约束应用到不同半监督聚类算法中。目前仍缺乏一种可以安全有效地扩充成对约束，并且可以将扩充后的成对约束应用到不同半监督聚类算法中的方法。因此，本文提出了一种基于安全性的成对约束扩充方法。该算法将传递闭包中的最大局部连通距离作为安全值，并且依据安全值来修改传递闭包间的相似性，最后通过模块度算法合并相似的传递闭包达到扩充成对约束的目的。

3　成对约束的传递性

对于用户来说，在许多实际的聚类应用领域中，如基于 GPS 数据的道路检测[24]和蛋白质功能预测[29]等，确定样本所属的类是非常困难的，但是得到数据对象之间的关系却是相对容易的。Wagstaff 等[30]最先提出成对约束的概念（即必连约束（must-link）和勿连约束（cannot-link））来指导聚类过程。必连约束表示两个数据对象必须在同一个类结构中，勿连约束表示两个数据对象不能在同一个类结构中。下面将介绍关于成对约束的性质。

假设 $X=\{x_1,\cdots,x_n\}$ 表示具有 n 个数据对象的数据集合，$L=\{1,\cdots,k\}$ 是对应的类标签集合，其中 k 表示数据集对应的真实聚类数量。令 $M=\{(x_i,x_j):y_i=y_j,1\leqslant i,j\leqslant n\}$ 表示一组必连约束集合，$C=\{(x_i,x_j):y_i\neq y_j,1\leqslant i,j\leqslant n\}$ 表示一组勿连约束集合，其中 y_i 和 y_j 分别是数据对象 x_i 和 x_j 对应的类标签。由于必连约束表示的是两个数据对象的等价关系，它具有自反性、对称性和传递性[28]。基于这些性质，可以得到以下推论。

推论 1[28]　根据必连约束的传递性，对于数据对象 x_i，x_j 和 x_h 有：

$$(x_i,x_j)\in M,(x_j,x_h)\in M\Rightarrow(x_i,x_h)\in M$$

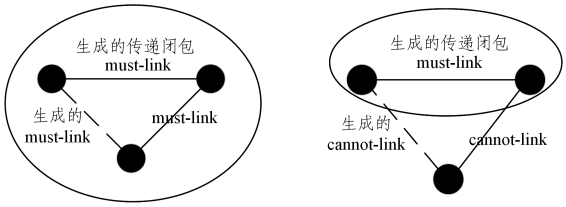
其中，$i,j,h\in\{1,2,\cdots,n\}$。

根据推论 1 可以发现，数据对象 x_i，x_j 与 x_h 之间存在等价关系，这意味着它们属于同一个类结构。因此，将所有的必连约束通过传递性进行扩充可以得到不同类结构对应的传递闭包。

推论 2[28]　根据必连约束和勿连约束的定义，成对约束具有以下性质：

$$(x_i,x_j)\in M,(x_j,x_h)\in C\Rightarrow(x_i,x_h)\in C$$

通过推论 2 可以发现，如果两个传递闭包中的数据对象之间存在一个或多个勿连约束，那么可以在两个传递闭包中所有的对象之间增加勿连约束来扩充勿连约束。图 1 展示了成对约束的性质。



(a)成对约束的传递性以及 must-link 约束的扩充　(b)扩充 cannot-link 约束

图 1　成对约束的性质

Fig. 1　Nature of pairwise constraints

虽然根据传递性和成对约束的性质可以很好地扩充成对约束，但是当成对约束数量过少时，相对于初始的成对约束，扩充后的成对约束对聚类结果性能的影响可能微乎其微。

4　基于安全性的成对约束扩充算法

为了能够有效地扩充成对约束，以解决成对约束稀疏性对半监督聚类算法性能的影响，并且可以将扩充后的成对约束应用到不同的半监督聚类算法中，本文提出了一种基于安全性的成对约束扩充算法。该算法通过修改不同传递闭包之间的相似性，以一种安全的方式将它们合并，从而达到扩充成对约束的目的。下面首先给出传递闭包相似性的定义。

定义 1　设 $R=\{r_1, r_2, \cdots, r_m\}$ 表示预先给定的成对约束通过推论 1 得到的 m 个传递闭包集合，$\boldsymbol{S}$ 表示 $m \times m$ 的传递闭包相似性矩阵。两个传递闭包 $r_i, r_j \in R(i \neq j)$ 之间的相似性定义如下：

$$S_{ij}=\exp\left(-\frac{d(r_i, r_j)}{2\sigma 2}\right)$$

其中，σ 表示高斯核参数；$d(r_i, r_j)$ 表示两个传递闭包之间的距离，其计算方式为：

$$d(r_i, r_j)=\min(dist(x_{a(r_i)}, x_{b(r_j)}))$$

其中，$a \in (1, \cdots, n_{r_i})$，$b \in (1, \cdots, n_{r_j})$，$x_{a(r_i)}$ 和 $x_{b(r_j)}$ 表示 r_i 和 r_j 中第 a 个和第 b 个数据对象，$dist(x_{a(r_i)}, x_{b(r_j)})$ 表示两个数据对象之间的距离。

由于传递闭包是通过成对约束的传递性扩展得到的，每个传递闭包内的数据对象都具有相同的类结构，即每个传递闭包都可以看作一个已知的类。而成对约束表示两个数据对象之间的关系，无法根据成对约束来推断数据对象的标签，因此不同的传递闭包也可能属于相同的类。根据定义 1 可以发现，通过计算传递闭包的相似性将相似的传递闭包进行合并，并利用成对约束的性质可以达到扩充约束的目的。同时，我们有效地利用了传递闭包的连通性，将两个传递闭包内数据对象之间的最短距离作为它们之间的距离。为了更好地反映数据对象在样本空间中的差异性，我们使用测地距离作为数据对象之间距离的度量。下文将介绍两个数据对象之间测地距离的定义。

定义 2[31]　设 x_i 和 x_j 为 X 中的两个数据对象，它们之间的测地距离为：

$$dist(x_i, x_j)=Dijkstra(G_{\text{knn}}(x_i, x_j))$$

其中，x_i 和 x_j 之间的测地距离 $dist(x_i, x_j)$ 是通过构建数据对象的 k 近邻图，计算图中两点之间的最短路径距离得到的。与欧氏距离相比，测地距离可以更好地反映两个数据对象之间的全局连通性。

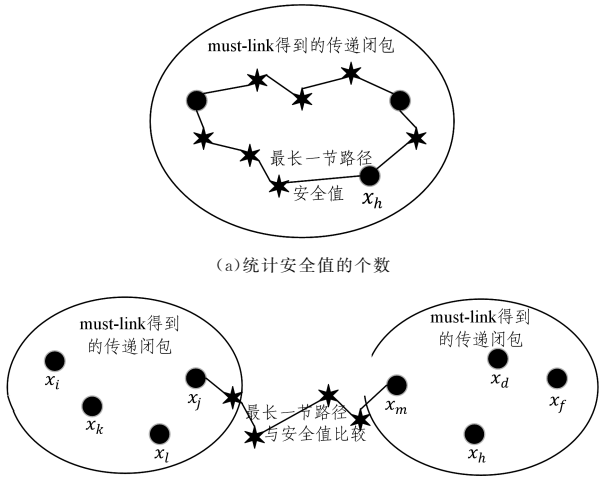
虽然两个传递闭包的距离越近表示它们越相似，但是直接合并两个传递闭包再扩充成对约束可能会存在风险。假如两个传递闭包不属于同一个类结构，则合并闭包并扩充成对约束会产生大量错误的约束，从而影响聚类结果。为了能够安全地合并传递闭包，尽可能减少存在风险的合并，本文使用安全值来判断是否合并传递闭包。下面给出安全值的定义。

定义 3　设 $x_{a(r_i)}$ 和 $x_{b(r_i)}$ 表示传递闭包 r_i 中的第 a 个和第 b 个数据对象。扩充成对约束时使用的安全值定义为：

$$sa=\max\ path(dist(x_{a(r_i)}, x_{b(r_i)}))$$

其中，$a, b \in (1, \cdots, n_{r_i})$，$i \in (1, \cdots, m)$。

根据定义 3 可知，安全值对应所有传递闭包中数据对象之间测地距离中最长一节路径的距离，也可以看作所有传递闭包对应的数据对象之间的最大局部连通距离。将两个传递闭包之间测地距离中最长一节路径的距离与安全值进行比较，如果最长路径的距离高于安全值说明合并两个传递闭包存在一定风险，因此将两个传递闭包之间的距离设为无穷大，否则保留原始测地距离。我们通过修改传递闭包之间的距离来修改相似性，减少传递闭包的合并带来的风险。图 2 展示了安全值的获取和修改传递闭包之间距离的过程。



(a)统计安全值的个数

(b)根据安全值更改传递闭包的相似性

图 2　基于安全性的成对约束扩充

Fig. 2　Extended of pairwise constraints based on security

在得到传递闭包之间的相似性之后，可以通过聚类方法划分传递闭包。由于根据成对约束得到的传递闭包无法判断最终的划分个数，因此本文使用一种代表性的图聚类算法(即模块度算法[32])对传递闭包进行划分，最后根据推论 1 和推论 2，在合并后的每个传递闭包块内所有数据对象之间增加必连约束来扩充必连约束。同时，当两个传递闭包中的数据对象之间存在一个或多个勿连约束时，需在两个传递闭包中所有的对象之间增加勿连约束来扩充勿连约束。因此，基于上述思想，本文提出一种基于安全性的成对约束扩充算法，如算法 1 所示。

算法 1　基于安全性的成对约束扩充算法

输入：X, M, C

输出：$\tilde{M}$, $\tilde{C}$

1. 通过推论 1 得到 M 对应的传递闭包 R；

2. 通过定义 1 计算相似性矩阵 **S**；

3. 通过定义 3 统计安全值 sa；

4. 通过比较任意成对约束之间测地距离最长的一节路径与 sa 来修改 **S**；

5. 在 **S** 上执行图聚得到传递闭包划分$\widetilde{\mathrm{R}}$；

6. 在$\widetilde{\mathrm{R}}$上执行推论 1 和推论 2 得到$\widetilde{\mathrm{M}}$，$\widetilde{\mathrm{C}}$；

7. 返回$\widetilde{\mathrm{M}}$，$\widetilde{\mathrm{C}}$。

5　实验分析

本文在 6 个 UCI 数据集与 2 个图像数据集上，分别将原始成对约束和按照传递性扩充得到的成对约束作为两种基于成对约束的半监督聚类算法的输入进行比较，测试了所提算法的有效性。表 1 列出了所使用的数据集。本文使用了两种被广泛使用的有效性指标来评价所提算法的聚类性能，分别为标准互信息(Normalized Mutual Information，NMI)[33]和调整兰德系数(Adjusted Rand Index，ARI)[34]。这两个指标用于估计数据集对应的真实划分和聚类结果之间的相似性。给定一组具有 N 个数据对象的数据集 X，以及该数据集所对应的两种划分，分别为$\mathbb{C}=\{c_1,c_2,\cdots,c_k\}$(聚类结果)和$\mathbb{P}=\{p_1,p_2,\cdots,p_k\}$(数据集的真实划分)。令$n_{ij}=|c_i\cap p_j|$表示第$c_i$组划分和第$p_j$组划分共有数据对象的数量，同时设$b_i=\sum_{j=1}^{N}n_{ij}$，$d_j=\sum_{i=1}^{N}n_{ij}$。标准互信息[33]的定义为：

$$NMI=\frac{2\sum_i\sum_j n_{ij}\log\frac{n_{ij}N}{b_i d_j}}{-\sum_i b_i\log\frac{b_i}{N}-\sum_j d_j\log\frac{d_j}{N}}$$

调整兰德系数[34]的定义为：

$$ARI=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{b_i}{2}\sum_j\binom{d_j}{2}\right]\Big/\binom{N}{2}}{\frac{1}{2}\left[\sum_i\binom{b_i}{2}+\sum_j\binom{d_j}{2}\right]-\left[\sum_i\binom{b_i}{2}\sum_j\binom{d_j}{2}\right]\Big/\binom{N}{2}}$$

如果聚类结果与真实的划分越相似，则对应的 *NMI* 值和 *ARI* 值越大。

表 1　实验数据集

Table 1　Experiment datasets

Dataset	# Instance	# Features	# Classes
Iris	150	4	3
Wine	178	13	3
Heart	270	13	2
ORL	400	1 024	40
Breast	569	31	2
Bank	1 372	4	2
Isolet	1 560	617	26
USPS	2 000	256	10

为了展示算法的有效性，本文选取经典的基于 K-Means 的成对约束聚类算法[12](Constrained K-means，COP)和基于成对约束的标签传播算法算法[25](Label Propagation Based on Pairwise Constraints，PCLP)作为测试算法。首先，随机抽取数据对象总数的 15%，20% 和 25% 对应数量的成对约束 10 组，其中所抽取的必连约束与勿连约束的数量一致。将原始成对约束(Pairwise Constraints，PC)[30]、根据传递性扩充的成对约束(Pairwise Constraints Transfer Algorithm，PCTA)[28]以及根据本文算法扩充的成对约束(PCES)分别作为先验信息来执行测试算法。最后，通过计算 *NMI* 和 *ARI* 的平均值来展示成对约束扩充算法的有效性。

为了安全地扩充成对约束，本文并没有直接将传递闭包进行合并，而是提出了一个最大局部连通距离作为安全值，降低了合并传递闭包和扩散成对约束过程中的风险。由于测地距离可以更好地反映两个数据对象之间的全局连通性，因此本文选择测地距离作为两点之间的最短路径距离。在实验过程中，我们统计了传递闭包中数据对象之间的测地距离，并将其中最长的一节路径距离作为安全值。此外，我们还统计了传递闭包之间的测地距离中最长一节路径的距离，如果该距离大于安全值，则认为合并两个传递闭包存在一定的风险，否则可以安全合并。实验结果表明，该方法可以安全有效地扩充成对约束。

本文提出的算法参数分别为计算测地距离使用的近邻参数以及构造传递闭包相似性矩阵时的高斯核参数 σ。我们将近邻参数固定为 3，同时将集合$\{s/2,s/5,s/10,s/20,s/30\}$中的值分别作为 σ 参数的值来测试不同 σ 值对应的聚类结果，并选择每个数据集对应最高的 *NMI* 和 *ARI* 值作为比较结果。

表 2 和表 3 列出了不同成对约束在两个测试算法上对应的评价指标。可以看出，当测试算法为 COP 算法与 PCLP 算法时，相比于原始成对约束(PC)，根据传递性扩充得到的成对约束(PCTA)在一些数据集上的聚类结果的提升很少甚至没有。但是，相较于原始的成对约束，通过本文提出的算法扩充后得到的成对约束(PCES)对聚类结果有较为明显的提升，尤其在 Iris，Wine，Heart 和 Bank 数据集上。此外，表 2 和表 3 展示了每个算法在所有测试数据集上的平均聚类结果。

表 2　基于成对约束的半监督聚类算法的 *NMI* 值

Table 2　*NMI* values of semi-supervised clustering algorithms based on pairwise constraints

Datasets	percent/%	COP			PCLP		
		PC	PCTA	PCES	PC	PCTA	PCES
Iris	15	0.7361	0.7361	**0.7515**	0.7642	0.7642	**0.7817**
	20	0.7495	0.7602	**0.7973**	0.7777	**0.7777**	**0.7986**
	25	0.7798	0.7862	**0.8122**	0.7907	0.7924	**0.8154**
Wine	15	0.8130	0.8130	**0.8763**	0.8823	0.8823	**0.8917**
	20	0.8585	0.8707	**0.8915**	0.8635	0.8794	**0.9182**
	25	0.8607	0.8981	**0.9195**	0.8794	0.8926	**0.9050**
Heart	15	0.3255	0.3295	**0.3702**	0.2419	0.2419	**0.2715**
	20	0.3760	0.3863	**0.3911**	0.2449	0.2449	**0.2856**
	25	0.4437	0.4705	**0.4752**	0.2466	0.2551	**0.2923**
ORL	15	0.7232	0.7260	**0.7260**	0.7689	0.7735	**0.7786**
	20	0.7371	0.7375	**0.7402**	0.7799	0.7806	**0.7849**
	25	0.7494	0.7503	**0.7569**	0.7861	0.7883	**0.7914**
Breast	15	0.6788	0.6795	**0.6837**	0.6710	0.6710	**0.7142**
	20	0.6901	0.6932	**0.6977**	0.6823	0.6823	**0.7464**
	25	0.7113	0.7125	**0.7214**	0.6945	0.6945	**0.7490**
Bank	15	0.0286	0.0322	**0.1836**	0.9041	0.9041	**0.9235**
	20	0.0580	0.0659	**0.4374**	0.9317	0.9317	**0.9359**
	25	0.1256	0.1843	**0.5006**	0.9359	0.9359	**0.9686**
Isolet	15	0.7623	0.7631	**0.7685**	0.7897	0.7897	**0.7933**
	20	0.7666	0.7688	**0.7688**	0.7903	0.7930	**0.7930**
	25	0.7762	0.7861	**0.7903**	0.7976	0.7987	**0.8016**
USPS	15	0.6355	0.6380	**0.6433**	0.7622	0.7651	**0.7675**
	20	0.6389	0.6395	**0.6493**	0.7677	0.7677	**0.7732**
	25	0.6480	0.6500	**0.6508**	0.7770	0.7779	**0.7803**
All Datasets	Average	0.6114	0.6199	**0.6668**	0.7304	0.7327	**0.7526**

表 3 基于成对约束的半监督聚类算法的 *ARI* 值

Table 3 *ARI* values of semi-supervised clustering algorithms based on pairwise constraints

Datasets	percent/%	COP			PCLP		
		PC	PCTA	PCES	PC	PCTA	PCES
Iris	15	0.7158	0.7158	**0.7494**	0.7309	0.7309	**0.7514**
	20	0.7611	0.7764	**0.8167**	0.7445	0.7445	**0.7927**
	25	0.7883	0.8022	**0.8339**	0.7811	0.7856	**0.8105**
Wine	15	0.8413	0.8413	**0.9058**	0.9134	0.9134	**0.9215**
	20	0.8727	0.8892	**0.9186**	0.8894	0.9062	**0.9390**
	25	0.8818	0.9168	**0.9337**	0.9062	0.9149	**0.9303**
Heart	15	0.4087	0.4136	**0.4681**	0.3061	0.3061	**0.3490**
	20	0.4780	0.4882	**0.4935**	0.3103	0.3103	**0.3668**
	25	0.5472	0.5749	**0.5806**	0.3133	0.3186	**0.3713**
ORL	15	0.3156	0.3214	**0.3214**	0.4557	0.4583	**0.4605**
	20	0.3590	0.3594	**0.3608**	0.4575	0.4650	**0.4732**
	25	0.3894	0.3902	**0.4026**	0.4690	0.4764	**0.4817**
Breast	15	0.7839	0.7858	**0.7901**	0.7607	0.7607	**0.7856**
	20	0.7990	0.8011	**0.8053**	0.7668	0.7668	**0.8372**
	25	0.8181	0.8197	**0.8245**	0.7730	0.7730	**0.8438**
Bank	15	0.0392	0.0440	**0.2493**	0.9425	0.9425	**0.9653**
	20	0.0801	0.0912	**0.5458**	0.9624	0.9624	**0.9653**
	25	0.1669	0.2106	**0.6056**	0.9653	0.9653	**0.9855**
Isolet	15	0.5291	0.5295	**0.5389**	0.5863	0.5863	**0.5899**
	20	0.5268	0.5284	**0.5375**	0.5964	0.5992	**0.5994**
	25	0.5548	0.5688	**0.5803**	0.6094	0.6103	**0.6156**
USPS	15	0.5533	0.5548	**0.5570**	0.6350	0.6382	**0.6398**
	20	0.5583	0.5624	**0.5738**	0.6388	0.6397	**0.6433**
	25	0.5844	0.5872	**0.5877**	0.6874	0.6881	**0.6951**
All Datasets	Average	0.5564	0.5655	**0.6242**	0.6751	0.6776	**0.7002**

上述实验结果表明，相比于根据传递性扩充成对约束的方法，本文提出的算法能够安全且有效地扩充成对约束，并且可以将扩充后的成对约束应用到不同的半监督聚类算法中。我们还发现，相比于根据传递性的扩充方法，本文的扩充方法在少部分数据集上对聚类结果的提升不是特别明显，但是其优势在于能够在低风险的情况下扩充成对约束，而不是一味地扩充成对约束。

结束语 本文提出了一种基于安全性的成对约束扩充算法。该算法统计所有传递闭包内不同数据对象之间的测地距离以找到安全值(最大局部连通距离)，通过比较不同传递闭包之间测地距离的最大路径值与安全值来修改传递闭包间的相似性，尽可能地减少传递闭包合并的风险，最终通过成对约束的性质达到扩充成对约束的目的。该算法不仅可以安全有效地扩充成对约束，同时可以将扩充后的成对约束应用到不同的半监督聚类算法中。最后，我们通过大量的实验证明了该算法的有效性。

若先验信息是粗粒度的，即不是数据对象的精确标签，这类先验信息往往无法直接作为约束来执行半监督聚类算法。如何能够在粗粒度先验信息的情况下执行标签传播将是今后研究的一个方向。

参 考 文 献

[1] WU X, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge andInformation Systems, 2008, 14(1): 1-37.

[2] JAIN A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[3] GALLEGO A J, CALVO-ZARAGOZA J, VALERO-MAS J J, et al. Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation[J]. Pattern Recognition, 2018, 74: 531-543.

[4] PATERLINI S, KRINK T. Differential evolution and particle swarm optimisation in partitional clustering[J]. Computational-Statistics & Data Analysis, 2006, 50(5): 1220-1247.

[5] LU Y, WAN Y. PHA: A fast potential-based hierarchical agglomerative clustering method[J]. Pattern Recognition, 2013, 46(5): 1227-1239.

[6] KUMAR K M, REDDY A R M. A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method[J]. Pattern Recognition, 2016, 58: 39-48.

[7] DINLER D, TURAL M K. Robust semi-supervised clustering with polyhedral and circular uncertainty[J]. Neurocomputing, 2017, 265: 4-27.

[8] ZHOU X, BELKIN M. Semi-supervised learning[J]. Academic Press Library in Signal Processing, 2014, 1: 1239-1269.

[9] ZHOU D, BOUSQUET O, LAL T N, et al. Learning with local and global consistency[J]. Advances in Neural Information Processing Systems, 2003, 16(3): 321-328.

[10] WANG H, LI T, LI T, et al. Constraint neighborhood projections for semi-supervised clustering[J]. IEEETransactions on Cybernetics, 2014, 44(5): 636-643.

[11] ZHOU Z H. Ensemble methods : foundations and algorithms [M]. Boca Raton: CRC Press, 2012: 72-73.

[12] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained k-means clustering with background knowledge[C]// Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 577-584.

[13] YANG Y, TAN W, LI T, et al. Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems[J]. Knowledge-Based Systems, 2012, 32: 101-115.

[14] WEI S, LI Z, ZHANG C. A semi-supervised clustering ensemble approach integrated constraint-based and metric-based[C]// Proceedings of the 7th International Conference on Internet Multimedia Computing and Service. New York: Association for Computing Machinery, 2015: 1-6.

[15] LI T, DING C, JORDAN M I. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization[C]// Seventh IEEE International Conference on Data Mining (ICDM 2007). Piscataway: IEEE, 2007: 577-582.

[16] REN Y, HU X, SHI K, et al. Semi-supervised denpeak clustering with pairwise constraints[C]// Pacific Rim International Conference on Artificial Intelligence. Berlin: Springer, 2018: 837-850.

[17] MIYAMOTO S, TERAMI A. Constrained agglomerative hierarchical clustering algorithms with penalties [C]// 2011 IEEE

International Conference on Fuzzy Systems. Piscataway：IEEE，2011：422-427.

[18] KAMVAR K，SEPANDAR S，KLEIN K，et al. Spectral learning [C]// International Joint Conference of Artificial Intelligence. Stanford Info Lab，2003：561-566.

[19] XU Q，DESJARDINS M，WAGSTAFF K. Constrained spectral clustering under a local proximity structure assumption[C]// Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference. Clearwater Beach，Florida，USA：DBLP，2005：866-867.

[20] JI X，XU W. Document clustering with prior knowledge[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：Association for Computing Machinery，2006：405-412.

[21] WANG F，DING C，LI T. Integrated KL (K-means－Laplacian) clustering：a new clustering approach by combining attribute data and pairwise relations[C]// Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. 2009：38-48.

[22] KULIS B，BASU S，DHILLON I，et al. Semi-supervised graph clustering：a kernel approach [J]. MachineLearning，2009，74(1)：1-22.

[23] DING S，JIA H，DU M，et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model[J]. Information Sciences，2018，429：215-228.

[24] KLEIN D，KAMVAR S D，MANNING C D. From instance-level constraints to space-level constraints：Making the most of prior knowledge in data clustering[C]// International Conference on Machine Learning. Stanford：IMLS，2002：307-314.

[25] LU Z，PENG Y. Exhaustive and efficient constraint propagation：A graph-based learning approach and its applications[J]. International Journal of Computer Vision，2013，103(3)：306-325.

[26] WANG L，BO L F，JIAO L C. Density-sensitive semi-supervised spectral clustering[J]. Journal of Software，2007，18(10)：2412-2422.

[27] WEI S，LI Z，ZHANG C. Combined constraint-based with metric-based in semi-supervised clustering ensemble[J]. International Journal of Machine Learning and Cybernetics，2018，9(7)：1085-1100.

[28] YU Z，KUANG Z，LIU J，et al. Adaptive ensembling of semi-supervised clustering solutions[J]. IEEE Transactions on Knowledge and Data Engineering，2017，29(8)：1577-1590.

[29] EISENBERG D，MARCOTTE E M，XENARIOS I，et al. Protein function in the post-genomic era [J]. Nature，2000，405(6788)：823-826.

[30] WAGSTAFF K，CARDIE C. Clustering with instance-level constraints[J]. AAAI/IAAI，2000，1097：577-584.

[31] TENENBAUM J B，DE SILVA V，LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science，2000，290(5500)：2319-2323.

[32] CLAUSET A，NEWMAN M E J，MOORE C. Finding community structure in very large networks[J]. Physical Review E，2004，70(6)：066111.

[33] DOM B E. An information-theoretic external cluster-validity measure[J]. Uncertainty in Artificial Intelligence，2001，27(3)：137-145.

[34] HUBERT L，ARABIE P. Comparing partitions[J]. Journal of Classification，1985，2(1)：193-218.

YANG Fan，born in 1995，postgraduate. Her main research interests include semi-supervised learning and so on.

BAI Liang，born in 1982，Ph.D，professor，is a member of China Computer Fedaration. His main research interest include cluster analysis and so on.