基于成对约束的半监督凝聚层次聚类算法

盛俊杰，谢丽聪
（福州大学 数学与计算机学院，福建 福州 350108）

摘 要：半监督聚类就是利用样本的监督信息来帮助提升无监督学习的性能。在半监督聚类中，成对约束（must-link 约束和 cannot-link 约束）作为样本的先验知识被广泛地使用。凝聚层次聚类（AHC）也叫合成聚类，是层次聚类法的一种。提出了一种基于成对约束的半监督凝聚层次聚类算法（PS-AHC），该算法利用成对约束来改变聚类簇之间的距离，使聚类簇之间的距离更真实。在 UCI 数据集上的实验表明，PS-AHC 能有效地提高聚类的准确率，是一种有前景的半监督聚类算法。

关键词：半监督聚类；成对约束；凝聚层次聚类

中图分类号：TP18 **文献标识码**：A **文章编号**：1674-7720(2012)24-0067-03

Semi-supervised agglomerative hierarchical clustering based pairwise constraints

Sheng Junjie, Xie Licong
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: Semi-supervised clustering uses the samples' supervised information to aid unsupervised learning. In the semi-supervised clustering, pairwise constraints information (must-link constraints and cannot-link constraints) are widely used as samples' prior knowledge. Agglomerative hierarchical clustering (AHC) is one kind of hierarchical clustering. This paper presents a semi-supervised agglomerative hierarchical clustering algorithm based on pairwise constraints (PS-AHC). The algorithm uses pairwise constraints to change distances of clusters. It makes distances of clusters closer to the truth. The results of experiments on the UCI data sets confirm that PS-AHC algorithm can improve the accuracy of clustering effectively and that it is a promising semi-supervised clustering algorithm.

Key words: semi-supervised clustering; pairwise constraints; agglomerative hierarchical clustering

聚类即根据数据集中数据的不同特征将其划分为不同簇的过程，使得同一个簇中的样本之间具有较高的相似度，而不同簇中的样本之间具有高度相异度。聚类过程中通常没有类别标签等监督信息，因而是一种无监督的学习。传统的无监督学习通常只利用无标签样本进行学习，而监督学习只利用有标签样本进行学习，半监督学习的优越性体现在其同时利用无标签样本和有标签样本进行学习。半监督聚类算法研究如何利用少量的监督信息来提升聚类性能[1]，使用的监督信息既可以是类标签，也可以是一对样本是否属于同一类的约束信息。半监督聚类算法对聚类性能的提高主要依赖于监督信息，监督信息的选取非常关键。

对于现实世界的无监督学习算法，例如人的语音识别、GPS 道路检测等，以成对约束形式出现的监督信息更实际。对用户而言，要确定样本类标签比较困难，但是获得关于两个样本是否属于同一类的约束信息则较为容易[2]。其中涉及两类成对点约束，分别是 must-link 和 cannot-link。

与监督学习相比，无监督聚类过程缺少用户或分类器（如类标签信息）的指导，因此不能产生理想的聚类结果。使用某种监督形式，例如成对约束，可以显著提高无监督聚类的质量。本文将监督信息的信息含量应用到聚类中，提出一种基于成对约束的半监督凝聚层次聚类算法。

1 成对约束概念

半监督聚类使用的成对约束表示两个样本一定被分到同一个簇或者一定被分到不同的簇。两个广泛使用的成对约束方法是 must-link 约束和 cannot-link 约束，其中，must-link 约束表示两个样本一定被分配到同一个簇，cannot-link 约束代表两个样本一定被分到不同的簇[3]。令 $Con(i,j)$ 表示样本 x_i 和样本 x_j 之间的成对约束，如下

表示：

$$\mathrm{Con}(i,j)=\begin{cases}1 & x_i \text{ 和 } x_j \text{ 一定属于同一个簇}\\ -1 & x_i \text{ 和 } x_j \text{ 一定属于不同的簇}\\ 0 & \text{其他}\end{cases} \tag{1}$$

很明显地可以看出，如果 $\mathrm{Con}(i,j)=1$，则 $\mathrm{Con}(j,i)=1$；如果 $\mathrm{Con}(i,j)=-1$，则 $\mathrm{Con}(j,i)=-1$。

2 凝聚层次聚类算法(AHC)

层次聚类方法是根据给定的簇间距离度量准则，构造和维护一棵由簇和子簇形成的聚类树，直至满足某个终结条件为止。根据层次分解是自底向上还是自顶向下形成，层次聚类方法可以分为凝聚的(Agglomerative)和分裂的(Divisive)两种[4]。一个纯粹的层次聚类方法的聚类质量受限于如下的特点：一旦一个合并或分裂被执行，就不能修正。

凝聚层次聚类 AHC(Agglomerative Hierarchical Clustering)采用自底向上的策略，首先将每个样本作为一个簇，然后合并这些原子簇为越来越大的簇，直至所有的样本都在一个簇中，或者满足某个终结条件。绝大多数层次聚类方法都属于这一类，它们只是在簇间距离的定义上有所不同。图 1 是凝聚层次聚类的一个简单例子。

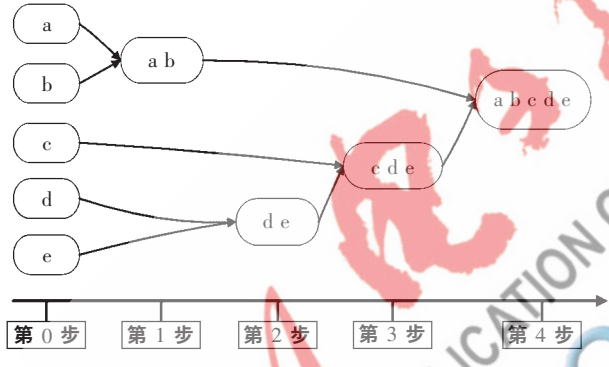

图 1 凝聚层次聚类举例

本文凝聚层次聚类的簇间距离的度量采用了中心点的方法。设 C 是一个聚类簇，x_k 是 C 中的样本，则 C 的中心点为：

$$M(C)=\frac{1}{|C|}\sum_{x_k\in C}x_k \tag{2}$$

聚类簇 C_1 和聚类簇 C_2 的距离为：

$$d(C_1,C_2)=||M(C_1)-M(C_2)||^2 \tag{3}$$

AHC 算法的步骤如下：

输入：未知分布的样本集 $S=\{x_1,x_2,\cdots,x_N\}$

输出：最优的聚类分组 $\{C_1,C_2,\cdots,C_K\}$

(1)假设初始聚类分组为：$g=\{C_1,C_2,\cdots,C_N\}$，初始的聚类簇个数为 $Y=N$，计算所有的距离 $d(C,C')$，其中 C、$C'\in g$。

(2)寻找距离最近的两个聚类簇：$(C_p,C_q)=\arg\min_{C、C'\in g} d(C,C')$。$C_p$ 和 C_q 合并：$C_r=C_p\cup C_q$。把 C_r 添加到 g 中，同时将 C_p 和 C_q 从 g 中删除。$Y=Y-1$。如果 $Y=K$，算法停止，输出结果；如果 $Y>K$，算法跳转到步骤(3)。

(3)计算所有的距离 $d(C_r,C'')$，其中 $C''\in g$。跳转到步骤(2)。

3 基于成对约束的半监督凝聚层次聚类算法

在 PS-AHC 中，P 代表成对约束 Pairwise Constraints，S 代表半监督 Semi-supervised。

3.1 近邻度的定义

本文提出了近邻度这个新的概念，其思想是基于 k 近邻分类算法的。k 近邻分类算法的思想是：找出距离待分类样本最近的 k 个有标记样本，在这 k 个有标记样本中，哪个类别的样本占的数目最多，待分类样本就属于哪个类别。在 KNN 算法中，待分类样本的类别由它附近最近的 k 个样本决定。

对于每一个样本 x_i，都有一个近邻度 α_i，$\alpha_i\geqslant 0$，α_i 的定义如下：

$$\alpha_i=\frac{1}{k}\sum_{m=1}^{k}||x_i-x_m|| \tag{4}$$

其中，k 是给定的参数，x_m 是距离样本 x_i 最近的 k 个样本。

如图 2 所示，有 x_1、x_2 和 x_3 三个样本，圆的半径代表相应样本的近邻度。近邻度大，说明该样本附近的样本分布比较稀疏，样本之间的距离比较远；反之，近邻度小，说明该样本附近的样本分布比较密集，样本之间的距离比较近。

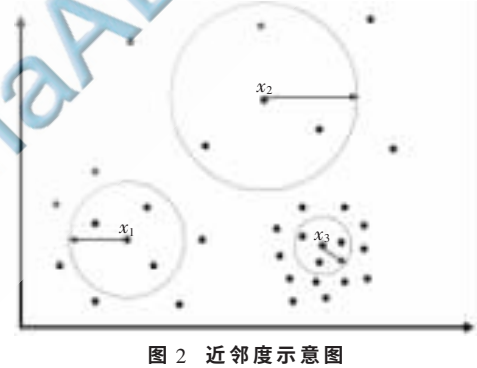

图 2 近邻度示意图

3.2 利用成对约束改变聚类簇之间的距离

首先，定义两个集合 $ML(C;x)$ 和 $CL(C;x)$。$ML(C;x)$ 指在聚类簇 C 中与样本 x 具有 must-link 约束关系的样本的集合，$CL(C;x)$ 指在聚类簇 C 中与样本 x 具有 cannot-link 约束关系的样本的集合，表示如下：

$$ML(C;x)=\{\varepsilon|\varepsilon\in C,(\varepsilon,x)\in ML\}$$

$$CL(C;x)=\{\varepsilon|\varepsilon\in C,(\varepsilon,x)\in CL\}$$

其次，在 $ML(C;x)$ 和 $CL(C;x)$ 的基础上定义集合的并 $ML(C;C')$ 和 $CL(C;C')$：

$$ML(C;C')=\bigcup_{x\in C'}ML(C;x)$$

$$CL(C;C')=\bigcup_{x\in C'}CL(C;x)$$

最后，用 $K(C;C')$ 表示所有 $ML(C;C')$ 和 $CL(C;C')$

的近邻度之差：

$$K(C;C')=\sum_{x_k\in ML(C;C')}\alpha_k-\sum_{x_l\in CL(C;C')}\alpha_l \tag{5}$$

其中，$\sum_{x_k\in ML(C;C')}\alpha_k$ 的实际意义是代表聚类簇 C 和 C' 的 must-link 约束程度，而 $\sum_{x_l\in CL(C;C')}\alpha_l$ 的实际意义是代表聚类簇 C 和 C' 的 cannot-link 约束程度。当 $K(C;C')>0$ 时，则称 C 是 must-link 约束于 C'；当 $K(C;C')<0$ 时，则称 C 是 cannot-link 约束于 C'。值得注意的是，$K(C;C')$ 不具有对称性，即 $K(C;C')\neq K(C';C)$。

例如，$C=\{x_1,x_2,x_3\}$ 和 $C'=\{x_4,x_5,x_6\}$ 是两个聚类簇，(x_1,x_4) 和 (x_3,x_5) 是 must-link 约束，(x_2,x_6) 是 cannot-link 约束。那么，$ML(C;C')$、$CL(C;C')$ 和 $K(C;C')$ 的计算结果如下：

$ML(C;C')=\{x_1,x_3\}$

$CL(C;C')=\{x_2\}$

$K(C;C')=\alpha_1+\alpha_3-\alpha_2$

同样，$ML(C';C)$、$CL(C';C)$ 和 $K(C';C)$ 的计算结果如下：

$ML(C';C)=\{x_4,x_5\}$

$CL(C';C)=\{x_6\}$

$K(C';C)=\alpha_4+\alpha_5-\alpha_6$

有了 $K(C;C')$ 的定义，聚类簇 C_1 和聚类簇 C_2 的距离被调整为：

$$d(C_1,C_2)=\begin{cases}\left[||M(C_1)-M(C_2)||-\dfrac{K(C_1;C_2)}{|C_1|}-\dfrac{K(C_2;C_1)}{|C_2|}\right]^2 \\ \text{当}||M(C_1)-M(C_2)||>\dfrac{K(C_1;C_2)}{|C_1|}+\dfrac{K(C_2;C_1)}{|C_2|}\text{时} \\ 0 \qquad \text{其他}\end{cases} \tag{6}$$

其中，$|C_1|$ 和 $|C_2|$ 分别表示 C_1 的样本数和 C_2 的样本数。

PS-AHC 算法的步骤如下：

输入：未知分布的样本集 $S=\{x_1,x_2,\cdots,x_N\}$ 和样本的成对约束信息 $\mathrm{Con}(i,j)$

输出：最优的聚类分组 $C=\{C_1,C_2,\cdots,C_K\}$

(1)假设初始聚类分组为：$g=\{C_1,C_2,\cdots,C_N\}$，初始的聚类簇个数为 $Y=N$，利用式(5)计算所有的 $K(C;C')$，C、$C'\in g$，再利用式(6)计算所有的距离 $d(C,C')$，C、$C'\in g$。

(2)寻找两个距离最近的聚类簇：$(C_p,C_q)=\arg\min_{C、C'\in g} d(C,C')$。$C_p$ 和 C_q 合并：$C_r=C_p\cup C_q$。把 C_r 添加到 g 中，同时把 C_p 和 C_q 从 g 中删除。$Y=Y-1$。如果 $Y=K$，算法停止，输出结果；如果 $Y>K$，算法跳转到步骤(3)。

(3)利用式(5)计算所有的 $K(C;C')$，C、$C'\in g$，再利用式(6)计算所有的距离 $d(C,C')$，C、$C'\in g$。跳转到步骤(2)。

4 实验结果与分析

为了验证本文提出的 PS-AHC 算法的有效性，对 AHC 和 PS-AHC 这两个算法进行了对比实验。从 UCI 数据集[5]上选择了 5 个完整的数据集，分别是 haberman、balance-scale、iris、tae(teaching assistant evaluation)和 pid(pima indians diabetes)。在每个数据集 S 中，随机地选择一些样本对，对这些样本对生成 must-link 约束和 cannot-link 约束。成对约束的数量设置为总样本集数量的 3 倍，即为 $3\times|S|$。所有算法各运行 30 次，取平均的聚类准确率，实验结果对比如表 1 所示。

表 1 AHC 和 PS-AHC 的聚类准确率比较

UCI 数据集	AHC 聚类准确率/%	PS-AHC 聚类准确率/%
haberman	73.5	85.2
balance-scale	63.5	91.8
iris	92.6	99.3
tae	36.4	79.5
pid	65.1	88.4

从实验结果可以看出，PS-AHC 表现出了比 AHC 更优越的性能。这是因为 PS-AHC 引进了样本的成对约束信息。PS-AHC 利用成对约束信息改变聚类簇之间的距离，使有 must-link 约束的两个聚类簇的距离变得更近，而有 cannot-link 约束的两个聚类簇的距离变得更远，从而改变层次聚类的树结构。实验结果同时表明，本文所提出的近邻度概念是有效的。

成对约束是样本的一种监督信息。本文利用成对约束来指导聚类过程，提出了一种基于成对约束的半监督凝聚层次聚类算法(PS-AHC)。PS-AHC 利用成对约束来改变聚类簇之间的距离，使聚类簇之间的距离更真实。在 UCI 数据集上的实验表明，PS-AHC 能有效地提高聚类的准确率，是一种有前景的半监督聚类算法。

参考文献

[1] BILENKO M, BASU S, MOONEY R J. Integrating constraints and metric learning in semi-supervised clustering[C]. Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning. New York: ACM Press, 2004: 81-88.

[2] 李昆仑，曹峥，曹丽萍，等. 半监督聚类的若干新进展[J]. 模式识别与人工智能，2009，22(5)：735-742.

[3] BASU S, BANERJEE A, MOONEY R J. Active semi-supervision for pairwise constrained clustering[C]. Proc. of the SIAM Int'l Conf. on Data Mining. Cambridge: MIT Press, 2004: 333-344.

[4] Han Jiawei, KAMBER M. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2004：1-262.

[5] NEWMAN D J, HETTICH S, BLAKE C L, et al. UCI repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

(收稿日期：2012-07-31)

作者简介：

盛俊杰，男，1988 年生，硕士研究生，主要研究方向：数据挖掘，数据集成。

谢丽聪，女，1964 年生，副教授，主要研究方向：数据挖掘，数据集成。