基于约束动态更新的半监督层次聚类算法

周晨曦 [1] 梁 循 [1] 齐金山 [1, 2]

摘 要 提出了一种基于约束动态更新的半监督层次聚类算法. 与现存的半监督层次聚类算法类似, 该算法也使用了必连和不连约束. 但不同的是, 该算法并不是在对满足必连约束的数据样本点进行预先划分的基础上依据不连约束进行聚合操作, 而是首先将约束扩展为一个闭包, 然后在这此基础上直接依据不连约束进行聚合操作, 并在聚合的过程中依据聚类结果动态地更新必连和不连约束, 以保证最终的聚类结果同时满足必连和不连约束. 该算法的优势在于省略了对必连约束的数据样本点进行预先划分的步骤, 这一改进能够保证数据样本点获得更为合理的聚合顺序, 从而得到更为准确的聚类结果. 本文具体给出了该算法基于 Ward 层次聚类算法的实现, 提出了 C-Ward 算法. 实验表明, 与其他同类算法相比, 无论是在人工模拟数据集还是在现实数据集上, 本文提出的算法都表现出了更高的准确性和更强的稳定性.

关键词 半监督聚类, 层次聚类, 约束, 动态更新, Ward 算法

引用格式 周晨曦, 梁循, 齐金山. 基于约束动态更新的半监督层次聚类算法. 自动化学报, 2015, **41**(7): 1253−1263

DOI 10.16383/j.aas.2015.c140859

A Semi-supervised Agglomerative Hierarchical Clustering Method Based on Dynamically Updating Constraints

ZHOU Chen-Xi[1] LIANG Xun[1] QI Jin-Shan[1, 2]

Abstract A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints is proposing in this research. Following the existing semi-supervised clustering algorithm, this method uses the must-link and cannot-link constraints. Instead of using the idea that the instances with must-link constraints are pre-clustered before agglomerating with the others, this method employs a more general and reasonable process. Firstly, must-link and cannot-link constraints are expanded to compose a constraints closure. Then, a standard agglomeration instructed by cannot-link constraints is processed. During this procedure, the must-link and cannot-link are dynamically updated according to the intermediate clustering results. This updating process guarantees the validity of the final results. The fundamental advantage of this method is omitting the pre-clustering process of the instances with must-link constraints. This modification ensures that data points gain a more reasonable agglomeration order, which may result in a significant improvement on the clustering results. This research also introduces an implementation of this model based on Ward′s method, leading to the C-Ward algorithm. The experimental analyses on both artificial simulated datasets and real world datasets show that this method is much better than the others.

Key words Semi-supervised clustering, agglomerative hierarchical clustering, constraints, dynamically updating, Ward′s method

Citation Zhou Chen-Xi, Liang Xun, Qi Jin-Shan. A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints. *Acta Automatica Sinica*, 2015, **41**(7): 1253−1263

无监督学习和监督学习是机器学习领域中发展最为成熟、应用最为广泛的两类学习模型. 无监督学习算法一般较为简单, 易于实现, 并且基本不需要任何关于数据集的先验知识, 因而其可应用的范围十分广泛. 但正是由于算法的思想简单, 无先验知识, 其聚类结果有时不太理想, 特别是对于较高维度的数据集. 监督学习算法则通过对带有标记的数据样本的学习能够在一定程度上获得一个泛化能力较强的模型. 然而, 这类算法一般需要大量的标记样本数据集作为训练集, 并且要求训练数据集在整个样本空间的分布比较均匀, 否则学习得到的很可能是一个有偏的分类器, 因而其泛化能力依然不够强. 由于获取数据样本标签的成本较高, 监督学习对于有

收稿日期 2014-12-12 录用日期 2015-03-20
Manuscript received December 12, 2014; accepted March 20, 2015
国家自然科学基金 (71271211), 北京市自然科学基金 (4132067), 中国人民大学品牌计划 (10XNI029) 资助
Supported by National Natural Science Foundation of China (71271211), National Natural Science Foundation of Beijing (4132067), and Brand Plan of Renmin University of China (10XNI029)
本文责任编委 封举富
Recommended by Associate Editor FENG Ju-Fu
1. 中国人民大学信息学院 北京 100872 2. 淮阴师范学院计算机科学与技术学院 淮安 223300
1. School of Information, Renmin University of China, Beijing 100872 2. School of Computer Science and Technology, Huaiyin Normal University, Huaian 223300

标记样本数据集的要求在现实情况中很难满足. 在这种情况下, 半监督学习成为一个较为理想的选择.

半监督学习介于无监督学习和监督学习之间, 使用少量的有标记数据样本来辅助学习, 并且对于这些有标记样本的分布并无特别的要求[1−2]. 半监督学习算法一般基于无监督学习算法或监督学习算法, 试图通过这些少量的有标记数据样本来完成对应的学习过程, 以提升这些算法的性能. 实际上, 甚至可以认为无监督学习和监督学习只是半监督学习的两种特例, 前者是有标记数据样本数量为零的情况而后者则是有标记数据样本数量足够多的情况. 理论研究表明, 即使是少数的有标记样本数据也能大幅度提升对应的无监督学习或监督学习模型的效果[3−5].

依据主要是基于无监督学习还是监督学习的思想, 半监督学习算法可以划分为两大类: 半监督聚类算法和半监督分类算法. 半监督聚类算法在无监督学习中引入了点对间的约束 (Pair-wise constraints), 一般是必连约束 (Must-link) 和不连约束 (Cannot-link)[6]. 满足必连约束的两个数据样本点必须被划分为同一个类别, 而满足不连约束的两个数据样本点则必须被划分为不同的类别, 这一类算法主要包括 COP-KMEANS 算法[6]、C-DBSCAN 算法[7]、半监督层次聚类算法[8]、集成式的半监督学习模型[9−10] 等. COP-KMEANS 算法在 K-means 算法中加入了上述两类约束, 类似的基于 K-means 算法的半监督学习算法还包括的 PCK-means 算法[11]、CMWK-means 算法[12]、基于成对约束的判别性半监督聚类方法[13] 等. C-DBSCAN 算法在 DBSCAN 算法中加入了上述两类约束. 在半监督层次聚类算法中, 除了上述提及的两类约束, 学者们还引入了基于距离的两类新约束, 即类别内最大距离约束 (ε-constraints) 和类别间最小距离约束 (δ-constraints).

半监督分类算法的种类也非常多, 其主要的两种思路是在监督学习算法的学习过程中选择置信度较高的数据样本点逐步加入训练数据集进行重复训练以指导学习过程, 以及利用少量的有标记的数据样本进行参数估计[14]. 代表算法包括生成式模型[15−19]、低密度分隔算法[20−24]、半监督图算法[25−28]、自学习和协同学习[29−33]. 生成式模型算法将整个数据样本空间的分布假设为一个混合模型, 一般为混合高斯分布, 然后利用有标记的样本数据进行参数估计. 低密度分隔算法将支持向量机扩展为半监督学习模型, 要求分类超平面不仅是有标记样本的最大间隔分割面, 同时还尽量是未标记的数据样本的最大间隔分割面. 半监督图算法将数据集合转化为图, 从而半监督学习问题也对应转化为图分割问题. 自学习和协同学习算法迭代式地学习一个或多个分类器, 在每次迭代的过程中选择置信度较高的数据样本点加入训练数据集, 重复这一过程直到算法收敛.

本文对层次聚类提出了一个新的半监督化的算法. 在文献 [8] 的半监督层次聚类算法中, 作者首先将满足必连约束的数据样本分别聚合为一类, 然后再和其他的数据样本进行聚合. 由于将其中一部分数据样本点首先进行聚合会改变这些样本点与其他样本点之间的关系, 因而这种聚合方式实际上在很大程度上影响了整个聚合过程. 与之不同, 本文的算法并不是在对满足必连约束的数据样本点进行预先划分的基础上依据不连约束进行聚合操作, 而是首先将约束扩展为一个闭包, 然后在此基础上直接依据不连约束进行聚合操作, 在层次聚类的过程中依据聚类的中间结果动态更新必连约束和不连约束, 以保证最终的聚类结果同时满足必连和不连约束. 聚合过程与没有约束的层次聚类算法基本相同. 在这种更新机制下, 类别的聚合顺序与没有约束时的类别聚合顺序能够保持较高程度的一致性, 从而能够得到更为理想的聚类结果. 该算法的优势在于省略了对必连约束的数据样本点进行预先划分的步骤, 这一改进能够保证数据样本点获得更为合理的聚合顺序, 从而得到更为准确的聚类结果. 本文的主要贡献就是将原来的必连约束和不连约束所隐含的样本点之间的是否属于同一类的关系进行了显性化表达, 使算法不需要在最开始就处理不连约束, 从而保证了层次聚类过程能够按照样本点间距离的远近进行, 从而能够改进以往算法中存在的仅利用初始必连约束和不连约束带来的不准确性. 这种限制条件的动态更新机制引申出了一种全新的半监督层次聚类模式. 在人工模拟数据集和现实数据集上的实验都表明本文提出的算法比文献 [8] 的方法更为有效, 而且比同类算法 COP-KMEANS 和 C-DBSCAN 也更具优势.

本文第 1 节简单介绍了半监督层次聚类算法, 并分析了文献 [8] 中的方法所存在的问题. 第 2 节详细介绍了本次研究提出的半监督层次聚类算法, 并将该算法运用于 Ward 层次聚类算法上, 提出了 C-Ward. 第 3 节是相关的实验设计和结果, 将 C-Ward 算法与文献 [8] 中的方法、COP-KMEANS 以及 C-DBSCAN 算法做了对比分析. 最后一节是对本文的简要总结和展望.

1 半监督层次聚类算法

层次聚类算法最开始将所有的样本数据点分别作为一类, 然后从中选择距离最近的两个类别将其聚合形成一个新的类别, 并删除被选择的两个类别.

算法重复这一聚合过程直到终止条件被满足. 层次聚类算法的一般流程见算法 1. 它们的聚类结果能够使用一个树状图来表示. 主要的层次聚类算法包括 Ward 算法[34]、BIRCH 算法[35]、CURE 算法[36]、CHAMELEON 算法[37] 等.

算法 1. 标准层次聚类算法的一般流程

输入. 数据样本集 X.

输出. 表示聚类结果的树状图.

步骤 1. 初始化类别. 将数据样本点各自作为一类, 即 $C_i = \{x_i\}, \forall i$. 并构建树状图 Dendrogram.

步骤 2. 计算任意两个类别间的距离, 即 $d_{i,j} = D(C_i, C_j), \forall i, j$.

步骤 3. for k =1 to $|S|$　　//$|S|$ 表示数据集 X 的样本数量.

1) 寻找距离最小的两个类别 C_a 和 C_b, 即 a, b = $\mathrm{argmin}_{i,j}\, d(i,j)$.

2) 将类别 C_a 和 C_b 合并为新的类别 C_n, 并计算 C_n 与除 C_a 和 C_b 之外的其他类别之间的距离.

3) 删除类别 C_a 和 C_b, 加入类别 C_n, 并构建树状图 Dendrogram.

步骤 4. 返回树状图 Dendrogram.

1.1　半监督层次聚类算法

Davidson 等将半监督的思想运用于层次聚类算法, 提出了半监督层次聚类算法[8]. 该算法同样也利用了必连和不连两种约束. 算法首先对所有涉及到必连约束的数据样本点集合划分类别, 得到 M_1, $M_2, \cdots, M_r$. 其具体方法见文献[38]. 对于不涉及必连约束的数据样本集合, 即 $\widetilde{X} = X - \cup_{i=1}^{r} M_i$, 每一个数据样本点各自作为一个类别, 即 $C_i = \{x_i\}, \forall i \in \widetilde{X}$. 从而, 算法初始化的类别为 $M_1, M_2, \cdots, M_r, C_1, C_2, \cdots, C_{|\widetilde{X}|}$. 算法然后在此初始化的基础上运行一个层次聚类算法. 与标准层次聚类算法唯一不同的是, 在步骤 3 中寻找距离最小的两个类别 C_a 和 C_b 时, C_a 和 C_b 中的所有数据样本点要求不存在满足不连约束的点对, 否则转而寻找距离次小的两个类别, 以此类推直到寻找到满足条件的两个类别进行合并. 如果不存在这样的两个类别, 则算法终止. 这里需要说明的是, 作者还引入了基于距离的类别内最大距离约束 (ε-constraints) 和类别间最小距离约束 (δ-constraints), 而在本次研究中不考虑这两种约束.

1.2　存在的问题

Davidson 等的半监督层次聚类算法首先将涉及必连约束的数据样本点进行划分的做法, 对于整个层次聚类过程的影响是比较大的. 具体来说, 层次聚类算法的基本思想是, 迭代地选择距离最近的两个类别进行合并, 而最开始具有必连约束的两个数据样本点之间的距离并不一定是最近的; 相反, 它们之间的距离甚至可能非常远. 而将它们合并之后产生的新的类别与其他的数据样本点之间的距离经过重新计算后可能都产生了极大的变化, 从而必连约束相关的那部分数据样本点的合并过程也将随之产生极大的变化, 这种变化也将波及到整个合并过程. 当然, 这种变化在特定的情况下也可能反而会有利于聚类效果, 但大多情况下却不是如此. 如图 1 所示是这种合并方式下的两个示例. 假设两个类别之间的距离用它们中心点之间的距离来表示.

对于图 1 (a), 正方形和圆圈分别代表不同的两类, 记为类别 C_1 和 C_2. 其中实心圆圈 A 和 B 代表的数据样本点之间具有必连约束, 从而在算法开始时将 A 和 B 合并为一类. M 代表的位置是 A 和 B 的中点, 从而也是 A 和 B 所属类别的中心. 不难看出, 在这种假设情况下, M 与属于类别 C_2 的数据样本点之间的距离非常近, 甚至远比属于类别 C_1 的数据样本点之间的距离近, 因而 A 和 B 很快会和 C_2 中的数据样本点进行合并操作. 这种不同类别间的

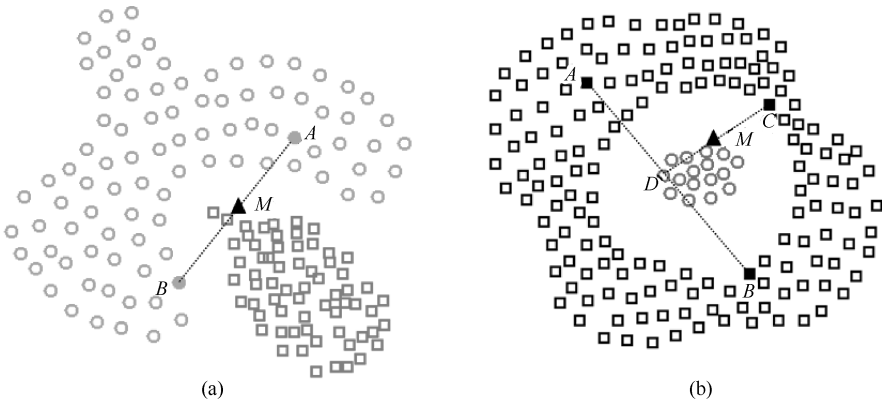


图 1　直接合并必连约束中的数据样本点所存在的问题的两个实例

Fig. 1　Two examples to illustrate the problems of directly merging instances contain must-link constraints at the beginning of the algorithm

数据样本点的过早合并会导致最终的聚类效果十分不理想. 图 1 (b) 所示意的情况与图 1 (a) 类似, 不同的是此时必连约束所涉及的数据样本点的数量为 3. 其中, D 为 A 和 B 的中点, M 为 D 和 C 的中点, 从而 M 代表点 A、B 和 C 组成的类别的中心. 这时, 同样 A、B 和 C 会过早和另一个类别中的数据样本点进行合并.

2 基于约束动态更新的半监督层次聚类算法

为了解决前面提出的半监督层次聚类算法中存在的问题, 本文提出了一种新型的半监督层次聚类算法. 该算法在整个层次聚类过程中不断动态更新约束, 从而避免了在最开始就合并满足必连约束的数据样本点.

2.1 算法原理

在介绍约束动态更新的半监督层次聚类算法之前, 这里首先定义约束的传递性特征. 具体见定义 1 同类传递和定义 2 异类传递. 这里必连约束集合记为 Ω, 不连约束集合记为 Θ.

定义 1 (同类传递). $\forall(x_1, x_2) \in \Omega$, $(x_1, x_3) \in \Omega$, 则 $(x_2, x_3) \in \Omega$.

定义 2 (异类传递). $\forall(x_1, x_2) \in \Omega$, $(x_1, x_3) \in \Theta$, 则 $(x_2, x_3) \in \Theta$.

接下来给出约束集合闭包、同类闭包和异类闭包的定义.

定义 3 (闭包). 约束集合 Ω 和 Θ 是一个闭包是指 Ω 是一个同类闭包同时 Θ 是一个异类闭包.

定义 4 (同类闭包). 必连约束集合是一个同类闭包是指所有能够依据同类传递推断得到的必连约束均被包含在了该必连约束集合中.

定义 5 (异类闭包). 不连约束集合是一个异类闭包是指所有能够依据异类传递推断得到的不连约束均被包含在了该不连约束集合中, 这里需给定对应的必连约束集合.

约束动态更新的半监督层次聚类算法首先将给定的必连约束集合 Ω 和不连约束集合 Θ 扩展为一个约束闭包. 具体方法见算法 2.

算法 2. 将约束集合扩展为闭包的方法

输入. 数据样本集 X, 必连约束集合 Ω 和不连约束集合 Θ.

输出. 同类闭包 Ω 和异类闭包 Θ.

步骤 1. 将必连约束中包含的所有数据样本点划分成 r 个类别, 得到 $M_1, M_2, \cdots, M_r$[38]. 其中 $M_1, M_2, \cdots, M_r$ 满足条件, 对 $\forall a, b, i, j, x_i \in M_a, x_j \in M_b, a \neq b$, 有 $(x_i, x_j) \in \Omega$.

步骤 2. 对 $k=1$ 至 r, 对所有的 $x_i, x_j \in M_k, x_i \neq x_j$, 将 (x_i, x_j) 加入 Ω, 即 $\Omega = \Omega \cup (x_i, x_j)$.

步骤 3. 对 $k = 1$ 至 r, 对所有的 $x_u \in M_k$, 寻找所有与 x_u 满足不连约束的数据样本点, 记为集合 M_Θ 对所有的 $x_i \in M_k$, $x_j \in M_\Theta$, $x_i \neq x_j$, 将 (x_i, x_j) 加入 Ω, 即 $\Omega = \Omega \cup (x_i, x_j)$.

步骤 4. 返回扩展后的必连约束集合 Ω 和不连约束集合 Θ.

将约束集合扩展为闭包是为了保证任意两个满足必连约束的数据样本点最终都会被划分到同一个类别. 为说明这一问题, 现在考虑以下情景: $(A, B) \in \Omega$, $(A, C) \in \Omega$, $(B, D) \in \Omega$, 当约束集合 Ω 和 Θ 不是闭包时, 有可能 $(C, D) \notin \Theta$, 即二者可以合并, 当它们合并后, A、B 和 C 就不可能全部被划分为同一个类别; 而当约束集合 Ω 和 Θ 被扩展为闭包, 一定有 $(C, D) \in \Theta$, 那么这一问题就得以解决.

约束动态更新的半监督层次聚类算法的关键部分是如何动态更新约束. 现在考虑层次聚类过程中的一次合并过程: 假设类别 C_1 和 C_2 被选择为将要合并的两个类别, 合并后的类别记为 $C_\cup$, 即 $C_\cup = C_1 \cup C_2$, 而类别 C_3、C_4、C_5、C_6 是除 C_1 和 C_2 之外的类别中的四个, 并满足以下条件:

1) $(C_1, C_3) \in \Theta$;
2) $(C_1, C_4) \in \Omega$;
3) $(C_2, C_5) \in \Theta$;
4) $(C_2, C_6) \in \Omega$.

这里需要注意的是, 最开始时每个数据样本点分别构成一个类别, 因而数据样本点满足某一约束时, 其对应的两个类别也满足该约束. 在约束更新之后, 其代表的不仅包括数据样本之间的约束关系, 同时还代表类别之间的约束关系. 依据以上条件能够推断得到的关系包括以下 10 种:

1) $[1] \rightarrow (C_\cup, C_3) \in \Theta$;
2) $[2] \rightarrow (C_\cup, C_4) \in \Omega$;
3) $[3] \rightarrow (C_\cup, C_5) \in \Theta$;
4) $[4] \rightarrow (C_\cup, C_6) \in \Omega$;
5) $[1]+[2] \rightarrow (C_3, C_4) \in \Theta$;
6) $[1]+[3] \rightarrow$ 当数据样本为两类时 $(C_3, C_5) \in \Omega$, 否则不能得到任何结论;
7) $[1]+[4] \rightarrow (C_3, C_6) \in \Theta$;
8) $[2]+[3] \rightarrow (C_4, C_5) \in \Theta$;
9) $[2]+[4] \rightarrow (C_4, C_6) \in \Omega$;
10) $[3]+[4] \rightarrow (C_5, C_6) \in \Theta$.

上述各种关系的证明比较简单, 其中 2)、4)、5)、6) 和 9) 由同类传递特征得到, 而 1)、3)、7)、8) 和 10) 则由异类传递特征得到. 在层次聚类算法的执行过程中, 必连和不连约束正是依据上述的 10 种推断进行更新. 但需要注意的是将 Ω 和 Θ 被扩展为闭包时, 5) 和 10) 已经自动被满足, 所以不需更新. 当不能确定类别为 2 时, 6) 不是有

效的推断. 从而 Ω 和 Θ 的具体更新方式见算法 3.

算法 3. 约束更新方法

输入. 给定需要合并的类别 C_1 和 C_2, 合并结果 $C_\cup$, 必连约束集合 Ω 和不连约束集合 Θ.

输出. 合并之后的必连约束集合 Ω 和不连约束集合 Θ.

步骤 1. 对 $\forall C_k$ 满足 $(C_1, C_k) \in \Omega$ 或 $(C_2, C_k) \in \Omega$, 则将 $(C_\cup, C_k)$ 加入 Ω, 即 $\Omega = \Omega \cup (C_\cup, C_k)$;

步骤 2. 对 $\forall C_k$ 满足 $(C_1, C_k) \in \Theta$ 或 $(C_2, C_k) \in \Theta$, 则将 $(C_\cup, C_k)$ 加入 Θ, 即 $\Theta = \Theta \cup (C_\cup, C_k)$;

步骤 3. 分别计算与 C_1 和 C_2 满足必连约束的集合 Ω_1 和 Ω_2, 以及满足不连约束的集合 Θ_1 和 Θ_2, 对于 $\forall C_k \in \Omega_1$, $\forall C_s \in \Theta_2$, 将 (C_k, C_s) 加入 Θ, 即 $\Theta = \Theta \cup (C_k, C_s)$, 对于 $\forall C_k \in \Omega_2$, $\forall C_s \in \Theta_1$, 将 (C_k, C_s) 加入 Θ, 即 $\Theta = \Theta \cup (C_k, C_s)$ 对于 $\forall C_k \in \Omega_1$, $\forall C_s \in \Omega_2$, 将 (C_k, C_s) 加入 Ω, 即 $\Omega = \Omega \cup (C_k, C_s)$;

步骤 4. 从 Ω 和 Θ 中将所有关于 C_1 或 C_2 的约束去除;

步骤 5. 返回更新之后的必连约束集合 Ω 和不连约束集合 Θ.

约束动态更新的半监督层次聚类算法的主体过程和标准的层次聚类算法相似, 只是在执行过程中选择合并的类别时需要考虑不连约束, 并且在两个类别合并完成之后需要对约束进行更新. 算法流程见算法 4.

算法 4. 约束动态更新的层次聚类算法

输入. 数据样本集 X, 必连约束集合 Ω 和不连约束集合 Θ.

输出. 表示聚类结果的树状图.

步骤 1. 初始化类别. 将数据样本点各自作为一类, 即 $C_i = \{x_i\}$, $\forall i$. 并构建树状图 Dendrogram.

步骤 2. 计算任意两个类别间的距离, 即 $d_{i,j} = D(C_i, C_j), \forall i, j$.

步骤 3. 将约束集合 Ω 和 Θ 扩展为闭包.

步骤 4. while 存在可合并的类别

//"存在可合并的类别" 表示至少有两个类别间不存在不连约束

1) 寻找满足不连约束类别对中距离最小的两个类别 C_a 和 C_b.

2) 将类别 C_a 和 C_b 合并为新的类别 C_n, 并计算 C_n 与除 C_a 和 C_b 之外的其他类别之间的距离.

3) 删除类别 C_a 和 C_b, 加入类别 C_n, 并构建树状图 Dendrogram.

4) 更新约束集合 Ω 和 Θ.

步骤 5. 返回树状图 Dendrogram.

2.2 C-Ward 算法

与标准的层次聚类算法一致, Ward 层次聚类算法初始将各个样本数据点看作单独的一个类别, 通过不断地合并距离最近的两个类而产生新的分类结果, 直至满足终止条件. 其中, 任意两个类的距离使用 Ward 连接 (Ward linkage) 来表示. 例如, 对于使用欧氏距离的两个数据样本点 x_i 和 x_j, 其 Ward 连接可以表示为

$$W(x_i, x_j) = \frac{\|x_i - x_j\|^2}{2} \tag{1}$$

假设 A 和 B 为两个类别, 并且其均值分别为 μ_A 和 μ_B, 则 A 和 B 两个类之间的 Ward 连接可以表示为

$$W(A, B) = \frac{|A|\,|B|\,\|\mu_A - \mu_B\|^2}{|A| + |B|} \tag{2}$$

其中, $|S|$ 表示集合 S 的元素个数. 显见, 当 $A = \{x_i\}$, $B = \{x_j\}$ 时, 有 $W(A, B) = W(x_i, x_j)$. Lance-Williams 公式[39] 给出了如下 Ward 连接的递推关系.

假设 A、B 和 C 为三个类别, A 和 B 两个类别合并后得到的类别记为 A', 则 A' 和 C 之间的 Ward 连接可以表示为式 (3).

这一递推关系式可以非常有效地计算两个类别在合并之后与其他类别之间的距离. 而其他类别之间的距离是不变的. 从而半监督 Ward 层次聚类算法 C-Ward 的算法流程可以归结为算法 4. 其中, 步骤 2 中计算任意两个类别间的距离使用式 (1); 步骤 3 计算 C_n 与其他类别之间的距离使用式 (3).

上述算法在图 1 中避免了不同类别间的数据样本点过早合并. 虽然图 1 (a) 中 M 代表的位置是 A 和 B 的中点, 从而也是 A 和 B 所属类别的中心, 但由于 M 与属于类别 C_2 的数据样本点之间的距离非常近, M 归入了 C_2, 避免了图 1 (a) 中不合理的过早合并的情形. 在图 1 (b) 中的情况也类似.

对 C-Ward 复杂度在不同约束条件数量的计算

$$W(A', C) = \frac{(|A| + |C|)W(A, C) + (|B| + |C|)W(B, C) - |C|\,W(A, B)}{|A| + |B| + |C|} \tag{3}$$

量可以如下考虑. 在第 k 步迭代时, 一共为 $N-k$ 类 (N 为样本数量). Ward 连接数量为 $(1/2)(N-k)^2-(N-k)$, 现需要从中选择满足约束条件的最大连接值, 平均计算次数为 $(1/4)(N-k)^2-(1/2)(N-K)$. 在选择出满足条件的连接值后, 即确定了需要合并的两个类, 更新必连约束和不连约束分别最多需要 $(1/2)(N-k)^2-(N-k)$ 次计算. 因而在第 k 步迭代时, 计算次数最多为 $(5/4)(N-k)^2-(5/2)(N-k)$, 这个数随 k 增大而逐渐变小, 所以算法时间复杂度为比 $\mathrm{O}(N^3)$ 略小一点的量级.

3 实验分析

我们对本文提出的约束动态更新的半监督层次聚类算法 C-Ward 进行了相关的实验研究, 对比算法包括 Wagstaff 等的 COP-KMEANS 算法、Ruiz 等的 C-DBSCAN 算法以及 Davidson 等的半监督层次聚类算法基于 Ward 算法的实现 AHCC-Ward 算法.

3.1 数据集和测试标准

实验使用的数据集包括 3 组人工模拟数据集和 5 组现实数据集. 人工模拟数据集来自于测试聚类算法常用的一些数据集, 而现实数据集则来自 UCI Machine learning repository[40]. 表 1 列出了算例名称、来源、数据数量、属性数量、类别数量等信息. 图 2 给出了三个人工模拟数据集的图示.

在一次实验中, 30 % 的数据样本将被随机选择以生成两类约束, 约束的生成也是完全随机的. 具体办法为: 从备选数据样本点中随机选择两个, 如果这两样本点的真实类别相同则加入必连约束集合, 否则加入不连约束集合, 重复这个过程 N 次以产生 N 个约束. N 的取值为 100, 200, 300, $\cdots$, 2 000. 对应 N 每个取值一共进行 30 次独立的实验. 各次的实验结果基于这 30 次实验的统计结果.

本次实验采用的 F-measure[46] 作为测试标准. F-measure 取值在 0 和 1 之间, 值越高表明聚类结果越好. F-measure 通过点对的召回率和准确率来计算. 其计算方法如下:

$$F=\frac{(1+\beta^2)\times P\times R}{\beta^2\times P+R} \tag{4}$$

其中, P 为点对聚类的准确率, R 为点对聚类的召回率, β 为用来加权准确率和召回率参数, 以体现对二

表 1 实验数据集

Table 1 Summary of the dataset for the experimental analysis

Number	Dataset	Instances	Features	Classes
1)	aggregation[41]	788	2	7
2)	compound[42]	399	2	6
3)	path[43]	300	2	3
4)	banknote[40]	1 372	4	2
5)	ionosphere[40]	351	34	2
6)	tic-tac-toe[40]	958	9	2
7)	libras-movement[44]	360	90	15
8)	urban-land-cover[45]	168	147	9

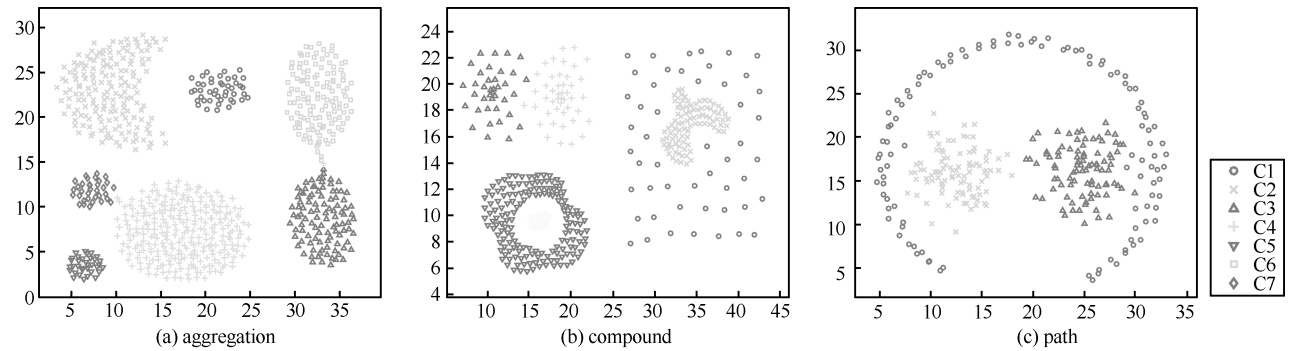


图 2 模拟数据集图示: (a) aggregation, (b) compound 及 (c) path

Fig. 2 Illustrations of the artificial simulated datasets: (a) aggregation, (b) compound, and (c) path

者的重视程度. 在本次研究中, β 取值为 1. P 和 R 的计算公式分别为

$$P = \frac{\rho}{m} \tag{5}$$

$$R = \frac{r}{m} \tag{6}$$

其中, m 为真实类别条件下点对的总数, ρ 为真实类别和聚类类别均相同的点对数量, r 为聚类类别条件下点对的总数, 计算方法分别为

$$m = \sum_{i=1}^{k} \frac{|c_i|\left(|c_i|-1\right)}{2} \tag{7}$$

$$\rho = \sum_{i=1}^{N}\sum_{j=i+1}^{N} \theta\left(\langle y_i, y_j\rangle, \langle \tilde{y}_i, \tilde{y}_j\rangle\right) \tag{8}$$

$$r = \sum_{i=1}^{\tilde{k}} \frac{|\tilde{c}_i|\left(|\tilde{c}_i|-1\right)}{2} \tag{9}$$

其中, k 为真实的类别总数, $|c_i|$ 为真实类别 c_i 的元素个数; y_i 和 y_j 分别为数据样本点 x_i 和 x_j 的真实类别, $\tilde{y}_i$ 和 $\tilde{y}_j$ 分别为数据样本点 x_i 和 x_j 的聚类类别. 如果 $y_i = y_j$ 且 $\tilde{y}_i = \tilde{y}_j$, 则 $\theta\left(\langle y_i, y_j\rangle, \langle \tilde{y}_i, \tilde{y}_j\rangle\right) = 1$, 否则等于 0, N 为样本数量; $\tilde{k}$ 为聚类类别的总数, $|\tilde{c}_i|$ 为聚类类别 $\tilde{c}_i$ 的元素个数.

3.2　实验结果

表 2～4 给出了 COP-KMEANS 算法、C-DBSCAN 算法 AHCC-Ward 算法和 C-Ward 算法 30 次独立实验的统计结果. 表 2、表 3 和表 4 分别是约束数量为 100、200 和 2 000 时的实验结果. 其中, $a \pm b$ 表示均值为 a, 标准差为 b; CW-COp-value 为 C-Ward 算法和 COP-KMEANS 算法实验结果配对 T 检验的 p 值; CW-CDp-value 为 C-Ward 算法和 C-DBSCAN 算法实验结果配对 T 检验的 p 值; CW-CO p-value 为 C-Ward 算法和 AHCC-Ward 算法实验结果配对 T 检验的 p 值.

表 2　约束数量为 100 时的实验结果统计

Table 2　Experimental results when the number of the constraints is 100

Dataset	Algorithm						
	COP-KMEANS	C-DBSCAN	AHCC-Ward	C-Ward	CW-CO p-value	CW-CD p-value	CW-AH p-value
aggregation	0.752 ± 0.039	0.799 ± 0.085	0.944 ± 0.042	0.954 ± 0.031	0.000	0.000	0.133
compound	0.590 ± 0.064	0.633 ± 0.099	0.679 ± 0.087	0.890 ± 0.087	0.000	0.000	0.000
path	0.548 ± 0.033	0.681 ± 0.085	0.643 ± 0.103	0.912 ± 0.116	0.000	0.000	0.000
banknote	0.534 ± 0.019	0.626 ± 0.122	0.688 ± 0.124	0.872 ± 0.161	0.000	0.000	0.000
ionosphere	0.565 ± 0.027	0.510 ± 0.071	0.532 ± 0.090	0.262 ± 0.224	0.000	0.000	0.000
tic-tac-toe	0.533 ± 0.011	0.350 ± 0.030	0.362 ± 0.039	0.215 ± 0.194	0.000	0.000	0.000
libras-movement	0.334 ± 0.021	0.152 ± 0.012	0.370 ± 0.017	0.363 ± 0.026	0.141	0.000	0.000
urban-land-cover	0.507 ± 0.026	0.248 ± 0.016	0.485 ± 0.042	0.405 ± 0.056	0.000	0.000	0.000

表 3　约束数量为 200 时的实验结果统计

Table 3　Experimental results when the number of the constraints is 200

Dataset	Algorithm						
	COP-KMEANS	C-DBSCAN	AHCC-Ward	C-Ward	CW-CO p-value	CW-CD p-value	CW-AH p-value
aggregation	0.743 ± 0.051	0.870 ± 0.070	0.969 ± 0.049	0.985 ± 0.015	0.000	0.000	0.060
compound	0.627 ± 0.096	0.696 ± 0.093	0.667 ± 0.094	0.939 ± 0.042	0.000	0.000	0.000
path	0.531 ± 0.040	0.732 ± 0.092	0.677 ± 0.096	0.961 ± 0.086	0.000	0.000	0.000
banknote	0.536 ± 0.023	0.701 ± 0.133	0.713 ± 0.099	0.972 ± 0.086	0.000	0.000	0.000
ionosphere	0.545 ± 0.034	0.500 ± 0.074	0.614 ± 0.093	0.394 ± 0.355	0.026	0.133	0.003
tic-tac-toe	0.537 ± 0.018	0.304 ± 0.036	0.344 ± 0.051	0.153 ± 0.183	0.000	0.000	0.000
libras-movement	0.336 ± 0.023	0.163 ± 0.015	0.364 ± 0.020	0.370 ± 0.027	0.323	0.000	0.000
urban-land-cover	0.502 ± 0.027	0.254 ± 0.020	0.450 ± 0.053	0.354 ± 0.042	0.000	0.000	0.000

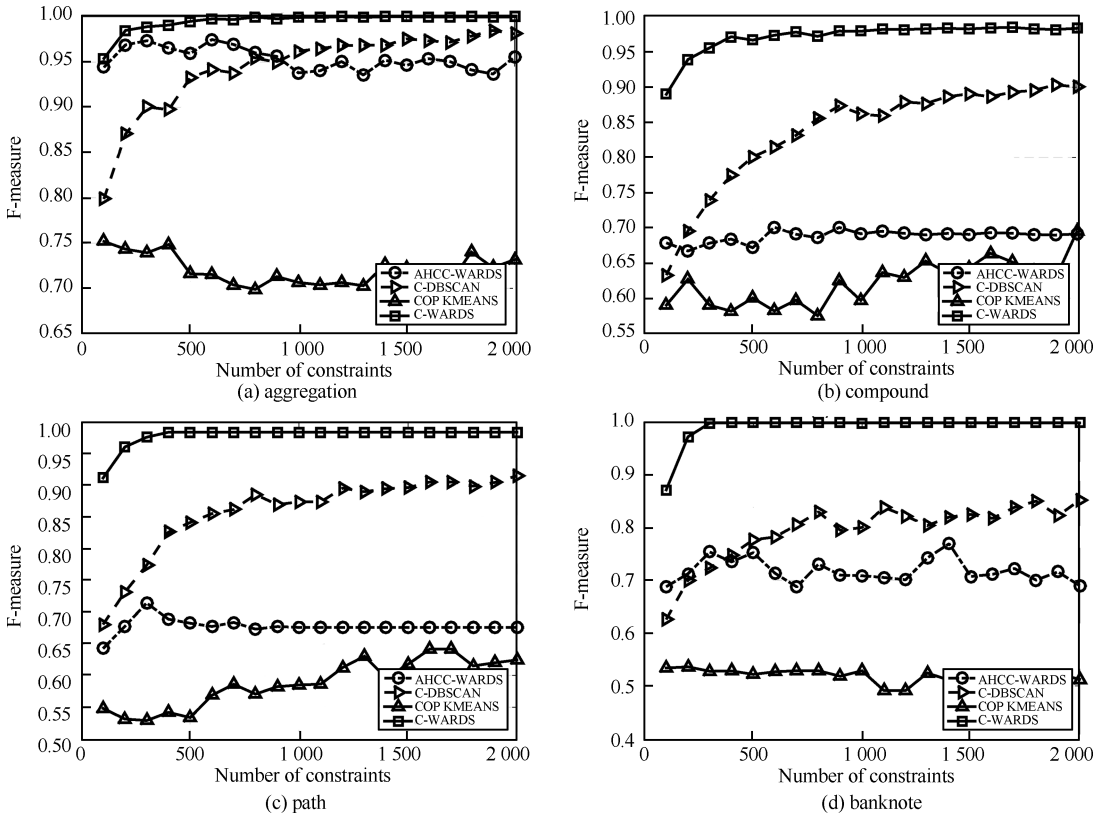
表 4 约束数量为 2 000 时的实验结果统计

Table 4 Experimental results when the number of the constraints is 2 000

Dataset	Algorithm						
	COP-KMEANS	C-DBSCAN	AHCC-Ward	C-Ward	CW-CO p-value	CW-CD p-value	CW-AH p-value
aggregation	0.732 ± 0.067	0.981 ± 0.012	0.956 ± 0.061	1.000 ± 0.000	0.000	0.000	0.000
compound	0.695 ± 0.098	0.900 ± 0.035	0.692 ± 0.033	0.984 ± 0.014	0.000	0.000	0.000
path	0.624 ± 0.061	0.915 ± 0.052	0.676 ± 0.039	0.984 ± 0.012	0.000	0.000	0.000
banknote	0.513 ± 0.049	0.852 ± 0.115	0.690 ± 0.117	0.999 ± 0.004	0.000	0.000	0.000
ionosphere	0.608 ± 0.106	0.615 ± 0.093	0.695 ± 0.019	0.856 ± 0.031	0.000	0.000	0.000
tic-tac-toe	0.513 ± 0.060	0.181 ± 0.016	0.644 ± 0.037	0.920 ± 0.011	0.000	0.000	0.000
libras-movement	0.376 ± 0.031	0.297 ± 0.037	0.432 ± 0.038	0.637 ± 0.039	0.000	0.000	0.000
urban-land-cover	0.626 ± 0.026	0.406 ± 0.034	0.567 ± 0.039	0.645 ± 0.032	0.000	0.041	0.000

从表中可以看出, 本文提出的 C-Ward 算法在大多数情况下明显优于其他三种算法. 对 aggregation、compound、path、banknote、libras-movement 的实验在约束数量为 100、200 和 2 000 时, C-Ward 算法所得到的 F-measure 值一直高于其他算法. 当约束数量为 100 时, COP-KMEANS 在 ionosphere, tic-tac-toe 和 urban-land-cover 实验上的表现要优于 C-Ward 算法. 当约束数量为 200 时, AHCC-Ward 算法在 ionosphere 实验上的表现要优于 C-Ward 算法, 而 COP-KMEANS 算法在 tic-tac-toe 和 urban-land-cover 的实验上的表现要优于 C-Ward 算法. 配对 T 检验结果也充分说明了这一问题. 当取置信度为 5 % 时, 仅仅当约束数量为 100 和 200 时, C-Ward 在 aggregation 实验上的表现不显著优于 AHCC-Ward 算法, p 值分别为 0.133 和 0.06, 而其他情况下这种优势都是显著的.

图 3 给出了 30 次实验 F-measure 均值随约束数量增加而变化的图示. 对于实验 aggregation、compound、path 和 banknote, C-Ward 算法的 F-measure 值始终高于其他三种方法. 并且, 在约束数量达到 500 后, 实验结果基本已经趋于稳定. 对于实验 ionosphere, 当约束的数量低于 400



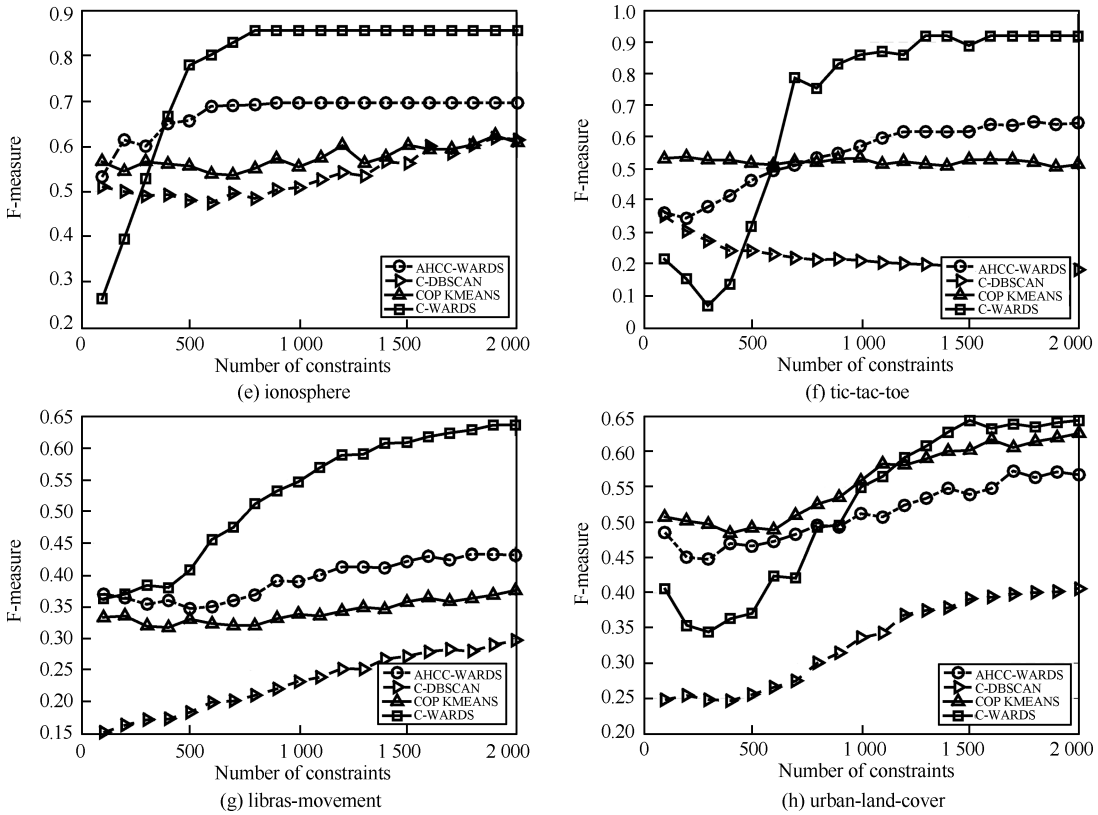


图 3　30 次实验 F-measure 均值随约束数量增加而变化的图示: (a) aggregation, (b) compound, (c) path, (d) banknote, (e) ionosphere, (f) tic-tac-toe, (g) libras-movement 和 (h) urban-land-cover

Fig. 3　Changes on the average F-measure values in 30 experiments with the increase of constraints number: (a) aggregation, (b) compound, (c) path, (d) banknote, (e) ionosphere, (f) tic-tac-toe, (g) libras-movement, and (h) urban-land-cover

时, 其他三种方法一般都优于 C-Ward, 特别是 AHCC-Ward 算法; 而当约束的数量高于 400 时, C-Ward 算法则变得比较优秀. 实验 tic-tac-toe 的情况比较类似, 当约束的数量少于 600 时, C-Ward 算法并不比其他三种算法优秀, 但当该数量高于 600 时则优势十分明显.

总的来说, 本文提出的 C-Ward 算法总体上大大优于进行对比的其他三种方法. 不管是人工模拟数据集还是现实数据集都支持这一结论. 这种优势一方面体现在总体的聚类效果上, 因为 F-measure 的均值一般高于其他的方法; 另一方面还体现在聚类结果的稳定性上, 因为 30 次实验结果的标准差也一般较小. 另外, 本文提出的 C-Ward 算法在约束数量较少时, 在某些算例的表现不如另外的三种算法. 但随着约束数量的增加, 这种优势变得非常明显.

4　结束语

针对 Davidson 等的半监督层次聚类算法存在的主要问题, 本文提出了一种新的基于约束动态更新的半监督层次聚类算法, 具体地给出了该算法基于 Ward 算法的实现, 提出了 C-Ward 算法.

本文在人工模拟数据集和现实数据集上的实验均表明, C-Ward 算法比 Davidson 等的半监督层次聚类算法基于 Ward 算法的实现 AHCC-Ward 算法表现地更为优秀和稳定. 同时, 同类算法 COP-KMEANS 算法和 C-DBSCAN 算法也加入了实验对比, 结果表明 C-Ward 仍然相对表现较好. 一般来看, 当约束数量在 500 左右时, C-Ward 算法的聚类结果就达到了比较理想的效果, 表现出了较高的准确性和较强的稳定性.

层次聚类是一种很实用的简单算法, 而将其半监督化是可以极大地提高其性能. 本文的下一步工作是, 进一步从不同的视角进行改进, 并选用一些更大规模的数据进行实验研究.

References

1　Chapelle O, Schölkopf B, Zien A. *Semi-Supervised Learning*. Massachusetts: MIT Press, 2006. 2−5

2　Zhu X J. Semi-supervised learning. In: Proceedings of the 2010 Encyclopedia of Machine Learning. US: Springer, 2010.

892−897

3 Balcan M F, Blum A. A PAC-style model for learning from labeled and unlabeled data. In: Proceedings of the 2005 Learning Theory. Berlin, Heidelberg: Springer, 2005. 111−126

4 Kääriäinen M. Generalization error bounds using unlabeled data. In: Proceedings of the 2005 Learning Theory. Berlin, Heidelberg: Springer, 2005. 127−142

5 Singh A, Nowak R D, Zhu X J. Unlabeled data: now it helps, now it doesn′t. In: Proceedings of the 2008 Advances in Neural Information Processing Systems 21 (NIPS). Vancouver, Canada, 2008. 1513−1520

6 Wagstaff K, Cardie C, Rogers S, Schrodl S. Constrained k-means clustering with background knowledge. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. 577−584

7 Ruiz C, Spiliopoulou M, Menasalvas E. C-DBSCAN: density-based clustering with constraints. In: Proceedings of the 11th International Conference. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Toronto, Canada: Springer, 2007. 216−223

8 Davidson I, Ravi S S. Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: Proceedings of the 2005 Knowledge Discovery in Databases: PKDD 2005. Berlin, Heidelberg: Springer, 2005. 59−70

9 Deng Chao, Guo Mao-Zu. Tri-training and data editing based semi-supervised clustering algorithm. *Journal of Software*, 2008, **19**(3): 663−673 (in Chinese)

10 Wang Hong-Jun, Li Zhi-Shu, Qi Jian-Huai, Cheng Yang, Zhou Peng, Zhou Wei. Semi-supervised cluster ensemble model based on Bayesian network. *Journal of Software*, 2010, **21**(11): 2814−2825 (in Chinese)

11 Basu S, Banerjee A, Mooney E R, Banerjee A, Mooney R J. Active semi-supervision for pairwise constrained clustering. In: Proceedings of the 2004 SIAM International Conference on Data Mining. Lake Buena Vista, FL: SIAM, 2004. 333−344

12 de Amorim RC. Constrained clustering with Minkowski weighted k-means. In: Proceedings of the 13th IEEE International Symposium Computational Intelligence & Informatics. Budapest: IEEE, 2012. 13−17

13 Yin Xue-Song, Hu En-Liang, Chen Song-Can. Discriminative semi-supervised clustering analysis with pairwise constraints. *Journal of Software*, 2008, **19**(11): 2791−2802 (in Chinese)

14 Wang Y Y, Chen S C, Zhou Z-H. New semi-supervised classification method based on modified cluster assumption. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(5): 689−702

15 Lin B B, Zhang C Y, He X F. Semi-supervised regression via parallel field regularization. In: Proceedings of the 2011 Advances in Neural Information Processing Systems 24 (NIPS). Granada, Spain, 2011. 433−441

16 Xiang S M, Nie F P, Zhang C S. Semi-supervised classification via local spline regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(11): 2039−2053

17 Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000, **39**(2−3): 103−134

18 Nigam K P. Using Unlabeled Data to Improve Text Classification, Technical Report, CMU-CS-01-126, Carnegie Mellon University, Pittsburgh, 2001.

19 Cozman F G, Cohen I, Cirelo M C. Semi-supervised learning of mixture models. In: Proceedings of the 20th International Conference on Machine Learning. Washington D. C., 2003.

20 Bennett K P, Demiriz A. Semi-supervised support vector machines. In: Proceedings of the 1998 Advances in Neural Information Processing Systems. Cambridge: MIT Press, 1998. 368−374

21 Fung G, Mangasarian O. Semi-supervised Support Vector Machines for Unlabeled Data Classification, Technical Report, 99-05, Data Mining Institute, University of Wisconsin Madison, 1999

22 Chapelle O, Zien A. Semi-supervised learning by low density separation. In: Proceedings of the 10th Int Workshop Artificial Intelligence & Statistics, 2005. 57−64

23 Chapelle O, Sindhwani V, Keerthi S S. Branch and bound for semi-supervised support vector machines. In: Advances Neural Information Processing Systems (NIPS). Vancouver, Canada, 2006. 217−224

24 De Bie T, Cristianini N. Convex methods for transduction. In: Proceedings of the 2003 Advances Neural Information Processing Systems 16. Vancouver, Canada, 2003. 73−80

25 Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. 19−26

26 Joachims T. Transductive learning via spectral graph partitioning. In: Proceedings of the 20th International Conference on Machine Learning. Washington D. C., 2003. 290−297

27 Zhu X J, Ghahramani Z. Towards Semi-supervised Classification with Markov Random Fields, Technical Report, CMU-CALD-02-106, Carnegie Mellon University, 2002.

28 Xiao Yu, Yu Jian. Semi-Supervised clustering based on affinity propagation algorithm. *Journal of Software*, 2008, **19**(11): 2803−2813 (in Chinese)

29 Culp M, Michailidis G. An iterative algorithm for extending learners to a semi-supervised setting. In: Proceedings of the 2007 Joint Statistical Meetings. Salt Lake, Utah, 2007.

30 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison: ACM, 1998. 92−100

31 Zhou Y, Goldman S. Democratic co-learning. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, FL: IEEE, 2004. 594−602

32 Zhou Z H, Li M. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge & Data Engineering*, 2005, **17**(11): 1529−1541

33 Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 2007, **37**(6): 1088−1098

34 Murtagh F, Legendre P. Ward′s hierarchical clustering method: clustering criterion and agglomerative algorithm. arXiv preprint arXiv, 2011: 1111.6285

35 Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1996. 103−114

36 Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1998. 73−84

37 Karypis G, Han E H, Kumar V. Chameleon: hierarchical clustering using dynamic modeling. *Computer*, 1999, **32**(8): 68−75

38 Davidson I, Ravi S S. Clustering with constraints: feasibility issues and the *k*-means algorithm. In: Proceedings of the 2005 SIAM International Conference on Data Mining. Lake Buena Vista, FL: SIAM, 2005. 138−149

39 Cormack R M. A review of classification. *J Royal Statistical Society. Series A (General)*, 1971, **134**(3): 321−367

40 Bache K, Lichman M. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml, November 3, 2013

41 Gionis A, Mannila H, Tsaparas P. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 2007, **1**(1): Article No. 4

42 Zahn C T. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 1971, **C-20**(1): 68−86

43 Chang H, Yeung D Y. Robust path-based spectral clustering. *Pattern Recognition*, 2008, **41**(1): 191−203

44 Dias D B, Madeo R C B, Rocha T, Biscaro H H, Peres S M. Hand movement recognition for Brazilian Sign Language: a study using distance-based neural networks. In: Proceedings of the 2009 International Joint Conference on Neural Networks. Atlanta, GA: IEEE, 2009. 697−704

45 Johnson B, Xie Z X. Classifying a high resolution image of an urban area using super-object information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2013, **83**: 40−49

46 Hripcsak G, Rothschild A S. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 2005, **12**(3): 296−298

周晨曦 中国人民大学硕士研究生. 主要研究方向为数据挖掘.
E-mail: chnx.zhou@gmail.com
(**ZHOU Chen-Xi** Master student at Renmin University of China. His main research interest is data mining.)

梁 循 中国人民大学信息学院教授. 主要研究方向为互联网信息分析, 数据挖掘, 商务智能, 社会计算. 本文通信作者.
E-mail: xliang@ruc.edu.cn
(**LIANG Xun** Professor at the School of Information, Renmin University of China. His research interest covers internet information analysis, data mining, business intelligence, and social computing. Corresponding author of this paper.)

齐金山 中国人民大学博士研究生. 主要研究方向为数据挖掘, 社会计算.
E-mail: qijinshan@sina.com
(**QI Jin-Shan** Ph. D. candidate at Renmin University of China. His research interest covers data mining and social computing.)