doi:10.3969/j.issn.1672-1292.2018.02.008

# 一种主动半监督 *K*-means 聚类算法的改进策略

吕　峰,柴变芳,李文斌,王　垚

(河北地质大学信息工程学院,河北 石家庄 050031)

[**摘要**]　经典的 APCKmeans(active pairwise constrained *K*-means)算法通过主动学习的方式构造 must-link 约束集和 cannot-link 约束集作为监督信息进行半监督聚类,提高了结果的准确性. 但该算法在样本指派的过程中可能出现指派不是当前最优的问题. 提出一种优先指派标签样本的方法,应用于 APCKmeans 算法,使用改进后的 APCKmeans_I 算法实现了使用较少的监督信息取得更好的聚类结果. 将改进策略应用于 PCKmeans(pairwise constrained *K*-means)算法,提出改进后的 PCKmeans_I 算法. 通过在 UCI 基准数据集的实验表明,改进后算法的性能得到明显提升.

[**关键词**]　主动半监督聚类,成对约束聚类,改进算法

[**中图分类号**]TP181　[**文献标志码**]A　[**文章编号**]1672-1292(2018)02-0056-07

# An Improved Strategy of Active Semi-supervision *K*-means Clustering Algorithm

Lü Feng,Chai Bianfang,Li Wenbin,Wang Yao

(School of Information Engineering,Hebei GEO University,Shijiazhuang 050031,China)

**Abstract**:The classic APCKmeans(active pairwise constrained *K*-means) algorithm constructs the must-link constraint set and the cannot-link constraint set as the supervised information by Semi-Supervised Clustering through the active learning method to improve the accuracy of the results. However,the algorithm may not be assigned to the current optimal problem during the sample assignment process. This paper proposes a method of assigning label samples to APCKmeans algorithm,and proposes an improved APCKmeans_I algorithm to achieve better clustering results with less supervisory information. The improved strategy is applied to PCKmeans(pairwise constrained *K*-means) algorithm,and PCKmeans_I algorithm is proposed. Experiments on the UCI reference data set show that the performance of the improved algorithm is obviously improved.

**Key words**:active semi-supervised clustering,pairwise constrained clustering,improved algorithm

随着数据采集技术和存储技术的发展,获取海量数据已变得相当容易. 这些数据可用于实现机器学习和数据挖掘任务,其中少量样本具有标记,大多为无标记. 半监督聚类通过将部分先验信息加入到无监督聚类过程用以改善聚类性能,挖掘数据潜在规律. 与只利用无标记样本的聚类算法相比,半监督聚类可有效提高聚类的效果[1-5].

近年来,半监督聚类已成为研究的热点. Basu 等提出了 Constrained *K*-means 和 Seeded *K*-means 算法,在经典 *K*-means 算法上通过少量的有标记样本形成种子集合,并从中选取 $k$ 个类别的样本来初始化中心点,从而形成半监督的 *K*-means 算法[1]. 方玲等提出了结合特征的相对偏好和少量标记样本的半监督聚类,并将这种建模方法应用到其他类似场景[6]. 张俊溪等针对复杂分布数据的特征,提出了一种结合元胞自动机距离变换算法的半监督聚类[7]. 高莹等改进了 *K*-means 聚类算法的初始聚类中心选择方法以及相

**收稿日期**:2017-12-10.

**基金项目**:国家自然科学基金(61503260)、河北省研究生创新资助项目(CXZZSS2017131)、河北地质大学教改项目(2017J04).

**通讯联系人**:李文斌,博士,教授,研究方向:机器学习、复杂网络等. E-mail:25304189@qq.com

似性度量方法,提出了一种半监督 K-means 多关系数据聚类算法[8]. 半监督聚类中的监督信息往往需要人工手动标注一些样本,这无疑会产生很大的时间花销,且监督信息若标记不当,可能会给实验结果造成较大的负面影响. 使用主动选择的先验信息可以进一步提高半监督聚类算法的性能.

当前,国内外对主动半监督聚类算法已做了大量开创性研究工作. Basu 等结合成对约束关系和距离,使用一种基于 K-means 的半监督聚类算法,将约束对获得方式改为主动选择,构造了一种主动成对约束半监督聚类(APCKMeans)算法[9]. Xiong S 等基于不确定性原理,运用一种用于计算各个样本相关不确定性的方法,提出了一种改进的主动半监督聚类算法,提高了算法的准确率[10]. Greene D 等提出了一种优先关注聚类结果模糊的样本来选择信息约束的新方法,实现了在更少的约束信息情况下提高聚类准确率[11]. Huang R 等提出了一种主动学习方法进行选择信息文档对,并构造出 3 种不同的模型用以提高文本聚类的准确率[12]. Mallapragada P K 等提出一种基于 Min-Max 标准的主动查询选择体系,提高了半监督聚类算法的性能[13]. Xu Q 等通过校验从相似矩阵导出的光谱特征向量,提出了一种主动学习的谱聚类方法,实验结果优于随机选择约束信息的谱聚类算法[14]. 在主动半监督聚类算法的研究中,Basu 等提出的 APCKmeans 算法最为经典,被相关学者大量引用. APCKmeans 算法提出一种最远距离优先的策略,用于主动学习成对约束信息,提高了半监督聚类算法的性能.

在 Basu 等提出的成对约束半监督聚类算法进行样本点指派时,可能会出现指派到的簇并非当前最优的情况. 对此,本文提出一种优化样本指派的改进策略,应用于 Basu 等人提出的 APCKmeans 算法,构建了改进后的 APCKmeans_I 算法. 此后,将改进策略应用于 PCKmeans 算法,提出改进后的 PCKmeans_I 算法,使算法的聚类准确率和收敛速度进一步提高.

# 1　APCKmeans_I 算法

## 1.1　APCKmeans 算法

APCKmeans 算法使用约束对标记形式作为先验信息,主动选择约束对进行标记,基于约束对生成标记节点集合. APCKmeans 相似度度量采用欧氏距离,两点的欧氏距离越大则表示这对样本点的相似度越低. 通过主动选择与已标记节点集合最不相似的节点,标记其与已标记节点的约束关系,进而确定其属于某个标记节点集合. 生成标记节点集合的过程包括两个阶段:初始架构构造阶段和类集合扩充阶段.

定义 $\{X_h\}_{h=1}^{k}$ 表示将数据集 $X$ 分成 $k$ 个互不相容的簇,每个簇的中心用 $\mu_h$ 表示;$K$ 表示要聚类的个数;标记节点集合 $\{N_p\}_{p=1}^{\lambda}$;must-link 约束集用 $M$ 表示,若 $(x_i,x_j)\in$ must-link,则规定 $x_i$ 和 $x_j$ 必须属于同一类;cannot-link 约束集用 $C$ 表示,若 $(x_i,x_j)\in$ cannot-link,则规定 $x_i$ 和 $x_j$ 必须不在同一类;$l_i$ 表示样本点 $x_i$ 被指派到的簇,其中 $l_i=\{h\}_{h=1}^{k}$;样本点违反约束集所要付出的代价用 $w$($w$ 为一个足够大的数)表示. 此外,定义一个指示函数 $\prod$,其中 $\prod[\text{true}]=1$,$\prod[\text{false}]=0$. 需要说明的是,在 must-link 和 cannot-link 约束集的所有约束对 $(x_i,x_j)$ 均满足无序性.

初始架构构造阶段基于最远距离优先的策略,选择与已标记节点集合相似度最低的节点,标记其与已标记节点的约束关系,进而确定其属于某个标记节点集合. 针对某个数据集 $X$,首先随机选取一个样本点作为初始类节点,将其作为 $N_1$ 的初始节点. 然后遍历数据集 $X$,选择与已有标记节点集合相似度最小的样本点,节点与已有 $N_k$ 的相似度值计算由式(1)计算可得. 根据与已有 $N_p$ 的关系,若与已有 $N_k$ 存在 must-link 约束关系则添加到该集合,若与所有 $\{N_k\}$ 均为 cannot-link 关系,则新构造一个集合,将该节点加入.

$$d(x,N_p)=\min_{y\in N_p}d(x,y), \tag{1}$$

式中,$d(x,y)$ 为样本点 $x$ 和样本点 $y$ 的欧氏距离.

初始架构构造阶段的结束条件是集合元素个数达到 $K$,或标记成对约束超过给定数目. 若初始骨架构造阶段在约束超限结束时,$\{N_k\}$ 个数未达到 $K$,则从未选择样本中随机选择样本加入 $\{N_k\}$,直至 $\{N_k\}$ 个数达到 $K$.

类集合扩充阶段是在初始架构已构造好且使用约束个数未超限情形下进行的. 根据式(1)继续在未标记节点中找到和 $\{N_k\}$ 集合相似度最小的样本,通过主动询问的方式得到与集合中元素的约束关系,并添加到 $\{N_k\}$ 集合中.

通过初始架构构造阶段和类集合扩充阶段可主动选择约束对进行标记,自动生成标记节点集合先验,

作为 APCKmeans 算法的先验. 初始架构构造阶段和类集合扩充阶段算法步骤如下:

**算法 1**　初始架构构造阶段和类集合扩充阶段

输入:数据集 $X=\{x_i\}_{i=1}^{n}$;聚类的个数 $k$;给定的可查询次数 $Q$.

输出:集合 $\{N_p\}_{p=1}^{k}$.

(1)$\lambda\leftarrow 1$,$\{N_p\}_{p=1}^{\lambda}\leftarrow$Null;

(2)随机选取一个样本 $x$,将其放入 $N_1$;

(3)找到距离集合 $\{N_p\}_{p=1}^{\lambda}$ 最远的样本 $x$;

(4)若 $x$ 和 $\{N_p\}_{p=1}^{\lambda}$ 均存在 cannot-link 约束,则 $\lambda\leftarrow\lambda+1$,$N_\lambda\leftarrow x$;

(5)否则,将 $x$ 添加到与之有 must-link 约束的 $N_p$;

(6)在查询次数允许且 $\lambda<k$ 时,循环执行(3)~(5),得到初始架构构造集合 $\{N_p\}_{p=1}^{\lambda}$;

(7)随机选取一个不存在于集合 $\{N_p\}_{p=1}^{\lambda}$ 的样本 $x$;

(8)$h\leftarrow 1$;

(9)若 $x$ 和 $N_h$ 有 must-link 约束关系,则将 $x$ 加入到 $N_h$;$h=h+1$;

(10)当 $h\leqslant k$ 时,循环执行(9);

(11)当查询次数允许时,循环执行(7)~(10),得到类集合扩充集合 $\{N_p\}_{p=1}^{k}$.

通过算法 1 得到一组标记节点集合作为先验信息,进而可使用该先验信息指导样本指派. 对于每个样本点 $x_i$,若 $l_i$ 使得目标函数(2)$J_{\text{pckm}}$ 的取值最小,则将 $x_i$ 指派到簇 $X_{l_i}$ 中:

$$J_{\text{pckm}}=\frac{1}{2}\sum_{x_i\in X}\|x_i-\mu_{l_i}\|^2+\sum_{(x_i,x_j)\in M}w[l_i\neq l_j]+\sum_{(x_i,x_j)\in C}w[l_i=l_j]. \tag{2}$$

### 1.2　APCKmeans 算法的缺陷及其改进算法 APCKmeans_I

通过初始架构构造阶段和类集合扩充阶段所得先验样本集合可用于指导 APCKmeans 算法进行样本指派. 在 APCKmeans 算法进行样本指派时,首先遍历样本集 $X$ 中的每个样本点 $x_i$,然后根据目标函数(2)将 $x_i$ 指派到相应的簇. 这样进行样本指派时可能会出现如图 1 所示的情况(为叙述方便,样本集在二维空间中表示,真实数据可能为多维). 不妨设 $N_p=\{x_1,x_2,x_3,x_4,\cdots,x_n\}$,假设 $x_k$ 为 $N_p$ 集合中第一个被指派的点,按照目标函数(2),$x_k$ 将会被指派到距离 $\{\mu_h\}_{h=1}^{k}$ 中最近的中心 $\mu_i$(图 1 中左侧的×所表示)所在的簇. 若 $N_p$ 中大部分的点均距离 $\mu_j$(图 1 中右侧的×所表示)较近,由于违反约束集的代价 $w$ 足够大,$N_p$ 中的所有样本点还是会被指派到 $\mu_i$ 所在的簇. 显然在图 1 所示的这种情况下,将 $N_p$ 中所有的点指派到 $\mu_j$ 是一个更好的选择.

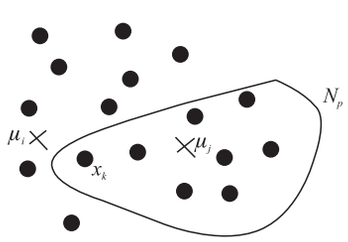


**图 1　示意图**

**Fig. 1　Schematic diagram**

针对上述情况,本文提出一种改进策略,其核心思想为:优先指派 $\{N_p\}_{p=1}^{k}$ 中的点,使目标函数(3)最小化,将节点标记集合 $\{N_p\}_{p=1}^{k}$ 中所有的点指派到相应的簇中,然后再指派其他无标记的样本点,从而更加充分地利用先验信息指导中心点的选择,提高算法的准确性,加快收敛速度:

$$J_{\text{ch}}=\frac{1}{2}\sum_{p=1}^{k}\sum_{x_i\in N_p}\|x_i-\mu_{h(h=1,\cdots,k)}\|^2. \tag{3}$$

改进后的 APCKmeans_I 算法步骤如下:

**算法 2**　APCKmeans_I 算法

输入:数据集 $X$;聚类的个数 $k$;给定的可查询次数 $Q$;违反约束集的代价值 $w$.

输出:将 $X$ 分成互不相容的 $k$ 个子集.

(1)通过算法 1 求得节点标记集合 $\{N_p\}_{p=1}^{k}$;

(2)根据节点标记集合初始化簇中心 $\{\mu_h\}_{h=1}^{k}$;

(3)根据式(3)指派节点标记集合中的样本点;

(4)根据式(2)指派其他样本点;

(5)更新簇中心;

(6)当两次簇中心之差未达到阈值时,循环执行(3)~(5),得到聚类结果$\{X_h\}_{h=1}^{k}$.

## 2　PCKmeans_I 算法

PCKmeans 算法是以两个约束对集合(must-link 集合和 cannot-link 集合)作为先验信息. 先验信息不是通过主动选择的方式得到,而是随机得到两个样本点,然后按照这两个样本点的约束关系添加到相应的约束集合. APCKmeans_I 算法即是在 PCKmeans 算法的基础上加入主动选择并改进样本指派方式得到. 由于 PCKmeans 是通过随机给定约束集合(先验信息),因而不用通过遍历数据集构建节点标记集合. 因此,PCKmeans 算法的时间复杂度要低于基于主动选择的 APCKmeans_I 算法,但准确率略低于 APCKmeans_I 算法.

在 PCKmeans 算法的样本指派阶段同样可能会遇到上文所提情况. 将上文的改进策略应用于 PCKmeans,提出 PCKmeans_I 算法,步骤如下:

**算法 3**　PCKmeans_I 算法

输入:数据集 $X=\{x_i\}_{i=1}^{n}$;must-link 约束集 $M=\{(x_i,x_j)\}$;cannot-link 约束集 $C=\{(x_i,x_j)\}$;聚类的个数 $k$;违反约束集的代价值 $w$.

输出:将 $X$ 分成互不相容的 $k$ 个子集$\{X_h\}_{h=1}^{k}$.

(1)用 $M$ 构建$\{N_p\}_{p=1}^{\lambda}$,然后扩充 $C$ 和 $M$;

(2)将$\{N_p\}_{p=1}^{\lambda}$按照 $N_p$ 中的元素个数降序排序;

(3)若 $\lambda\geqslant k$,则用$\{N_p\}_{p=1}^{k}$初始化 $k$ 个聚类中心 $\mu_1^{(0)}\sim\mu_k^{(0)}$;

(4)否则,用$\{N_p\}_{p=1}^{\lambda}$初始化 $\lambda$ 个聚类中心 $\mu_1^{(0)}\sim\mu_\lambda^{(0)}$,并从 $X-\{N_p\}_{p=1}^{\lambda}$中随机抽取 $k-\lambda$ 个不同的的样本点初始化 $\mu_{\lambda+1}^{(0)}\sim\mu_k^{(0)}$;

(5)根据式(3)指派节点标记集合中的样本点;

(6)根据式(2)指派其他样本点;

(7)更新簇中心;

(8)当两次簇中心之差未达到阈值时,循环执行(5)~(7),得到互不相容的 $k$ 个子集$\{X_h\}_{h=1}^{k}$.

其中,PCKmeans_I 算法步骤 1 中的构造和扩充过程如下:

首先求得约束集 $M$ 的传递闭包,然后将该图的 $\lambda$ 个连通分支构造$\{N_p\}_{p=1}^{\lambda}$集合,得到集合$\{N_p\}_{p=1}^{\lambda}$. 取每个 $N_p$ 的所有点的平均值作为一个初始中心点,得到 $\lambda$ 个初始中心点. 再用$\{N_p\}_{p=1}^{\lambda}$集合来扩充集合 $C$ 和 $M$. 根据$\{N_p\}_{p=1}^{\lambda}$集合扩充约束集 $C$ 和 $M$,扩充的方法为:若有 $x_i\in N_p$ 且$(x_i,x_j)\in C$,则对于$\forall x_t\in N_p$,将$(x_t,x_j)$添加到 $C$ 中;将$\{N_p\}_{p=1}^{\lambda}$的每个 $N_p$ 中所有样本点两两构成一个约束对,添加到 $M$ 中.

## 3　实验结果分析

使用 UCI 数据集对改进前后的算法进行测试,比较改进前后算法的优劣及加入主动学习后的半监督聚类的性能.

### 3.1　评价标准

本文采用两个常用的指标对聚类算法进行性能评估:归一化互信息[15](normalized mutual information,NMI)和迭代次数(iterative times,ITs).

NMI 常用于度量两个聚类结果的相似度. 设 $\boldsymbol{A},\boldsymbol{B}$ 分别为程序的聚类结果和真实的聚类结果:

$$I(\boldsymbol{A},\boldsymbol{B})=H(\boldsymbol{A})+H(\boldsymbol{B})-H(\boldsymbol{A},\boldsymbol{B}). \tag{4}$$

若用 $P_{\boldsymbol{A}}(a)$、$P_{\boldsymbol{B}}(b)$表示 $\boldsymbol{A}$、$\boldsymbol{B}$ 的概率分布,$P_{\boldsymbol{AB}}(a,b)$表示 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的联合概率分布,则

$$H(\boldsymbol{A})=-\sum_{a}P_{\boldsymbol{A}}(a)\ln P_{\boldsymbol{A}}(a), \tag{5}$$

$$H(\boldsymbol{B})=-\sum_{b}P_{\boldsymbol{B}}(b)\ln P_{\boldsymbol{B}}(b), \tag{6}$$

$$H(\boldsymbol{A},\boldsymbol{B})=-\sum_{a,b}P_{\boldsymbol{AB}}(a,b)\ln P_{\boldsymbol{AB}}(a,b), \tag{7}$$

式中,$H(\boldsymbol{A})$是 $\boldsymbol{A}$ 向量的熵,$I(\boldsymbol{A},\boldsymbol{B})$是 $\boldsymbol{A},\boldsymbol{B}$ 两向量的互信信息.

根据联合熵和个体熵之间的关系,归一化互信

$$\mathrm{NMI}=\frac{H(\boldsymbol{A})+H(\boldsymbol{B})}{H(\boldsymbol{A},\boldsymbol{B})},\tag{8}$$

NMI 在[0,1]范围内,代表和真实聚类结果的相似度,值越大表示聚类的准确性越高.

本文使用算法结果收敛时的迭代次数作为该算法的收敛速度. 迭代次数越少,说明算法在较少的迭代次数下得到了一个收敛的结果,即收敛速度更快.

### 3.2　实验结果及分析

本节实验使用了 4 组 UCI[16] 数据集:Iris、Wine、Seeds 和 Ecoli,每种数据集的信息如表 1 所示. 本节所用的实验结果数据均为程序运行 20 次的平均结果,具有一定的代表性. 图 2~图 9 所示为在不同的数据集上给定不同的约束对时算法的 NMI 和 ITs 值.

**表 1　UCI 数据集**

**Table 1　UCI dataset**

| 样本集 | 样本数目 | 属性数目 | 真实簇数目 |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Wine | 178 | 13 | 3 |
| Seeds | 210 | 7 | 3 |
| Ecoli | 336 | 7 | 8 |

图 2 和图 3 是 4 种算法在 Iris 数据集上的测试结果,其中 PCKmeans_I 为改进的 PCKmeans 算法,APCKmeans_I 为改进的 APCKmeans 算法. 随着给定约束对数的增加,实验结果的准确率也进一步提高,迭代速度也越来越快. 这是因为当给定的约束对数越多即监督信息越多时,得到的初始中心点越好,在指导无监督信息的样本进行聚类时可以使样本点更快地找到其最适合的簇. 改进后的算法明显比未改进算法的实验结果准确性要高,且迭代速度也有所提高. 可见,加了主动学习的半监督聚类比事先给定约束对的半监督聚类的准确性提升更快. 而图 2 中,随着约束对数的增加,APCKmeans 算法的准确性并未提高,这是由于前文所提问题导致的. 而 APCKmeans_I 算法已解决了该问题,故算法准确性和收敛速度均有所提高.

图 4~图 9 是 4 种算法分别在数据集 Wine,Seeds 和 Ecoli 上的 NMI 和 ITs 的值. 同样,随着约束对数的增加,4 种算法结果的 NMI 在不同数据集上均有不同程度的提升,在多数数据集上,迭代速度也会有一定的提升. 通过对比 NMI 值,改进后的算法明显比未改进算法的准确性更高且迭代速度更快,且加上主动学习的半监督聚类比随机产生约束对的半监督聚类在给定相同约束对的情况下 NMI 值更高,即准确性更高,迭代速度更快.

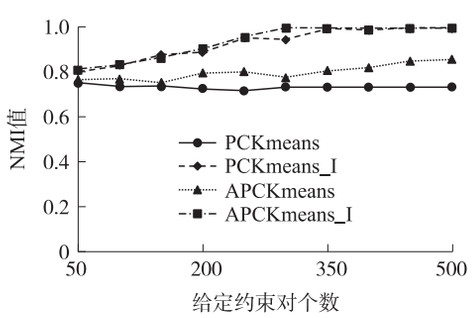


**图 2　在 Iris 数据集上给定不同约束对的 NMI 值**

**Fig. 2　NMI on Iris with different number of constraints**

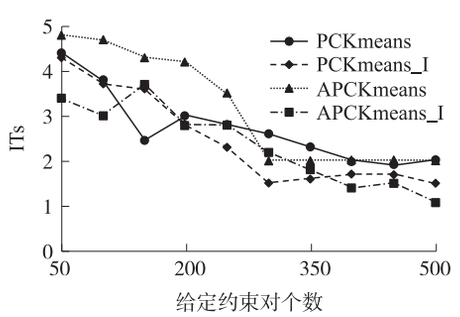


**图 3　在 Iris 数据集上给定不同约束对的 ITs 值**

**Fig. 3　ITS on Iris with different number of constraints**

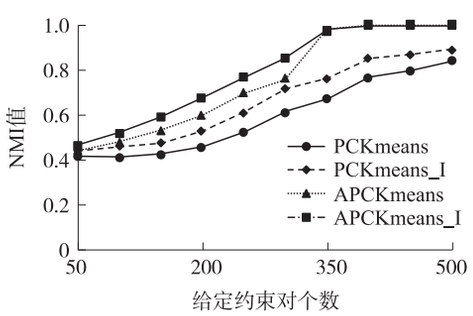


**图 4　在 Wine 数据集上给定不同约束对的 NMI 值**

**Fig. 4　NMI on Wine with different number of constraints**

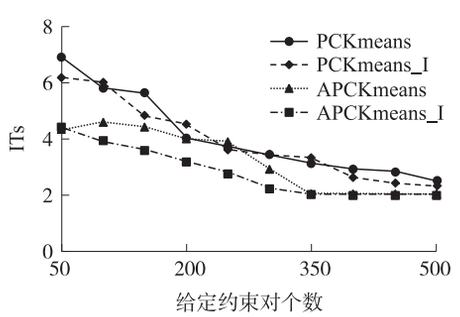


**图 5　在 Wine 数据集上给定不同约束对的 ITs 值**

**Fig. 5　ITs on Wine with different number of constraints**

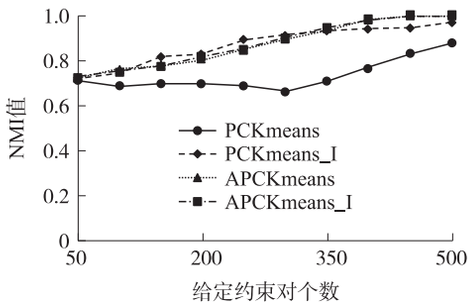


**图 6　在 Seeds 数据集上给定不同约束对的 NMI 值**

**Fig. 6　NMI on Seeds with different number of constraints**

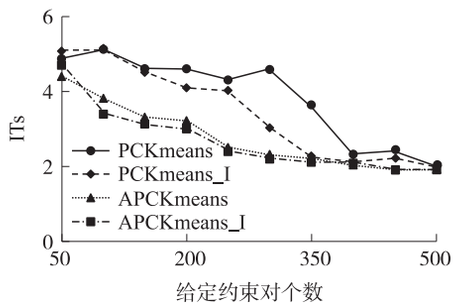


**图 7　在 Seeds 数据集上给定不同约束对的 ITs 值**

**Fig. 7　ITs on Seeds with different number of constraints**

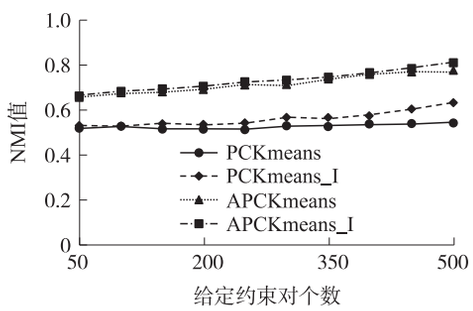


**图 8　在 Ecoli 数据集上给定不同约束对的 NMI 值**

**Fig. 8　NMI on Ecoli with different number of constraints**

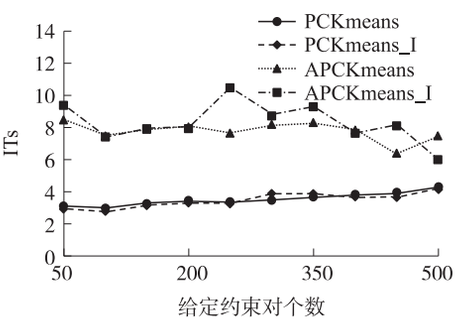


**图 9　在 Ecoli 数据集上给定不同约束对的 ITs 值**

**Fig. 9　ITs on Ecoli with different number of constraints**

## 4　结束语

本文提出了一种优化成对约束半监督聚类约束对的指派方式的改进策略. 通过在 UCI 基准数据集上测试,应用该改进策略的 PCKmeans_I 算法和 APCKmeans_I 算法相比于原算法,实验结果显示在收敛速度和 NMI 值上均有一定程度的提升. 后续研究将把本文中提到的两种改进算法在分布式集群上运行,提高其并行性,使其在处理大规模数据时更有效.

### [参考文献](References)

[1] BASU S,BANERJEE A,MOONEY R J. Semi-supervised clustering by seeding[C]//Nineteenth International Conference on Machine Learning. San Fransisco,USA:Morgan Kaufmann Publishers Inc,2002:19-26.

[2] WAGSTAFF K,CARDIE C. Clustering with instance-level constraints[C]//Seventeenth International Conference on Machine Learning. Stanford,CA,USA,2000:1103-1110.

[3] 王玲,薄列峰,焦李成. 密度敏感的半监督谱聚类[J]. 软件学报,2007,18(10):2412-2422.
WANG L,BO L F,JIAO L C. Density-sensitive semi-supervised spectral clustering[J]. Journal of software,2007,18(10):2412-2422.(in Chinese)

[4] 尹学松,胡恩良,陈松灿. 基于成对约束的判别型半监督聚类分析[J]. 软件学报,2008,19(11):2791-2802.
YIN X S,HU E L,CHEN S C. Discriminative semi-supervised clustering analysis with pairwise constraints[J]. Journal of software,2008,19(11):2791-2802.(in Chinese)

[5] 肖宇,于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报,2008,19(11):2803-2813.
XIAO Y,YU J. Semi-supervised clustering based on affinity propagation algorithm[J]. Journal of software,2008,19(11):2803-2813.(in Chinese)

[6] 方玲,陈松灿. 结合特征偏好的半监督聚类学习[J]. 计算机科学与探索,2015,9(1):105-111.
FANG L,CHEN S C. Semi-supervised clustering learning combined with feature preferences[J]. Journal of frontiers of computer science and technology,2015,9(1):105-111.(in Chinese)

[7] 张俊溪,吴晓军,蒋江红. 复杂分布数据的半监督阶段聚类[J]. 计算机科学与探索,2016,10(7):1003-1009.
ZHANG J X,WU X J,JIANG J H. Semi-supervised stage clustering for complex distribution data[J]. Journal of frontiers of computer science and technology,2015,10(7):1003-1009.(in Chinese)

[8] 高莹,刘大有,齐红,等. 一种半监督 $K$ 均值多关系数据聚类算法[J]. 软件学报,2008,19(11):2814-2821.
GAO Y,LIU D Y,QI H,et al. Semi-supervised *K*-means clustering algorithm for multi-type relational data[J]. Journal of software,2008,19(11):2814-2821.(in Chinese)

[9] BASU S,BANERJEE A,MOONEY R J. Active semi-supervision for pairwise constrained clustering[C]//Proceedings of the SIAM International Conference on Data Mining. Lake Buena Vista,FL,2004:333-344.

[10] XIONG S,AZIMI J,FERN X Z. Active learning of constraints for semi-supervised clustering[J]. IEEE transactions on knowledge and data engineering,2013,26(1):43-54.

[11] GREENE D,CUNNINGHAM P. Constraint selection by committee:an ensemble approach to identifying informative constraints for semi-supervised clustering[M]//Machine Learning:ECML 2007. Berlin Heidelberg:Springer-Verlag,2007:140-151.

[12] HUANG R,LAM W. Semi-supervised document clustering via active learning with pairwise constraints[C]//IEEE International Conference on Data Mining. Omaha,Nebraska,USA:IEEE,2007:517-522.

[13] MALLAPRAGADA P K,JIN R,JAIN A K. Active query selection for semi-supervised clustering[C]//International Conference on Pattern Recognition. Anchorage,AK,USA:IEEE,2008:1-4.

[14] XU Q, DESJARDINS M, WAGSTAFF K L. Active constrained clustering by examining spectral eigenvectors[C]//International Conference on Discovery Science. Berlin Heidelberg:Springer-Verlag,2005:294-307.

[15] WU M R, SCHOLKOPF B. A local learning approach for clustering[C]//Proceedings of the Conference on Neural Information Processing Systems. Cambridge,MA,USA:MIT Press,2006:1529-1536.

[16] ASUNCION A,NEWMAN D. UCI machine learning repository[EB/OL][2014-02-18]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[责任编辑:严海琳]