基于图收缩的半监督聚类算法

兰远东[1,2] 邓辉舫[1] 陈 涛[1]
(华南理工大学计算机科学与工程学院 广州 510640)[1] (惠州学院计算机科学系 惠州 516007)[2]

摘 要 为了在只有少量已知标记的数据集中获得较好的聚类效果,提出了一种基于图收缩的半监督聚类算法。首先将整个样本空间中的数据表达为一个带权图,再根据给出的must-link约束,对图进行边收缩的修改,进而增强must-link约束。在此基础上引入图拉普拉斯算子,结合cannot-link约束将样本空间投影到一个特征子空间。最后在子空间上进行聚类分析。实验结果表明,该方法不仅提高了对复杂数据的聚类结果,而且在约束对数量较少时也能获得较好的结果。

关键词 半监督聚类,图拉普拉斯算子,聚类分析,样本空间,机器学习
中图法分类号 TP391.4 **文献标识码** A

Semi-supervised Clustering Algorithm Based on Graph Contraction

LAN Yuan-dong[1,2] DENG Hui-fang[1] CHEN Tao[1]
(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China)[1]
(Department of Computer Science, Huizhou University, Huizhou 516007, China)[2]

Abstract In order to get a good clustering performance in data set with a small number of labeled samples, a semi-supervised clustering algorithm based on graph contraction was proposed in this paper. At first, the whole data in sample space was represented as an edge-weighted graph. Then the graph was modified by contraction according to must-link constraints and graph theory. On this basis, we projected sample space into a subspace by combining graph laplacian with cannot-link constraints. Data clustering was conducted over the modified graph. Experimental results show that the method indeed reaches its goal for complex datasets, and it is acceptable when there has small amount of pairwise constraints.

Keywords Semi-supervised cluster, Graph laplacian, Clustering analysis, Sample space, Machine learning

1 引言

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程称为聚类。由聚类所生成的簇(cluster)是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象相异。聚类分析又称群分析,它是研究(样品或指标)分类问题的一种常用统计分析方法[1]。

半监督聚类能同时利用标记(labeled)数据的信息和未标记(unlabeled)数据中的隐含信息,来达到比仅使用一种数据信息更好的学习效果,因此,近年来它获得研究者的广泛关注[2]。半监督聚类方法大致可以分为以下3种[6]:(1)基于约束的半监督聚类;(2)基于距离的半监督聚类;(3)混合半监督聚类。基于约束的半监督聚类,是用一些样本对之间的约束条件来引导聚类过程[7]。基于距离的半监督聚类,是在某种距离测度方法下,用某种距离条件约束来引导样本空间的聚类过程[8,9]。混合半监督聚类,是用某种概率框架综合上述两种方法的聚类分析方法[10]。

基于图的学习,是指选定某种测度方法来度量样本间的相似性后,整个样本空间的数据可以用一个带权图来表达[3]。然后在图上根据图论的一些相关理论,寻找整个样本空间的一种划分(聚类)。文献[4]利用图的着色理论来对样本空间进行聚类分析。其他一些方法也有利用图的割集、图的最大流以及谱图等理论,比如文献[5]给出的谱聚类方法。

本文提出了一种基于图收缩的半监督聚类算法(Semi-Supervised Clustering Algorithm Based on Graph Contraction, SSCGC)。其算法分为以下4个步骤:(1)将样本空间的所有数据构建为一个带权图,图上的点代表样本空间中的数据,边代表连接的两个样本之间的相似性;(2)根据已知约束条件(must-link)按照图论中的图收缩理论对图进行收缩[11];(3)按照谱图理论中的图拉普拉斯算子(Graph Laplacian)来体现已知约束条件(cannot-link)[5];(4)综合步骤(2)、(3)可得到聚类目标函数。

2 图的构建与收缩

2.1 图的构建

设X是样本的集合,$|X|$是集合X的基数。令$G(V,E,W)$表示一个带权图,其中V表示结点集,即数据点集(包括标记数据和无标记数据);E为边的集合,可以看作是V上的

到稿日期:2011-07-12 返修日期:2011-10-20 本文受国家自然科学基金(61170193)资助。
兰远东(1975—),男,博士生,主要研究方向为模式识别与机器学习,E-mail:lanyuandong@gmail.com。

二元关系 $V\times V$ 的子集，边 $(v_i,v_j)\in E$ 当且仅当顶点 v_i,v_j 之间有边相连；当 $|X|=|V|=n$ 时，W 是一个 $n\times n$ 的矩阵，其中的每一个元素 w_{ij} 的值就是边 (v_i,v_j) 的权值，表示结点 i,j 之间的相似度，且 $w_{ij}=w_{ji}$。

在构建好样本空间的图之后，半监督聚类就是利用给出的一些约束条件(监督信息)来引导聚类过程，即半监督聚类可以定义如下：

在一个给定的样本空间 X 上，根据一些给定的约束条件，找到样本空间的一个划分 $T=\{t_1,\cdots,t_k\}$，$t_i(i=1,\cdots,k)$ 为某个聚类，其中 k 为聚类的数目。

约束条件的形式可以是多种多样的。根据前人的工作[7-9]，本文主要考虑以下两种约束条件：must-link 约束和 cannot-link 约束，其具体定义如下：

给定样本空间 X 和样本的一个划分 $T=\{t_1,\cdots,t_k\}$，must-link 约束 C_{ML} 和 cannot-link 约束 C_{CL} 是满足下面要求的一些顶点对的集合：

$$(x_i,x_j)\in C_{ML}\Rightarrow\exists\, t\in T,(x_i\in t\wedge x_j\in t) \tag{1}$$

$$(x_i,x_j)\in C_{CL}\Rightarrow\exists\, t_a,t_b\in T,t_a\neq t_b,(x_i\in t_a\wedge x_j\in t_b) \tag{2}$$

也就是说 must-link 约束指定两个顶点必须属于同一个聚类，而 cannot-link 约束指定两个顶点不能属于同一个聚类。在图上的反映是，must-link 约束指定的顶点对之间必须有边相连，而 cannot-link 约束指定的顶点对间不能有边相连。

2.2 图的收缩

在选定了某种距离测度(如欧几里德距离)来描述样本之间的相似性，并根据约束条件(must-link 和 cannot-link)创建好样本点的图(G)之后，根据图论中的边收缩方法来修改图 G。因为样本空间的每一个样本 $x\in X$ 都与图 G 上的某个顶点 $v\in V$ 相对应，因此，下面用样本空间的集合 X 来代替图 G 的顶点集 V。本文提出的基于图收缩的半监督聚类算法的思路如图 1 所示。

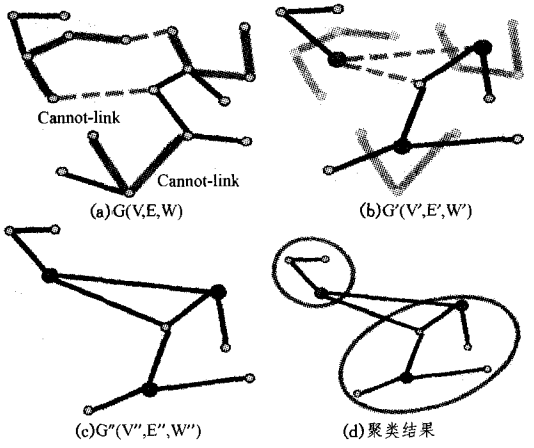

图 1 基于图收缩的聚类算法概要

图 1 中(a)为由样本空间及给定的约束条件构建的带权图，权值由 W 存储；(b)为经过收缩 must-link 边后得到的图 G'；(c)为融入 cannot-link 后得到的图 G''；(d)为聚类的最终结果。

式(1)中给出的 must-link 约束 C_{ML} 是满足传递性的，即对任意两个顶点对 $(x_i,x_j)\in C_{ML}$ 和 $(x_j,x_l)\in C_{ML}$，x_i 和 x_l 都应该属于同一个聚类中。本文按照图论中的边收缩方法来对图 G 进行修改得到图 G'，边收缩的定义如下[11]：

设 $e=(x_i,x_j)$ 是图 $G=(X,E)$ 的一条边，用 G/e 表示从 G 中删除 e 后，将 e 的两个端点用一个新的顶点 x_e 代替，使 x_e 关联 e 以外 x_i,x_j 关联的所有边，称为边 e 的收缩。边 e 收缩后得到的图记为 $G'=(X',E')$，其满足：

$$X'=(X\backslash\{x_i,x_j\})\cup\{x_e\} \tag{3}$$

$$E'=\{(u,v)\in E|\{u,v\}\cap\{x_i,x_j\}=\Phi\}\cup\{(x_e,u)|(x_i,u)\in E\backslash\{e\}\ \text{or}\ (x_j,u)\in E\backslash\{e\}\} \tag{4}$$

通过边收缩将 must-link 边收缩为一个顶点 x_e，并且 x_e 与 G 中原来 x_i,x_j 相邻的所有顶点都相邻。

因为整个样本空间 X 和已知的约束条件被表示为一个带权图 $G(V,E,W)$，所以将边 $e=(x_i,x_j)\in C_{ML}$ 收缩为一个顶点 x_e 后，必须重新计算图 G/e 的权值矩阵。权值矩阵 W 表示的是图 G 上顶点之间的相似性，因此重新计算权值的时候一定要保持原始的相似性，而 must-link 约束就是对相似性的加强。本文按照下面的方法来计算 G/e 中两个顶点间的权值：

$$w(x_e,u)=\max(w(x_i,u),w(x_j,u));\quad (x_i,u)\in E\ \text{or}\ (x_j,u)\in E \tag{5}$$

$$w(u,v)=w(u,v)\quad \text{其他} \tag{6}$$

式(5)保证了权值的单调递增，使得边收缩后相似性被 must-link 加强，式(6)使得没有受到边收缩影响的顶点对之间的权值保持不变。权值的修改如图 2 所示。

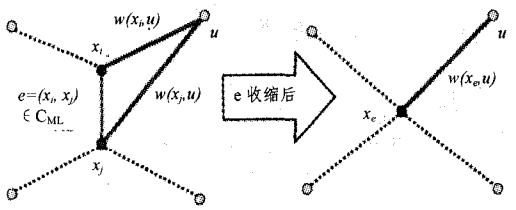

图 2 must-link 边的收缩

递归地将所有 must-link 边收缩后，图 G 被修改为图 $G'=(X',E',W')$，且令 $n'=|X'|$。

2.3 Cannot-link 的体现

为了在聚类过程中体现 cannot-link 约束，将聚类过程正则化为式(7)所示的目标函数，这样就将一个聚类问题转化为一个目标函数的优化问题。

$$J=\frac{1}{2}\{w'_{ij}\|f_i-f_j\|^2-\lambda\sum_{u,v\in C'_{CL}}w'_{uv}\|f_u-f_v\|^2\} \tag{7}$$

式中，i 和 j 指的是被收缩后的图 G' 中的任意两个顶点 x_i 和 x_j，w'_{ij} 是 (x_i,x_j) 在 G' 中的权值，C'_{CL} 是图 G' 中的 cannot-link 约束的集合，f_i 指的是聚类过程中 x_i 获得的标记值，$\lambda\in[0,1]$ 是控制超参数。$w'_{ij}\|f_i-f_j\|^2$ 保证了顶点在聚类过程中的平滑性。x_i 和 x_j 越相似，w'_{ij} 就越大，而 $\|f_i-f_j\|^2$ 就越小，从而 $w'_{ij}\|f_i-f_j\|^2$ 在总体上保证了图的平滑性。$\lambda\sum_{u,v\in C'_{CL}}w'_{uv}\|f_u-f_v\|^2$ 体现了 C'_{CL}(cannot-link 约束)在优化过程中的作用，设置 $\lambda\in[0,1]$，保证了式(7)是一个凸函数。调整超参数 λ 的大小，实际上就是把 C'_{CL} 作为优化过程中的一种软约束对待。

对于式(7)，根据谱图理论，可以做如下推导：

$$J=\frac{1}{2}\{w'_{ij}\parallel f_i-f_j\parallel^2-\lambda\sum_{u,v\in C_{CL}}w'_{uv}\parallel f_u-f_v\parallel^2\}$$
$$=f'D''f-f'w''f=f'L''f \tag{8}$$

式中，f' 是向量 f 的转置向量，L''是非正则化的图拉普拉斯算子，D''和 L''的具体含义见下面的推导。

首先根据 cannot-link 约束的集合 C'_{CL} 新建一个 $n'\times n'$的矩阵 C'（n'为 G'中的顶点数），其中的元素值为：

$$C'_{uv}=\begin{cases}1, & (x_u,x_v)\in C'_{CL}\\ 0, & \text{else}\end{cases} \tag{9}$$

然后令：

$$W^c=C'\otimes W',W''=W'-\lambda W^c \tag{10}$$

式中，$\otimes$为矩阵对应元素之间的乘法运算，这样 W^C 中除 cannot-link 约束对应的元素保持和 W'中相同外，其它元素全为 0。实际上就是通过上面的设定，将图 G'转化为图 G''，其对应的权值矩阵为 W''。

现在就可以给出式(8)中 D''和 L''的具体含义：

$$D''=diag(d''_1,\cdots,d''_n),L''=D''-W'' \tag{11}$$

其中，

$$d_i=\sum_{j=1}^{n'}w'_{ij},d_i^c=\sum_{j=1}^{n'}w^c_{ij},d''_i=d_i-\lambda d_i^c \tag{12}$$

然后根据文献[5]中的推导和式(11)、式(12)得到下面的优化函数：

$$J_{sym}=\sum_{i,j}w''_{ij}\left\|\frac{f_i}{\sqrt{d''_i}}-\frac{f_j}{\sqrt{d''_j}}\right\|^2 \tag{13}$$

从文献[5]可知，在图 G''上最小化式(13)，相当于求解 $L''h=\alpha D''h$ 的广义特征值问题，h 是广义特征向量，α 是广义特征值。

2.4 算法描述

综上所述，可以得到下面给出的算法。

算法 基于图收缩的半监督聚类(SSCGC)

Input 1 G(X,E,W);//样本空间的带权图表示

Input 2 C_{ML};//must-link 约束

Input 3 l;//子空间的维度

Input 4 k;//聚类数目

1. for each e$\in C_{ML}$ do
2. 收缩边 e 得到图 G/e
3. end for//得到收缩后的图 G'=(X',E',W')
4. 根据式(9)—式(12)计算 C'_{uv},W^c,W'',D''
5. $L''_{sym}=I-D''^{-\frac{1}{2}}W''D''^{-\frac{1}{2}}$
6. 用从小到大的非零特征值找出 L''_{sym}的 l 个特征向量，得到 H={h_1,…,h_l}
7. 用 spherical K-means 算法对数据聚类
8. 返回数据聚类

算法中的 1—3 行是根据 must-link 约束对图 G 进行边收缩，得到图 G'。4—6 行是对式(13)进行优化求解，其实际上是将问题转化为求解正则化的图拉普拉斯算子，也就是算法的第 5 行。对式(13)的优化求解，分别就相当于把样本空间 X 通过谱嵌入投影到子空间 $H=\{h_1,\cdots,h_l\}$。第 7 行是选择某种聚类算法（此处选择 spherical K-means[12]）对子空间中的数据进行聚类分析。第 8 行返回聚类结果。

3 实验结果与分析

本文采用比较流行的 20 Newsgroup data(20NG)作为实验数据集[6]，在该数据集上聚类，就相当于对文档进行分类。实验中，将 20NG 分为 3 组，如表 1 所列。

表 1 20 Newsgroup 数据集

数据集	包含的新闻组
Multi5	comp. graphics, rec. motorcycles, rec. sport. baseball, sci. space, talk. politics. mideast
Multi10	alt. atheism, comp. sys. mac. hardware, misc. forsale, rec. autos, rec. sport. hockey, sci. crypt, sci. med, sci. electronics, sci. space, talk. politics. guns
Multi15	alt. atheism, comp. graphics, comp. sys. mac. hardware, misc. forsale, rec. autos, rec. motorcycles, rec. sport. baseball, rec. sport. hockey, sci. crypt, sci. electronics, sci. med, sci. space, talk. politics. guns, talk. politics. mideast, talk. politics. misc

从每组中随机选择 50 个文档形成一个数据集，重复抽取，在每一组上选出 10 个数据集。对每一个数据集，用 PorterStemmer 和 MontyTagger 工具去除文档句子中的标点符号，并且挑选出 2000 个互信息(mutual information)最大的单词。

使用归一化互信息(Normalized Mutual Information, NMI)[13]，来对每一个数据集的聚类效果做评价，NMI∈[0,1]，其值越大，聚类效果越好。NMI 的定义如下：

$$NMI=\frac{I(\hat{T},T)}{(H(\hat{T})+H(T))/2}(\in[0,1]) \tag{14}$$

NMI 是聚类精度的一种反映，其中$\hat{T}$，T 是两个随机向量，$\hat{T}$是聚类过程中获得的值，T 是真实的值，$H(T)$是香农熵。

由于本文提出的基于图收缩的半监督聚类算法从根本上来说就是基于划分的聚类算法，因此作者选择了同样是基于划分的聚类算法 SCREEN[6]和 PCP[9]来做对比实验，并且在实验的时候，假定聚类数目 k 已知。

SCREEN[6]通过将特征空间投影到一个使得协方差最大的子空间来实现半监督聚类，对于像文档这样的高维数据，计算协方差的过程相当耗时。因此文献[6]中先用主成分分析(Principal Component Analysis, PCA)来减少特征空间的维度，在对比试验中也采用相同的方法。而 PCP[9]的对比实验，完全按照原文中的方法实现。

参数设定：本文所提出的算法，在实现时牵涉到的主要参数有(1)约束的数目(number of constraints)，也就是 must-link 和 cannot-link 的顶点对数目。在实验中，随机从每个数据集中挑选出一些样本来生成 must-link 约束和 cannot-link 约束，设定$|C_{ML}|=|C_{CL}|$，从 10～100 不断变化其具体值，每次递增 10 个。(2)子空间的维度 l，实验中将其设置为聚类的数目 k。(3)超参数 λ 经反复实验，最后设置为 0.6。

对数据集中的每一个样本 x 都进行归一化，使得 $x'x=1$，并使用欧几里德距离来计算 SCREEN[6]中样本间的相似性。在 PCP 和 SSCGC 中使用余弦相似性来构建初始的相似矩阵。对于样本空间的图表达，按照文献[9]中相同的方法，构建 m 近邻图(m-nearest neighbor graph)，并设置 $m=10$。

实验过程：在每一个数据集上，对每个不同数目的约束，重复执行 10 次聚类。约束$|C_{ML}|$和$|C_{CL}|$的数量从 10 变化到 100，共取了 10 个不同的值，每个值做 10 次聚类，也就是将算

法一共执行了100次。在每一个数据分组上生成了10个数据集，再对这10个数据集，按照上面相同的过程完成实验，最后给出的是在每个具体的约束数量时，每个组中的平均聚类效果。实验结果如图3和图4所示。

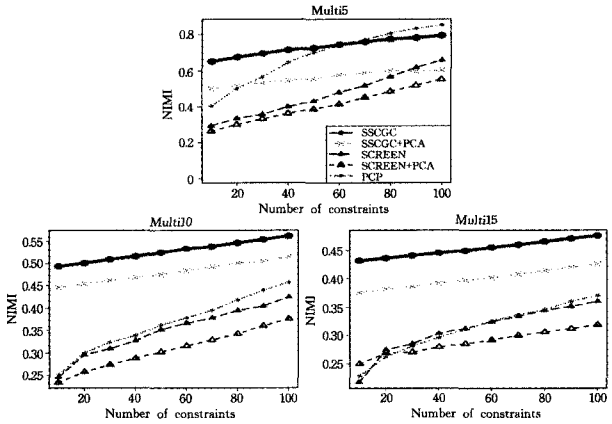

图3　在20-Newsgroup上的聚类结果(NMI)

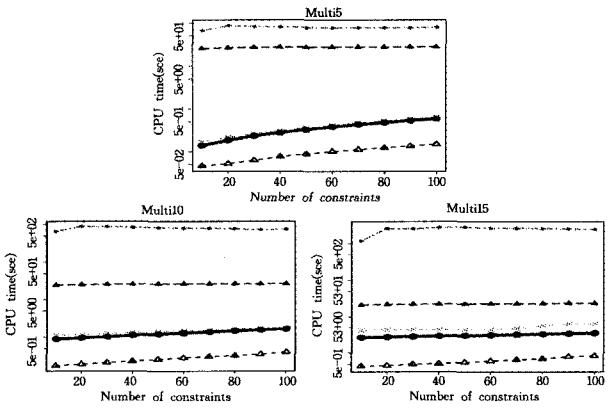

图4　在20-Newsgroup上的聚类时间

在图3的图例中，SCREEN+PCA是数据集先通过PCA预处理过后再应用SCREEN的情况，同样SSCGC+PCA也是先通过PCA预处理后再应用SSCGC的情况。为了显示的简洁，图3和图4中的所有图共用一个图例，如图3中所示。

从图3所示的实验结果可以看出，本文所提算法的NMI明显要优于PCP和SCREEN，平均高出约0.2。只有在Multi5中当约束在60以上时，PCP的NMI才略高于SSCGC。这是因为SSCGC不但使用图边收缩来加强must-link约束，而且还使用基于图的半监督学习的正则化框架将cannot-link作为一种权重可由参数λ调节的软约束对待。虽然在约束增多的情况下，PCP和SSCGC的聚类效果相当，但是从图4可以看出，PCP的聚类时间耗费要远远多于SSCGC。通过PCA预处理后使用SCREEN，会极大地提高聚类速度。从实验可知，本文提出的算法从聚类时间和聚类精度上来说都是比较好的；而且从图3和图4可以看出，本文提出的算法只需要较少的约束(监督信息)就能取得不错的聚类效果，这使得该算法能广泛应用于计算机辅助医学图像分析、Web网页推荐等众多领域。这些领域有一个共同的特点，即收集大量未标记的(unlabeled)示例已相当容易，而获取大量有标记的示例则相对较为困难，因为获得这些标记可能需要耗费大量的人力物力。

结束语　本文提出了一种基于图收缩的半监督聚类算法，它将样本空间的整个数据集先表示为一个带权图，然后通过边收缩，根据图论知识对图修改，并引入图拉普拉斯算子，其实际上就把样本空间投影到了一个子空间，在子空间上进行聚类分析。对图的收缩并不是为了减少样本的数量，而是为了增强must-link约束。本文的算法虽然也使用了图拉普拉斯，但是是在经过边收缩的图上(相当于子空间上)使用图拉普拉斯。实验证明，本文提出的算法具有较好的聚类精度和较快的聚类速度。

参考文献

[1] Duda R O, Hart P E, Stork D G. Pattern Classification(2nd Edition)[M]. New York: John Wiley & Sons, 2001: 517-580

[2] Chapelle O, Schölkopf B, Zien A. Semi-Supervised Learning [M]. Cambridge: MIT Press, 2006: 1-18

[3] Zhu X. Semi-Supervised Learning with Graphs [D]. Pittsburgh, Pennsylvania, USA: Carnegie Mellon University, 2005: 5-8

[4] Elghazel H, Yoshida T, Deslandres V, et al. A new greedy algorithm for improving b-coloring clustering [J]. Lecture Notes in Computer Science, 2007, 45(38): 228-239

[5] von Luxburg. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395-416

[6] Tang Wei, Xiong Hui, Zhong Shi, et al. Enhancing semi-supervised clustering: A feature projection perspective[C]// Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2007: 707-716

[7] 尹学松，胡恩良，陈松灿，等. 基于成对约束的判别型半监督聚类分析[J]. 软件学报，2008，11(19)：2791-2802

[8] 魏莱，王守觉. 基于流形距离的半监督判别分析[J]. 软件学报，2010，21(10)：2445-2453

[9] Li Zhen-guo, Liu Jian-zhuang, Tang Xiao-ou. Pairwise constraint propagation by semidefinite programming for semi-supervised classification[C]// Proceedings of the 25th international conference on machine learning. New York: ACM, 2008: 576-583

[10] Basu S, Bilenko M, Mooney R J. A probabilistic framework for semi-supervised clustering[C]// Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. Washington: ACM, 2004: 59-68

[11] 屈婉玲，耿素云，张立昂. 离散数学[M]. 北京：高等教育出版社，2008：273-280

[12] Zhong Shi. Efficient online spherical k-means clustering[C]// IEEE international joint conference on neural networks. Montreal: IEEE, 2005: 3180-3185

[13] Strehl A, Ghosh J. Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Machine Learning Research, 2002, 3(3): 583-617