Computer Science and Application 计算机科学与应用, 2016, 6(12), 794-801
Published Online December 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.612095

Hans汉斯

Research of Large-Scale Water Pump Classification Based on Artificial Fish Swarm Mixed Clustering

Li Shan[1*], Dongliang Yang[2], Lei Yao[2], Wenwen Yu[1], Changyu Duan[1]

[1]Anhui Key Laboratory of Large Submersible Pump Equipment, Hefei Hengda Jianghai Pump Co., Ltd., Hefei Anhui
[2]Hefei Sanyijianghai Intelligent Technology Co., Ltd., Hefei Anhui
Email: *13955166706@139.com

Received: Dec. 8th, 2016; accepted: Dec. 21st, 2016; published: Dec. 29th, 2016

Abstract

Artificial Fish Swarm (AFS) is a stochastic search optimization algorithm with a fast convergence speed, and it is widely used in different areas due to its tolerance for descriptions and mechanism model of questions. In this paper, we propose a new clustering algorithm called Artificial Fish Swarm Mixed Clustering (AFSMC) by performing a combination of AFS and similarity based K-means. And we further propose a relative entropy based optimization method for AFSMC. Finally, we evaluate our proposal by conducting classification on the real data of large scale water pump, which is a new thought for analyzing the potential correlations among the different pumps. The experimental results show that our proposal outperforms the existing clustering approaches.

Keywords

Artificial Fish Swarm Algorithm, Relative Entropy, Mixed Clustering, Large Water Pump

基于人工鱼群混合聚类的大型水泵分类研究

单　丽[1*]，杨栋梁[2]，姚　磊[2]，于文文[1]，段长余[1]

[1]合肥恒大江海泵业股份有限公司，大型潜水电泵装备安徽省重点实验室，安徽　合肥
[2]合肥三益江海智能科技有限公司，安徽　合肥
Email: *13955166706@139.com

*通讯作者。

文章引用: 单丽, 杨栋梁, 姚磊, 于文文, 段长余. 基于人工鱼群混合聚类的大型水泵分类研究[J]. 计算机科学与应用, 2016, 6(12): 794-801. http://dx.doi.org/10.12677/csa.2016.612095

收稿日期：2016年12月8日；录用日期：2016年12月21日；发布日期：2016年12月29日

摘　要

人工鱼群是一种随机搜索优化算法，具有较快的收敛速度，对问题的机理模型与描述无严格要求，具有广泛的应用范围。本文在该算法的基础上，结合传统的K-means聚类方法，提出了一种新的人工鱼群混合聚类算法(AFSMC)，并给出相对熵混合聚类优化方法。文中将几种算法应用于大型水泵数据的分类，算例分析表明，基于相对熵的AFSMC算法在准确率与时间效益上都具有较明显的优越性。

关键词

人工鱼群算法，相对熵，混合聚类，大型水泵

1. 引言

大型水泵是一种重要的防水、治水设备，是水利、矿山等部门安全生产的重要保障。目前对于大型水泵的物联网建设及实时运行数据管理，主要是从使用方的角度，即是基于某一泵站若干台设备进行的，实现实时监控、智能预警、报警等功能[1] [2] [3]。而从大型水泵制造商的角度，若能建立动态数据库，掌握其所有水泵产品在全生命周期的运行状态，则一方面可以通过实时监控提升服务质量，另一方面还能利用大数据分析促进产品技术革新升级，具有重要的战略意义。而对多种产品规格，多种数据来源，海量规模的大型水泵数据进行大数据研究，一个关键任务就是对大型水泵数据准确、快速的聚类分析。

K-means 算法是一种经典的聚类分析算法[4] [5] [6]，该算法的最大优势在于简洁和快速，但主要不足之处在于聚类结果极大依赖聚类中心的选取，聚类过程易陷入局部最优等。而与之相对的，源于对鱼群行为特点的研究[7] [8]，由李晓磊等[9] [10]提出并得到了大量后续的关注与改进[11] [12] [13]的人工鱼群算法(AFS)则具有良好的克服局部极值、获取全局极值的能力，不受初始化的随机性影响，且对搜索空间具有一定自适应能力。刘白等[11]根据 K-means 算法与 AFS 算法的相关特点，发展了一种混合聚类算法，即先使用 AFS 算法做初始化优化，然后再进入主体的 K-means 算法流程，取得了一定的优化效果。

本文以 AFS 为主体，融合了 K-means 算法，提出了一种新的人工鱼群混合聚类算法(AFSMC)。并根据大型水泵数据特点，采用相对熵作为距离度量，对该算法做新一步的优化。利用合肥恒大江海泵业股份有限公司的真实数据，构建仿真实验对几种聚类方法进行验证，算例分析表明，采用相对熵度量的 AFSMC 兼具聚类的准确性和快速性，在大型水泵数据的聚类分析上具有很好的应用前景。

2. 人工鱼群混合聚类算法

2.1. K-means 聚类算法

K-means 是一种常用的聚类分析算法[4] [5] [6]，是一种基于划分的方法，需要预先指定聚类数目和聚类中心，它能够使聚类域中的所有样品到聚类中心距离的平方和最小。算法基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类，通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。假设要把样本集分为 k 个类别，算法描述如表 1 所示。

该算法的最大优势在于简洁和快速。算法的关键在于初始中心的选择和距离公式。一般 K-means 聚类算法选择欧氏距离作为最常见的距离度量，衡量的是多维空间中各个点之间的绝对距离。

Table 1. K-means clustering algorithm
表 1. K-means 聚类算法

1: 适当选择 k 个类的初始中心；

2: 在某次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类；

3: 利用均值等方法更新该类的中心值；

4: 对于所有的 k 个聚类中心，如果利用 2、3 的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

2.2. 人工鱼群算法

人工鱼群算法[9] [10] (AFS)是根据鱼类的活动特点提出的一种基于动物行为的自治体寻优模式。在水域中，鱼生存数目最多的地方一般就是该水域中富含营养物质最多的地方，依据这一特点模仿鱼群觅食、聚群、追尾等行为，这些行为在不同的条件下会相互转换。鱼类通过对行为的评价，选择一种当前最优的行为进行执行，到达食物浓度最高的位置，从而实现全局最优。

采用面向对象技术重构人工鱼模型，将人工鱼封装成变量和函数两部分。表 2、表 3 分别给出人工鱼模型中变量部分和函数部分定义。

人工鱼基本行为描述

① 随机行为 *Rand()*

是作为人工鱼在觅食行为 *Prey()*中的缺省行为，是在解空间中随机寻找大于自身状态的行为。

② 觅食行为 *Prey()*

通过视觉或味觉来感知水中的食物量或浓度进而来选择趋向。设人工鱼 i 当前状态为 X_i^t，在其感知范围内随机选择一个状态 X_{try}：

$$X_{\text{try}} = X_i^t + \text{Visual}\cdot\text{Rand}(\ ) \tag{1}$$

若 $Y_i < Y_{\text{try}}$，则向该方向前进一步：

$$X_i^{t+1} = X_i^t + \frac{X_{\text{try}} - X_i^t}{\left\| X_{\text{try}} - X_i^t \right\|}\cdot\text{Step}\cdot\text{Rand}(\ ) \tag{2}$$

反之，再重新随机选择状态 X_{try}，判断是否满足前进条件，反复尝试 *Try_number* 次后，若仍不满足前进条件，则随机移动一步：

$$X_i^{t+1} = X_i^t + \text{Visual}\cdot\text{Rand}(\ ) \tag{3}$$

③ 聚群行为 *Swarm()*

鱼在游动过程中为保证群体的生存和躲避危害，会自然的聚集成群。在算法中对该行为中的人工鱼有如下规定：一是尽量向邻近伙伴的中心移动；二是避免过分拥挤。设人工鱼当前状态为 X_i^t，探索当前邻域内($d_{ij} < Visual$)的伙伴数目 n_f 及中心位置 $\boldsymbol{X}_c$。若 $Y_c/n_f > \delta Y_i$，表明伙伴中心有较多食物且不太拥挤，则朝伙伴的中心位置方向前进一步：

$$X_i^{t+1} = X_i^t + \frac{X_c - X_i^t}{\left\| X_c - X_i^t \right\|}\cdot\text{Step}\cdot\text{Rand}(\ ) \tag{4}$$

否则，执行觅食行为。

④ 追尾行为 *Follow()*

追尾行为是一种向邻近的有着最高适应度的人工鱼追逐的行为，在寻优算法中可以理解为是向附近

Table 2. Parameters of artificial fish-AF
表 2. 人工鱼参数

总数	N
个体状态	$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$
人工鱼条数	*Total*
移动最大步长	*Step*
视野	*Visual*
尝试次数	*Try_Number*
拥挤度因子	δ
人工鱼个体间距	$d_{ij}=\left\|X_i-X_j\right\|$

Table 3. Function of AF
表 3. 人工鱼函数

食物浓度	$Y=f(X)$ (Y 为目标函数值)
觅食行为	*Prey()*
聚群行为	*Swarm()*
追尾行为	*Follow()*
随机行为	*Rand()*
行为评价	*Evaluate()*

的最优伙伴前进的过程。设人工鱼当前状态为 X_i^t，探索当前邻域内(d_{ij} < Visual)的伙伴中 Y_j 为最大值的伙伴 $\boldsymbol{X}_j$。若 $Y_c/n_f > \delta Y_i$，表明伙伴 $\boldsymbol{X}_j$ 的状态具有较高的食物浓度且周围不太拥挤，则朝 $\boldsymbol{X}_j$ 的方向前进一步：

$$X_i^{t+1}=X_i^t+\frac{X_j-X_i^t}{\left\|X_j-X_i^t\right\|}\cdot\text{Step}\cdot\text{Rand}(\ ) \tag{5}$$

否则，执行觅食行为。

2.3. 改进的人工鱼群混合聚类算法

由 K-means 模型给出分类模型结构，AFS 算法将给出的模型在解空间中循环迭代到最优，由此消除 K-means 初始化随机性的影响，发挥 AFS 全局寻优特性。一条鱼代表一个 K-means 分类模型，若干鱼组成一个鱼群，在所有训练实例域空间中，每个鱼都各自按照一定的原则寻找最优值，循环修正自己的 K-means 模型，并将最优值对应的模型记录在公告板，直到达到迭代次数以后，公告板上模型即为最后的分类最优模型。算法流程如图 1 所示。

每条鱼中包含的参数除了基本 AFS 算法参数以外，还有 K-means 聚类模型 X_{k*n}，是一个具有 n 个属性的实例，将其分为 k 类的描述。将此模型编码，如下：

$$X_{k*n}=\begin{bmatrix}x_{11} & \cdots & x_{1n}\\ \vdots & \ddots & \vdots\\ x_{k1} & \cdots & x_{kn}\end{bmatrix} \tag{6}$$

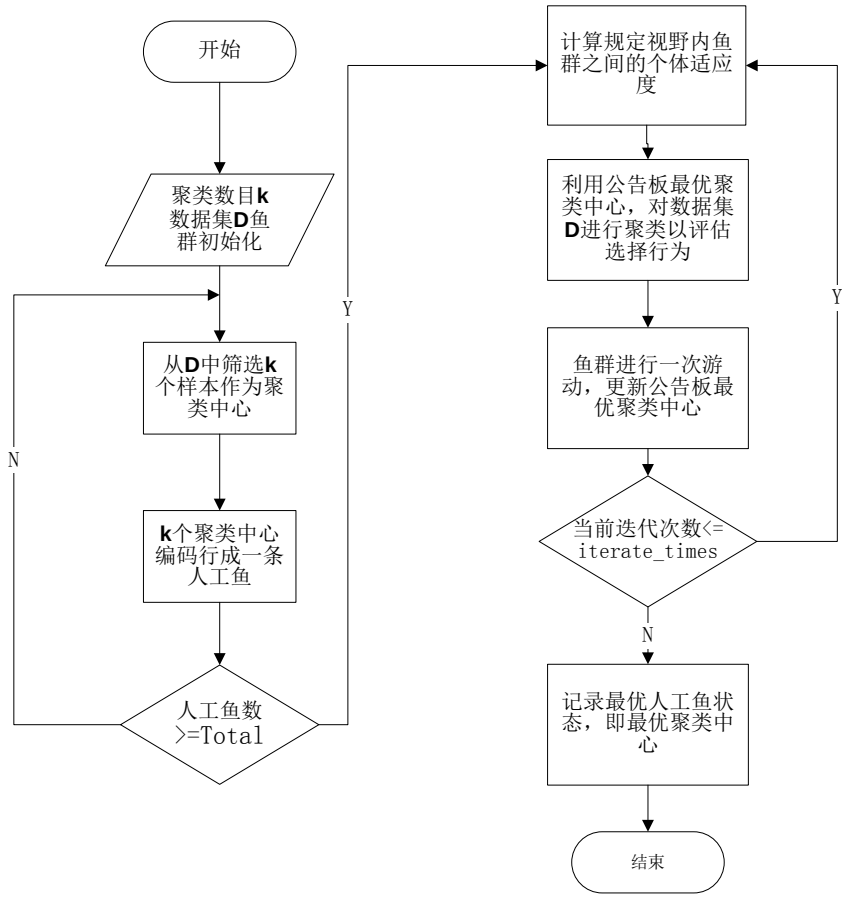

Figure 1. Flowchart of AFSMC algorithm
图 1. 人工鱼群混合聚类算法(AFSMC)流程图

人工鱼按照一定规则在解空间中游动，比较执行觅食、聚群、追尾等行为。为使得鱼群能够最大范围覆盖到最优解，鱼群需要保持一个合适的密度，即人工鱼之间要保持相应的距离。每条人工鱼以自己状态为中心，搜索视野范围 Visual 内的人工鱼并与其进行交互。与 K-means 聚类算法类似，可以选择欧氏距离作为距离度量，衡量各人工鱼之间的绝对距离。为了消除量纲选择造成的绝对数值的差异，使用相对熵[14] [15]描述人工鱼之间距离。

计算两条人工鱼的相对熵距离，假设 p，q 代表两条不同状态的人工鱼，每条鱼的 k 分类模型编码为 X_{k*n}，其中 k 为分类数，n 为数据维度； X_{k*n} 中第 $m(1\le m\le k)$ 分量结构为：

$$\begin{aligned}\left(\boldsymbol{X}_{\boldsymbol{mn}}\right)_p &= \left(x_{m1p}, x_{m2p}, \cdots, x_{mnp}\right)\\ \left(\boldsymbol{X}_{\boldsymbol{mn}}\right)_q &= \left(x_{m1q}, x_{m2q}, \cdots, x_{mnq}\right)\end{aligned} \tag{7}$$

则可计算两条鱼的概率分布

$$\begin{aligned}\boldsymbol{X}'_{\boldsymbol{mn}}\left(x\right)_p &= \left(\frac{x_{m1p}}{\sum_{i=1}^{n} x_{mip}}, \frac{x_{m2p}}{\sum_{i=1}^{n} x_{mip}}, \cdots, \frac{x_{mnp}}{\sum_{i=1}^{n} x_{mip}}\right)\\ \boldsymbol{X}'_{\boldsymbol{mn}}\left(x\right)_q &= \left(\frac{x_{m1q}}{\sum_{i=1}^{n} x_{miq}}, \frac{x_{m2q}}{\sum_{i=1}^{n} x_{miq}}, \cdots, \frac{x_{mnq}}{\sum_{i=1}^{n} x_{miq}}\right)\end{aligned} \tag{8}$$

进一步可计算得两条鱼之间的相对熵：

$$D\left(X_p \| X_q\right)=\sum_{m=1}^{k}\sum_{i=1}^{n} X'_{mn}\left(x_i\right)_p \log\frac{X'_{mn}\left(x_i\right)_p}{X'_{mn}\left(x_i\right)_q} \tag{9}$$

相对熵越大，两条人工鱼的状态模型差异越大；反之，差异越小。

用欧式距离或相对熵距离法判别视野范围 *Visual* 的其他人工鱼，在 *Visual* 范围内计算人工鱼基本行为，评估并执行最优的人工鱼基本行为，得到此人工鱼最新的参数模型 X_{k*n}，将实例在参数模型 X_{k*n} 下按照余弦相似度[14]分类，并计算其离差平方和[16]作为当前食物浓度，并将其与公告板上比较，更新公告板；继续迭代执行以上过程，达到迭代次数以后结束算法。

3. 混合聚类算法在大型水泵中的应用

3.1. 实验设置

1) 对大型水泵数据进行清理和集成，经过分析后选择其中 5 种不同类型水泵作为实验数据，汇总后数据集共包括 6 个属性，最后一个属性标记其真实类别；

2) 算法参数设置，人工鱼群的个体数 *Total* = 15，迭代次数 *iterate_times* = 2，尝试次数 *Try_number* = 30，聚类中心个数 *K* = 5，人工鱼群移动的最大步长，相对熵距离：*Step* = 60、视野范围 *Visual* = 200，欧式距离：*Step* = 10、视野范围 *Visual* = 2500，拥挤度因子 $\delta=9$，X_{k*n} 随机初始化为同一个值。

3.2. 实验结果与分析

分别实现基本的 K-means 算法，基于欧氏距离的 AFSMC 算法和基于相对熵的 AFSMC 算法，应用于大型水泵数据集。

三种算法得出的聚类中心结果分别如表 4、表 5 和表 6 所示。对比发现，三种算法给出的聚类中心大致吻合，证明了算法及仿真的可靠性。将聚类分析结果与水泵真实类型相比较，三种聚类方法的准确率分别为：K-means 算法 0.80；基于欧式距离 AFSMC 模型 0.91；基于相对熵 AFSMC 模型 0.96。证明三种算法均具有一定的有效性，且本文提出的基于相对熵 AFSMC 算法具有更高的准确率，并且其准确性也高于文献[11]中的相关结果。对两种不同距离度量的 AFSMC 算法进行比较，其收敛曲线如图 2 所示。可以看出，基于相对熵的 AFSMC 算法比基于欧式距离的 AFSMC 算法迭代收敛更快，其收敛精度也更高。

Table 4. The model of K-means clustering
表 4. K-means 算法聚类中心结果

中心	人工鱼(×1.0e+04)				
Center1	0.0576337	0.0545036	0.1276827	1.0260385	0.0097891
Center2	0.0219722	0.0571446	0.0502595	0.6014273	0.0057776
Center3	0.0003189	1.0512171	0.0130227	0.0380193	0.0257085
Center4	0.0968742	0.0311806	0.1210063	0.9876387	0.0089693
Center5	0.0253718	0.0575048	0.0587409	1.0003554	0.0043071

Table 5. AFSMC model: AF-distance based on Euclid
表 5. 基于欧式距离 AFSMC 聚类中心结果

中心	人工鱼(×1.0e+04)				
Center1	0.0651670	0.0314778	0.1071545	1.0332536	0.0086398
Center2	0.0227489	0.0486605	0.0470335	0.5873838	0.0054763
Center3	0.0003531	1.0664919	0.0133296	0.0390437	0.0264963
Center4	0.0506997	0.0706579	0.1404607	1.0190835	0.0104566
Center5	0.0181779	0.0852772	0.0683022	0.9989775	0.0050150

Table 6. AFSMC model: AF-distance based on relative-entropy
表 6. 基于相对熵 AFSMC 聚类中心结果

中心	人工鱼(×1.0e+04)				
Center1	0.0727024	0.0427193	0.1323379	0.9958218	0.0096672
Center2	0.0237896	0.0484455	0.0471812	0.6018432	0.0054278
Center3	0.0003610	1.0348928	0.0140044	0.0381490	0.0278141
Center4	0.0540214	0.0619098	0.1294160	1.0232266	0.0097573
Center5	0.0235000	0.0760597	0.0704117	0.9985826	0.0051498

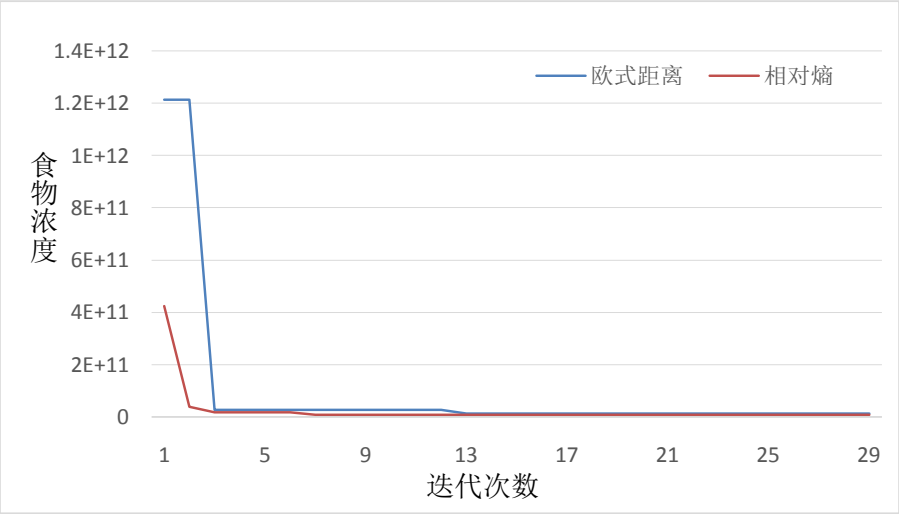

Figure 2. Comparison of convergence curves of two AFSMC algorithms
图 2. 两种人工鱼群混合算法(AFSMC)的收敛曲线对比

4. 总结

本文结合人工鱼群算法(AFS)和 K-means 聚类分析算法，提出一种新的人工鱼群混合聚类算法(AFSMC)，并给出利用相对熵作为距离度量的优化方法。将几种聚类算法应用于真实的大型水泵数据集，算例结果表明：

1) 几种算法在大型水泵的聚类分析上均具有相当的有效性；

2) AFSMC 在准确率上较传统的 K-means 聚类方法有明显的提高；

3) 利用相对熵优化的 AFSMC 比采用欧式距离的 AFSMC 具有更高的准确率和收敛速度。

致　谢

本项目为“安徽省自然科学基金资助项目，项目编号 1408085MKL82”，在此表示感谢。

参考文献 (References)

[1] 姚宇. 基于物联网的矿井主排水设备状态监测及寿命管理系统的开发[D]: [硕士学位论文]. 太原: 太原理工大学, 2016.

[2] 郭春春, 贺贵明. 大型泵站运行数据管理研究[J]. 工业控制计算机, 2006, 19(2): 15-16.

[3] 谭一川. 煤矿工业水泵自动化监控系统研究与应用[D]: [硕士学位论文]. 重庆: 重庆大学, 2009.

[4] Han, J.-W., Kamber, M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.

[5] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[6] 周涛, 陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用, 2012, 48(12): 100-111.

[7] Partridge, B.L. (1982) The Structure and Function of Fish Schools. *Scientific American*, **246**, 114-123. https://doi.org/10.1038/scientificamerican0682-114

[8] Tu, X. and Terzopoulos, D. (1994) Artificial Fishes: Physics, Locomotion, Perception, Behavior. *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, Orlando, 24-29 July 1994, 43-50. https://doi.org/10.1145/192161.192170

[9] 李晓磊, 钱积新. 人工鱼群算法: 自下而上的寻优模式[J]. 系统工程理论与实践, 2002, 22(3): 76-82.

[10] 李晓磊. 一种新型的智能优化算法——人工鱼群算法[D]: [博士学位论文]. 杭州: 浙江大学, 2003.

[11] 刘白, 周永权. 一种基于人工鱼群的混合聚类算法[J]. 计算机工程与应用, 2008, 44(18): 136-138.

[12] 陈祥生. 人工鱼群算法在聚类问题中的应用研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2010.

[13] 张梅凤. 人工鱼群智能优化算法的改进及应用研究[D]: [博士学位论文]. 大连: 大连理工大学, 2008.

[14] 吴军. 数学之美[M]. 北京: 人民邮电出版社, 2014.

[15] 易莉桦. 高维数据聚类算法的研究[D]: [硕士学位论文]. 秦皇岛: 燕山大学, 2012.

[16] Zhang, Z. and Zhou, J. (2012) Multi-Task Clustering via Domain Adaptation. *Pattern Recognition*, **45**, 465-473. https://doi.org/10.1016/j.patcog.2011.05.011