一种大数据环境下的新聚类算法

李 斌 王劲松 黄 玮
(天津理工大学计算机与通信工程学院 天津 300384)
(计算机病毒防治技术国家工程实验室 天津 300457)
(天津理工大学智能计算及软件新技术天津市重点实验室 天津 300191)

摘 要 提出了一种新的聚类算法NGKCA,该算法克服了经典聚类算法检测率和稳定性的不足,适用于解决大数据环境下的聚类问题。NGKCA聚类算法包括4个阶段:首先利用谱聚类NJW算法对大数据集进行列降维和数据归一化处理,其次引入对初始值不敏感的粒子群算法对数据集进行行降维从而选出临时的聚类中心集,接着通过全局Kmeans算法对最佳聚类中心集进行聚类以获取聚类中心点,最后使用粒子群算法对聚类中心点进行调整进而获取最终的聚类划分。在一些著名的机器学习数据集和国际标准的网络安全数据集KDDCUP99上进行实验,结果表明:提出的算法比谱聚类、Kmeans、粒子群、全局Kmeans等常见算法具有更好的稳定性和更高的检测率,与全局Kmeans算法相比具有更优的时间复杂度。

关键词 全局Kmeans,谱聚类,粒子群优化,聚类,KDDCUP99
中图法分类号 TP393.0 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.053

Novel Global Kmeans Clustering Algorithm for Big Data

LI Bin WANG Jin-song HUANG Wei
(School of Computer and Communication Engineering, Tianjin University of Technology, Tianjin 300384, China)
(National Engineering Laboratory for Computer Virus Prevention and Control Technology, Tianjin 300457, China)
(Tianjin Key Lab of Intelligent Computing & Novel Software Technology, Tianjin University of Technology, Tianjin 300191, China)

Abstract The clustering method for big data has attracted lots of interest in recent years. This paper proposed a novel global k-means clustering algorithm (NGKCA). The proposed clustering method comprises four phrases, namely row dimension reduction phrase, line dimension reduction phrase, global k-means clustering phrase and the adjustment of clustering center point. The row dimension reduction phrase is realized by means of spectral clustering method, while the line dimension reduction phrase is realized with the aid of particle swarm optimization. Both the row dimension reduction phrase and the line dimension reduction phrase are completed, and then the global k-means clustering phrase and the PSO phrase proceed. The experiments were carried out on some well-known machine learning data set and a standard network security data set KDDCUP99. Experimental results show that the proposed NGKCA leads to superior performance in comparison with some common algorithms reported in the literature and the time complexity of the NGKCA is better than the algorithm of global k-means.

Keywords Global Kmeans, Spectral clustering, PSO, Clustering, KDDCUP99

1 引言

近几年,随着互联网的发展以及信息传输量的急速提高,以大数据为主要特性的网络结构日益凸显出来。同时,网络是把双刃剑,它在方便我们日常工作和丰富我们生活的同时也为黑客提供了便利。扫描、蠕虫、DDoS拒绝服务等攻击变得更加恶劣。因此,在大数据面前,如何快速判断并挖掘出网络中的异常流量显得尤为重要。

KDDCUP99是国际标准的网络安全数据集,近年来很多安全研究人员对KDDCUP99数据集进行数据挖掘并提出了不少网络异常流量的检测算法。Macqueen提出的Kmeans算法是一种经典的聚类算法[1,2],但Kmeans算法的聚类结果依赖于初始值的选取,并且基于梯度下降进行搜索常常使算法陷入局部最优,这些缺陷极大限制了它的应用范围。姜大庆和夏士雄等[3]在对网络故障数据进行实验分析后,提出了一种基于谱聚类算法的网络故障检测算法。周文刚和陈雷霆等[4]指出了现有的基于端口映射的网络流量分类识别方法的准确性和效率受到制约,因此引入谱聚类方法,将网络流量分类识别问题转化为一个无向图的多路划分问题。但是谱聚类算法检测率相对较低,所以更多被用于精度要求不高的领域

到稿日期:2014-12-12 返修日期:2015-02-04 本文受国家自然科学基金项目(61272450),天津市科技支撑项目(14ZCZDGX00072)资助。
李 斌(1990-),男,硕士生,主要研究领域为网络安全,E-mail:898445911@qq.com;王劲松(1970-),男,博士,教授,博士生导师,主要研究领域为信息安全、计算机网络,E-mail:jswang@tjut.edu.cn;黄 玮(1980-),男,博士,副教授,主要研究领域为信息安全、人工智能,E-mail:huangwabc@163.com。

中。刘婧明、韩丽川[5]针对 Kmeans 算法容易陷入局部最优解的缺陷，利用粒子群算法快速收敛于全局最优解的特性，提出了基于粒子群的 Kmeans 聚类算法，这种方法虽然提升了聚类结果的稳定性，但是在检测率方面没有太大的提升。

本文提出了一种新的聚类算法(Novel Global K-means Clustering Algorithm, NGKCA)。该算法通过谱聚类的 NJW[8,10]算法对数据进行降维和归一化处理，并利用 PSO 快速全局收敛的特点对全局 Kmeans 算法[9]进行优化。本文算法与谱聚类、Kmeans、PSO 和全局 Kmeans 等常见算法相比具有更好的稳定性和更高的检测率，与全局 Kmeans 算法相比具有更优的时间复杂度。

2 一种新的聚类算法

2.1 相关背景

2.1.1 谱聚类算法

谱聚类是建立在谱图理论基础上的聚类算法，它是一种通过求解数据集的相似性矩阵来进行的聚类，所以常常被用于非测度空间。谱聚类算法只与数据点的个数有关而与数据点属性的个数无关，所以能够避免由特征向量引发的奇异性问题。谱聚类算法可以解决两类问题，并且通过递归地使用 2-way 方法可以处理多类聚类的划分。已经证明，同时使用多个特征向量就能够避免因算法不稳定而引起的重要特征信息遗漏。为了解决多类聚类问题，一种思想是构造 k-way 图划分标准；另外一种利用的是一种嵌入方法，其思想类似于 PCA 子空间划分方法，代表性算法有 NJW 算法，它本质上是利用相似矩阵的特征向量进行聚类，选取构造矩阵的前 k 个最大特征值所对应的特征向量，从而在 k 维空间中构成与原数据一一对应的表述，进而在 k 维空间中利用 Kmeans 或其他简单算法进行聚类。

NJW 算法的步骤[8,10]如下：

输入：数据集，聚类数目为 K

输出：聚类的划分

1. 构造相似矩阵 $A\in R^{n*n}$，其中 σ 是参数；
2. 构造矩阵 $L=D^{-1/2}AD^{-1/2}$，其中 D 是对角矩阵，对角元素为 $D_{ij}=\sum_{j=1}^{n}A_{ij}$；
3. 计算矩阵 L 的前 K 个特征值所对应的特征向量 $X_1,X_2,\cdots,X_k$，构造矩阵 $X=[X_1,\cdots,X_n]\in R^{n*k}$；
4. 规范化矩阵 X 的行向量，得到矩阵 $Y=X_{ij}/(\sum_j X_{ij}^2)^{1/2}$；
5. 把 Y 的每行当作 K 维空间中的点，采用 Kmeans 等聚类算法把 K 维子空间的数据聚成 K 类，即得到 K 个聚类中心点 $(m_1,\cdots,m_k)$，其中每一行的分类结果对应着原始数据的划分。

2.1.2 粒子群算法

粒子群算法[6,7](Particle Swarm Optimization, PSO)是由美国的 Kennedy 和 Eberhar 在 1995 年提出的。在 PSO 中，通过不断地调节自己的位置，粒子可以获取自己能够搜索到的最优解(记作 P_{id})和整个粒子群搜索到的最优位置(记作 P_{gd})。其中每个粒子都有自己的速度，记作 V：

$$V'_{id}=\omega V_{id}+C_1\,\mathrm{rand}()(P_{id}-X_{id})+C_2\,\mathrm{rand}()(P_{gd}-X_{id}) \quad (1)$$

其中，V_{id} 表示第 i 个粒子在 d 维上的速度，ω 为权重，C_1，C_2 为调节 P_{id} 和 P_{gd} 的参数，rand() 为随机数生成函数。通过式(2)的计算，能够获得粒子下一位置：

$$X'_{id}=X_{id}+V_{id} \quad (2)$$

从式(1)和式(2)可以看出粒子的移动方向包含 3 部分：粒子自己的速度 V_{id}、与粒子本身最佳经历的距离 $(P_{id}-X_{id})$ 和与群体最佳经历的距离 $(P_{id}-X_{id})$、权重 ω。

PSO 算法的步骤[6,7]如下：

1)初始化粒子群中粒子的位置 X 和速度 V；

2)根据适应度函数来计算每个粒子的适应度；

3)将每个粒子计算后的适应度值与自己保持的 P_{id} 的适应度值进行比较，如果更好，则更新 P_{id}；

4)计算群体的适应度值并与 P_{gd} 比较，如果更好，则更新 P_{gd}；

5)根据式(1)和式(2)计算粒子的新的速度和位置；

6)如果达到结束条件(足够好的位置或最大迭代次数)，则结束，否则转步骤 2)。

PSO 收敛速度快而且不会陷入局部最优解，是一种全局优化算法。PSO 算法规则简单，所以容易被编写实现。

2.1.3 全局 Kmeans 算法

针对 Kmeans 算法对初始值敏感的缺点，本文引入全局 Kmeans 算法对全局的优化的特性来克服初始值敏感问题。全局 Kmeans 算法是将 k 个簇的聚类问题转化为一系列的子聚类问题，得到 k 簇聚类问题的最佳解答。具体做法是：从 $k=1$ 开始，实现一个簇的聚类问题，得到一个簇的最佳质心。在此基础上，寻找 $k=2$ 的最佳聚类结果，方式是通过设定每一个样本作为第二个簇的初始中心并运行 Kmeans 聚类算法，以聚类效果最佳(即聚类误差平方和最小)的作为两个簇聚类的最优结果，从而得到两类聚类问题的最佳初始中心和最佳质心；然后再解决 3 个簇的聚类问题，直到实现 k 个簇的聚类问题。

2.2 算法描述

NGKCA 分为 4 个阶段，首先对数据集运行谱聚类 NJW 算法获取重要特征(相当于列降维，并归一化)，再利用 PSO 对数据集进行行降维生成临时聚类中心(相当于行降维)，然后将所有这些聚类中心作为新数据集的样本，引入全局 Kmeans 算法计算新的数据集的聚类中心点，最后利用 PSO 对聚类中心点进行调整以获取最终的聚类中心。

NGKCA 算法的步骤如下：

输入：数据集 $X=\{x_1,\cdots,x_n\}$，$x_n\in R^{n*n}$，聚类数目为 K

输出：聚类的划分

Step1 数据集预处理：

1)构造相似矩阵 $A\in R^{n*n}$，其中 $A_{ij}=\exp(-\frac{|x_i-x_j|^2}{2\sigma^2})$，$i\neq j$，$A_{ij}=0$。其中 σ 是参数。

2)构造矩阵 $L=D^{-1/2}AD^{-1/2}$，$D_{ij}=\sum_{j=1}^{n}A_{ij}$。

3)计算矩阵 L 的前 k 个特征值所对应的特征向量 $X_1,X_2,\cdots,X_k$，构造矩阵 $X=[X_1,\cdots,X_n]\in R^{n*k}$。规范化矩阵 X 的行向量，得到矩阵 $Y=X_{ij}/(\sum_j X_{ij}^2)^{1/2}$，$Y=[Y_1,\cdots,Y_n]\in R^{n*k}$。

Step2 生成临时聚类中心集：

1)根据 $Y=[Y_1,\cdots,Y_n]\in R^{n*k}$ 初始化粒子群，随机生成 $M=\{m_1^*,\cdots,m_t^*\}$ 聚类中心点，t 为人工设定值。并初始化粒子速度 V。

2)引入适应度函数 fit=1.0/[x * sin(10x * pi)+2.0]，计算每个粒子的适应度值；

3)更新 P_{id} 和 P_{gd} 并不断调整聚类中心的位置，直到找到一个足够好的位置或达到最大迭代次数，生成 $M=\{m_1,\cdots,m_t\}$，t 个聚类中心集。

Step3 获取聚类中心点：

1)以 $M=\{m_1,\cdots,m_t\}$ 为数据集进行全局 Kmeans 聚类，首先通过 Kmeans 算法对 K =1 的数据集进行 Kmeans 聚类，找出最佳聚类中心 m_1^*(1)。

2)通过依次将 M 中的点 m_t 设定为第 C 个聚类中心，即构造 t 组初始聚类中心点$(m_1^*(C),\cdots,m_c^*(C),m_t)m_t\in M$，并对每组依次执行 Kmeans 算法同时计算目标函数 $E(m_1,\cdots,m_M)=\sum_{i=1}^{N}\sum_{k=1}^{M}I\parallel x_i-m_k\parallel^2, x_i\in c_k$。找出使得 E 值最小的那一组再进行一次 Kmeans 计算获取新的聚类中心点$(m_1^*(C+1),\cdots,m_{c+1}^*(C+1))$。其中点$(m_1^*(C),\cdots,m_c^*(C))$是 K=C 时找出的 C 个最佳聚类中心。

3)如果 N =C +1，则停止，输出 $M=\{m_1^*,\cdots,m_k^*\}$；否则 C = C+1，返回 2)。

Step4　调整聚类中心点，得到最终的聚类划分：

1)初始化聚类中心点 $M=\{m_1^*,\cdots,m_k^*\}$，k 个聚类中心点的速度为 V；

2)引入适应度函数 fit=1.0/[x * sin(10x * pi)+2.0]，计算每个粒子的适应度值；

3)更新 P_{id}和 P_{gd}并不断调整聚类中心的位置，直到找到一个足够好的位置或达到最大迭代次数，生成 $M=\{m_1,\cdots,m_t\}$，k 个聚类中心点；输出聚类的划分。

算法流程如图 1 所示。

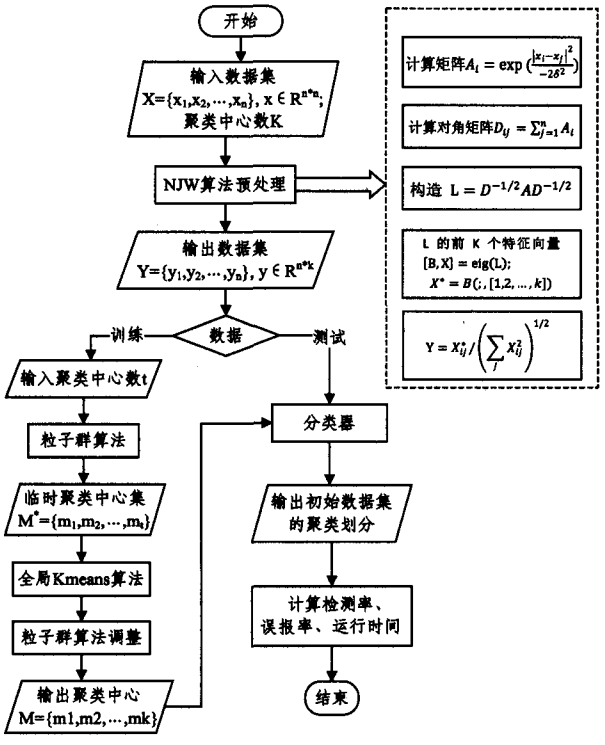

图 1　NGKCA 算法流程

3　实验及结果

实验采用机器学习数据集 UCI 中的 Iris、Ionosphere、Glass、Pendigits 4 个经典数据集和麻省理工 KDDCUP99 网络安全数据集进行测试，所选的数据集的属性和数据点个数具有从少到多的特点。实验是在 Intel 酷睿 i5 双核 CPU、存 4GB、Windows 8 操作系统、Matlab 2008 和 MyEclipse 的环境下进行的。表 1 给出了数据集的具体特征值。

表 1　UCI 数据集的特征

数据集	属性数	数据点总数	类的总数
Iris	4	150	3
Ionosphere	34	351	2
Glass	10	214	7
pendigits	16	7494	9

3.1　UCI 数据集实验

在机器学习数据集 UCI 中，本文实验采用谱聚类、Kmeans、基于 PSO 的 Kmeans 算法(K-PSO)和全局 Kmeans (G-K)算法进行对比实验，分别运行 5 次并将最终的检测率和误报率取平均，得到表 2(检测率)、表 3(误报率)和表 4(平均运行时间)所列结果。

表 2　检测率(%)

算法	Iris	Ionosphere	Glass	pendigits
谱聚类	72	52	67	78
Kmeans	88.7	58.6	72.4	81.8
K-PSO	89.6	63.2	75.2	86.7
G-K	94.2	82.3	84.9	91.4
NGKCA	95.4	86.7	88.2	94.7

表 3　误报率(%)

算法	Iris	Ionosphere	Glass	pendigits
谱聚类	28	52	33	22
Kmeans	11.3	41.4	27.6	18.2
K-PSO	10.4	36.8	24.8	14.3
G-K	5.8	17.7	15.1	8.6
NGKCA	4.6	13.3	11.8	5.3

表 4　平均运行时间(ms)

算法	Iris	Ionosphere	Glass	pendigits
谱聚类	28	46	35	1124
Kmeans	82	94	189	6745
K-PSO	138	143	262	8993
G-K	579	621	1041	74966
NGKCA	267	288	396	11065

通过实验结果可以得知，本文算法比 Kmeans、K-PSO、G-K 算法在检测率上更高并且误报率更低，而且从算法运行时间角度来看，本文算法的运行时间比全局 Kmeans 算法有明显的优势。

观察 UCI 数据集的变化，可以清楚得知 NGKCA 算法的执行过程。表 5 是对 UCI 数据集执行 NJW 算法进行列降维后其属性的变化情况。表 6 是对 UCI 数据集执行 PSO 算法进行行降维后其数据点的变化情况。

表 5　NJW 列降维后的数据集特征

数据集	属性数	数据点总数	类的总数
Iris	3	150	3
Ionosphere	2	351	2
Glass	7	214	7
pendigits	9	7494	9

表 6　PSO 行降维后的数据集特征

数据集	属性数	数据点总数	类的总数
Iris	3	50	3
Ionosphere	2	80	2
Glass	7	60	7
pendigits	9	300	9

通过表 5 得知执行 NJW 算法后属性数较表 1 有所减少，通过表 6 得知执行 PSO 算法后数据点总数有较为明显变化。执行全局 Kmeans 算法后，观察 UCI 中 Iris 数据集在求第二个聚类中心点($K=2$)和第三个聚类中心点($K=3$)处 $E(m_1,\cdots,m_M)=\sum_{i=1}^{N}\sum_{k=1}^{M}I\parallel x_i-m_k\parallel^2$ 值的变化，可以得到图 2 和图 3 所示结果。

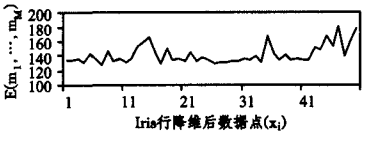

图 2　E 值变化($k=2$)

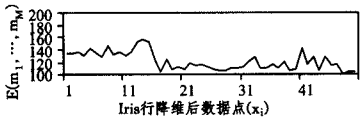

图3 E值变化($k=3$)

3.2 KDDCUP99 数据集实验

KDDCUP99 数据集是国际上标准的网络安全数据集,它是从一个模拟的美国空军局域网上采集来的 9 个星期的网络连接数据,大约有 480 万条记录。其中也提供了约 10%的训练数据集总共 494021 条记录,分成具有标识的训练数据和未加标识的测试数据。如表 7 所列,训练集中包含 1 种正常数据和 22 种攻击类型数据。

表 7 KDDCUP99 入侵检测实验数据

标识类型	含义	具有分类标识
Normal	正常记录	Normal
DoS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	Ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	Buffer_overflow、loadmodule、perl、rootkit

KDDCUP99 数据量大而且维数多,单纯用全局 Kmeans 算法时间开销大。所以本文算法使用 PSO 初步筛选出 500 个聚类中心点,然后再对这 500 个中心点运用全局 Kmeans 算法进行聚类,得到 5 个聚类中心,最后利用 PSO 算法对这 5 个聚类中心进行调整,得到最终的聚类划分。最终的检测率和误报率分别见表 8 和表 9,算法运行时间见表 10。

表 8 检测率(%)

类型	Kmeans	K-PSO	G-K	NGKCA
DoS	89.7	96.3	98.1	98.7
Probe	88.1	92.0	98.4	98.6
R2L	89.5	91.6	94.2	95.9
U2R	90.5	93.2	98.7	98.9
Normal	91.4	98.7	98.2	99.2
All	90.3	94.3	97.5	98.3

表 9 误报率(%)

类型	Kmeans	K-PSO	G-K	NGKCA
DoS	10.3	3.6	1.9	1.3
Probe	11.9	8.0	1.6	1.4
R2L	10.5	8.4	5.8	4.1
U2R	9.5	6.8	1.3	1.1
Normal	8.6	1.3	1.8	0.8
All	9.7	5.7	2.5	1.7

表 10 运行时间(s)

算法	Kmeans	K-PSO	G-K	NGKCA
时间	377.25	516.34	11654.82	1038.57

实验结果表明在 KDDCUP99 网络安全数据集上,NGKCA 算法的检测率优于 Kmeans、谱聚类、K-PSO 和 G-K 算法的。在算法运行时间上,本文算法运行时间为 1038s,与全局 Kmeans 算法的 11654s 相比,有着明显的优势。

结束语 本文采用经典机器学习数据集 UCI 和 KDDCUP99 网络安全数据集进行实验,可以证明 NGKCA 算法在大数据集上有着较高的检测率并且在运行时间上与全局 Kmeans 算法相比要低一个数量级,有效解决了全局 Kmeans 算法时间复杂度高的问题。

参 考 文 献

[1] Li M J, Ng M K, et al. Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(11): 1519-1534

[2] Tou J T, Gonzalez R C. Pattern recognition principle [M]. Addison Wesley, 1974

[3] 姜大庆,夏士雄,周勇. 基于半监督自动谱聚类算法的网络故障检测[J]. 计算机工程与应用,2012,48(30):89-94
Jiang Da-qing, Xia Shi-xiong, Zhou Yong. Network fault detection based on semi-supervised automatic spectral clustering algorithm[J]. Computer Engineering and Applications, 2012, 48(30): 89-94

[4] 周文刚,陈雷霆,董仕. 基于谱聚类的网络流量分类识别算法[J]. 电子测量与仪器学报,2013,27(12):1114-1119
Zhou Wen-gang, Chen Lei-ting, Dong Shi. Network traffic classification algorithm based on spectral clustering[J]. Journal of Electronic Measurement and Instrument, 2013, 27(12): 1114-1119

[5] 刘婧明,韩丽川,侯丽文. 基于粒子群的 K 均值聚类算法[J]. 系统工程理论与实践,2005(6):54-58
Liu Jing-ming, Han Li-chuan, Hou Li-wen. Cluster Analysis Based on Particle Swarm Optimization Algorithm[J]. Systems Engineering—Theory & Practice, 2005, 25(6): 54-58

[6] 张宇,吴昊,陈怀新. 一种新的基于粒子群密度的聚类算法[J]. 电讯技术,2008,48(8):17-21
Zhang Yu, Wu Hao, Chen Huai-xin. A Novel Particle Swarm Optimization Clustering Algorithm Based on Density[J]. Telecommunication Engineering, 2008, 48(8): 17-21

[7] 夏奇,郝顺义,董淼,等. 新的改进 K 均值粒子群算法在组合导航的应用[J]. 计算机应用,2014,34(5):1397-1399,1412
Xia Qi, Hao Shun-yi, Dong Miao, et al. Application of novel K-means particle swarm optimization algorithm in integrated navigation[J]. Journal of Computer Applications, 2014, 34(5): 1397-1399, 1412

[8] 施培蓓,郭玉堂,胡玉娟,等. 初始化独立的谱聚类算法[J]. 计算机工程与应用,2010,46(25):134-137
Shi Pei-bei, Guo Yu-tang, Hu Yu-juan, et al. Initialization independent spectral clustering algorithm[J]. Computer Engineering and Applications, 2010, 46(25): 134-137

[9] 谢鹏,张平伟,罗晟. 基于全局 K-means 的谱聚类算法[J]. 计算机应用,2010,30(7):1936-1937,1940
Xie Huang, Zhang Ping-wei, Luo Sheng. Spectral clustering based on global K-means[J]. Journal of Computer Applications, 2010, 30(7): 1936-1937, 1940

[10] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2001: 856-897