第 55 卷　第 4 期
2019 年 7 月

南京大学学报(自然科学版)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 55, No. 4
July, 2019

DOI:10.13232/j.cnki.jnju.2019.04.004

基于稳定性的三支聚类

杨　鑫[1],施　虹[1],王平心[2*],徐　刚[3]

(1.江苏科技大学计算机学院,镇江,212003;2.江苏科技大学理学院,镇江,212003;
3.江苏科技大学船舶与海洋工程学院,镇江,212003)

摘　要:二支聚类要求聚类结果必须具有清晰的边界,即每个对象要么属于一个类,要么不属于一个类. 然而在许多实际问题中,一个对象和类别可能会有三种关系:即确定属于、确定不属于和无法确定. 为了克服二支聚类的这一问题,三支聚类使用核心域,边界域和琐碎域来表示每个类别,较好地处理了具有不确定性对象的聚类问题. 给出一种基于样本稳定性的三支聚类算法. 首先使用聚类集成的结果计算出每个数据的稳定性,然后基于阈值将这些数据元素分为两部分:核与环. 对核中的数据采用硬聚类进行聚类,对环中的数据通过比较环中数据到聚类中心的距离将它们分到相应类的边界域中. 通过以上策略,可以得到三支聚类的核心域和边界域. 在UCI数据集上的实验结果显示,该方法能更好地显示出聚类的结构.

关键词:聚类集成,稳定性,二支聚类,三支聚类

中图分类号:TP391　　**文献标识码:**A

Three-way clustering based on sample´s stability

Yang Xin[1], Shi Hong[1], Wang Pingxin[2*], Xu Gang[3]

(1. School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, 212003, China;
2. School of Science, Jiangsu University of Science and Technology, Zhenjiang, 212003, China;
3. School of Naval Architecture and Ocean Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China)

Abstract: Two-way clustering algorithms produce clusters with clear and sharp boundaries, which does not truly reflect the fact that a cluster may not necessarily have a well-defined boundary in many real world situations. To tackle this deficiency, three-way clustering uses three regions through a pair of sets to represent a cluster instead of using two regions to represent a cluster by a single set, which reflects the three types of relationship between an object and a cluster, namely, belong-to definitely, uncertain and not belong-to definitely. In this paper, we propose a three-way clustering algorithm by using the stability of each sample. We use clustering ensemble results to compute the sample's stability and divide the universe into cluster core and cluster halo based on sample's stability. The elements in the cluster core are assigned into the core region of each cluster by using traditional clustering algorithm. The elements in the cluster halo are assigned into the fringe region of corresponding cluster according to distances between the elements and

基金项目:国家自然科学基金(61503160,61572242),江苏省高校自然科学研究重大项目(18KJA1300),江苏省高校自然科学研究项目(15KJB110004)

收稿日期:2019－05－22

*通讯联系人,E-mail:wangpingxin@just.edu.cn

the centers of the cluster core region. Therefore, a three-way clustering is naturally formed. Experimental results on UCI datasets show that this method can improve the structure of the clustering results.

Key words: clustering ensemble, stability, two-way clustering, three-way clustering

聚类是对一个数据对象的集合进行分析,它将数据集合分为多个簇,使簇内对象之间有较高的相似性,而不同簇中的对象有较大的差异.聚类分析是一种无监督的学习方法,事先不知道样本的标签,而是利用一些聚类算法将样本进行分类.经过多年发展,聚类已在机器学习、模式识别和数据挖掘中得到广泛应用.

由于数据集的不同,没有一个单一的聚类算法可以准确揭示数据内部的关系与结构,而集成聚类正是为了解决这一问题而被提了出来.集成聚类通过不同的聚类算法或者聚类算法参数的设置对同一个数据集进行集成,建立矩阵,然后通过层次聚类得到最终的结果.

传统的聚类方法都是一种二支决策,如果获取的信息不充分,直接运用传统的聚类算法可能会带来较高的决策风险.为了解决传统聚类算法存在的问题,许多新方法被提出.Hoppner et al[1]提出模糊聚类.Yao et al[2]用区间集来表示聚类结果中的一个类.Yu et al[3-4]提出三支决策方法,将类用核心域、边界域和琐碎域来表示.

所谓三支决策就是将一个研究对象分为三部分看待,即正域、负域和边界域.而三支聚类是在硬聚类的基础上发展而来,它采用了三支决策的思想,将研究对象分为核心域、边界域和琐碎域来表示.即对于一个数据集来说,核心域的点确定属于这个类,琐碎域的点确定不属于这个类,而边界域的点表示这个点可能属于这个类但也有可能属于其他类.

2019年,Li et al[5]提出了基于稳定性的集成算法.本文利用其中一种基于稳定性的方法将硬聚类转化为三支聚类,即利用稳定性把数据分为核与环,对核内数据进行传统的硬聚类,再对环中数据做三支聚类,从而进一步提高聚类质量,降低决策风险.

1　相关工作

1.1　三支决策聚类　2010年,Regina大学的姚一豫教授在研究粗糙集三个域和统计学中的假设验证基础上提出了三支决策理论[6-8],这个理论更精确地反映了粗糙集的近似原理,并可以用来解释实际应用中很多决策现象.三支决策将研究对象分为正域、负域和边界域.正域所对应的规则简称正规则,表示接收;负域对应的规则简称负规则,表示拒绝;边界域对应的规则简称边界规则,对应不做决定或者推迟决定.

现在,三支决策理论的发展越来越快,并在许多领域得到了应用.例如:Yu et al[9-11]提出了三支决策的框架,即用核心域和边界域来表示一个类.Zhang et al[12]提出了分类误差的三支决策模型.Li et al[13]提出了面向多粒度的三支认知概念学习.Hao et al[14]提出了基于序列三支决策的动态多尺度决策表的最优尺度选择.正是这些努力和研究,三支决策理论的内容越来越丰富.

李金海和邓硕[15]给出了三支决策的描述如下:设U是一个有限、非空实体集,其中A是有限条件集.基于有限条件集,三支决策主要的任务是将U划分成三个两两互不相交的域,这三个域分别称之为POS(正域)、NEG(负域)、BND(边界域).依据这三个域可以给出三支决策的规则:接受、拒绝以及不承诺规则.

传统的聚类大多是硬聚类,然而在许多实际问题中,一个对象和类别可能会有三种关系:即确定属于、确定不属于和无法确定.如果把无法确定的点强制划分到某类中可能会带来决策风险,这样的做法不十分合理.于是Yu et al[16]将三支决策思想引入到聚类中,提出了三支决策聚类方法.三支决策聚类用三个集合

C_i^P, C_i^B, C_i^N,分别表示类的核心域、边界域和琐碎域.核心域的点表示这些点确定属于这个类,边界域的点表示这些点可能属于这个类,而琐碎域的点表示这些点不属于这个类.

本文使用 C_i^d 和 C_i^u 分别表示类 i 的核心域与边界域.根据聚类结果的定义,C_i^d 和 C_i^u 须满足以下三个条件:

(1)$C_i^p \neq \varnothing, i=1,2,\cdots,k$;

(2)$\bigcup_{i=1}^{k}(C_i^p \bigcup C_i^B)=U$;

(3)$C_i^p \bigcup C_i^B \bigcup C_i^N = U$.

其中,条件(1)表示任意类簇都是非空的,条件(2)表示样本 $x_i \in U$ 至少属于一个类簇,条件(3)表示表示任意一个类簇的三个区域之并为 U.

1.2 稳定性 2002年 Strehl and Ghosh[17]提出聚类集成(Clustering Ensemble)的概念,给出聚类集成的定义:将两个或多个对同一组对象的数据划分得到的不同结果进行合并,而不使用对象原有的特征.现在对聚类集成问题的研究主要包括集成生成、集成选择和整体集成三个方面.

对于集成方法,可以通过不同的参数设置、不同的聚类方法、特征的不同表示以及弱的聚类等方式进行集成.通过对集成的结果构建矩阵,分析差异,寻找合适的算法对集成结果进行分析,最终得到较好的聚类结果.本文利用 k-means算法[18]来聚类集成.

Li et al[5]提出基于稳定性的集成算法,其主要思想如下:

1.2.1 关系矩阵 首先需要聚类集成来构建关系矩阵.假定 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 表示数据有 n 个样本.经过不同的聚类方法或聚类算法参数的设置,得到一组聚类结果 $\prod=\{C_1,C_2,\cdots,C_L\}$.然后以此聚类结果构建关系矩阵,其中任意两点的关系计算如下:

$$p_{ij}=\frac{1}{L}\sum_{l=1}^{L}\prod(C_l(x_i),C_l(x_j)) \tag{1}$$

L 表示不同的聚类结果,x_i 和 x_j 表示样本中的两个点,$C_l(x_i)$ 表示第 l 个聚类结果中的点 x_i 所在的簇编号.其中:

$$\prod(C_l(x_i),C_l(x_j))=\begin{cases}1 & C_l(x_i)=C_l(x_j)\\ 0 & C_l(x_i)\neq C_l(x_j)\end{cases}$$

此时,关系矩阵就可以通过式(1)求得.

1.2.2 稳定性求法 采用一种线性的方法来求稳定性.首先定义关于变量 p,t 的函数 f,其中 $p\in[0,1],t\in[0,1]$,定义如下:

(1)如果 $p<t$,$f'(p)<0$;如果 $p>t$,$f'(p)>0$.

(2)如果 $p_i<t<p_j$ 且 $\frac{t-p_i}{p_j-t}=\frac{t}{1-t}$,则 $f(p_i)=f(p_j)$.

其中,(1)表示当 $p<t$ 时,函数 f 的导数小于零,函数单调递减;当 $p>t$ 时,函数的导数大于零,函数 f 单调递增.(2)则表示存在 t,当 $p_i<t<p_j$ 且 $\frac{t-p_i}{p_j-t}=\frac{t}{1-t}$ 时,函数 $f(p_i)=f(p_j)$.

假定一个数据集含有 n 个样本,基于这个函数 f,对于每一个点 x_i,定义稳定性 $s(x_i)$ 如下:

$$s(x_i)=\frac{1}{n}\sum_{j=1}^{n}f(p_{ij}) \tag{2}$$

根据函数 f 的定义,一个线性的方法可以定义如下:

$$fl(p_{ij})=\begin{cases}\left|(p_{ij}-t)/t\right| & p_{ij}<t\\ \left|(p_{ij}-t)/(1-t)\right| & p_{ij}\geqslant t\end{cases} \tag{3}$$

这里,针对线性函数 f,求每个点的稳定性:

$$sl(x_i)=\frac{1}{n}\sum_{j=1}^{n}fl(p_{ij}) \tag{4}$$

在这里,采用 Otsu 算法[19]来求阈值 t. Otsu 算法的大致思想如下:

集合 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 含有 n 个元素.存在阈值 t 使得集合 X 被分为两部分,即:

$$X_1=\{x_i:x_i<t,1\leqslant i\leqslant n\}$$
$$X_2=\{x_j:x_j\geqslant t,1\leqslant j\leqslant n\}$$

此时需要学习阈值 t,根据 X_1 和 X_2 定义类间方差为:

$$\beta_t=\omega_0(\mu_0-\mu)^2+\omega_1(\mu_1-\mu)^2 \tag{5}$$

其中,

$$\omega_0=\frac{|X_1|}{|X|},\omega_1=\frac{|X_2|}{|X|}$$

$$\mu_0=\frac{\sum_{X_i\in X_1}x_i}{|X_1|},\mu_1=\frac{\sum_{X_i\in X_2}x_i}{|X_2|},\mu=\frac{\sum_{y_i\in X}x_i}{|X|}$$

随后,根据式(5)求得集合β_t的最大值,这样就得到阈值t:

$$t=\text{argmax}(\beta_t)$$

如此,就能求得每个点的稳定性.

1.2.3　核与环　接下来对于样本$X=\{x_1,x_2,x_3,\cdots,x_n\}$,通过式(4)求得每个点的稳定性$S^M=\{s_1^M,s_2^M,\cdots,s_n^M\}$,然后对这些点再次利用Otsu算法求得集合$S^M$的阈值$t_S$,通过$t_S$可以将集合$S^M$分为两部分,即核与环:

$$O=\{i|s_i^M>t_s,i=1,2,\cdots,n\} \tag{6}$$

$$H=\{i|s_i^M\leqslant t_s,i=1,2,\cdots,n\} \tag{7}$$

其中,集合O代表被分到核中的数据,即比较稳定的数据;H代表被分到环中的数据,即不稳定的数据.

寻找核与环的算法步骤如算法1所示.

算法1　寻找核与环

Step1. 给定一组样本数据集

$$S=\{s_1,s_2,\ldots,s_i,\ldots,s_n\}$$

其中$s_i\in R^l(i=1,2,\ldots,n)$;

Step2. 使用聚类集成求得关系矩阵$\boldsymbol{W}$;

Step3. for $i=1,2,3,\ldots,n$ do

利用式(4)和$\boldsymbol{W}$求得每个点的稳定性s_i^M

end for

所有点的稳定性集合$S^M=\{s_1^M,s_2^M,\cdots,s_n^M\}$;

Step4. 利用Otsu算法应用到S^M求得阈值t_S;

Step5. 最终利用式(6)和(7)求得核与环.

2　基于稳定性的三支聚类

三支聚类的关键问题在于如何计算核心域和边界域,本节给出了一种求核心域和边界域的算法.即基于稳定性的三支聚类算法.

基于稳定性的三支聚类算法的主要思想是:给定数据集$X=\{x_1,x_2,x_3,\cdots,x_n\}$,先利用聚类集成求出关系矩阵,这里的聚类集成通过$k$-means每次返回的结果进行集成.使用Otsu算法求出关于这个关系矩阵的阈值t,然后根据定义的线性函数(式(4))计算出每个点的稳定性$S^M=\{s_1^M,s_2^M,\cdots,s_n^M\}$,再对集合$S^M$中的数据点再次使用Otsu算法得出阈值$t_s$.比较集合$S^M$中的每一个数据,如果$s_i^M$比阈值$t_s$大,则把此点划分到核中,反之将它划分到环中,这样就求得核与环.随后对核中数据进行传统硬聚类k-means得到聚类结果$C_i,i=1,2,\cdots,k$.而对于环中数据,采用遍历的形式,依次计算环中的每个数据到聚类中心的距离d,先找出距离最小的值$d_{\min}$,将此距离最小的所对应的数据点划分为此类的上界,然后计算此点到其他聚类中心的距离与$d_{\min}$的差值d_{poor},如果这个距离d_{poor}小于指定的阈值p,则把此数据点划分为该类的上界,直至环中数据全部遍历完成.最终得到三支聚类结果.算法步骤如算法2所示.

算法2　基于稳定性的三支聚类

输入:由算法1得到的稳定性数据O,不稳定数据H和关系矩阵$\boldsymbol{W}$,聚类数目k,阈值p

输出:聚类结果

Step1. 对稳定性的数据进行k-means聚类得聚类结果$C_i,i=1,2,\ldots,k$.

Step2. 取不稳定的数据H,进行遍历.

for $i=1,2,3,\ldots,|H|$ do

计算不稳定点H_i到每一个聚类C的聚类中心的距离$d=\{d_1,d_2,\cdots,d_k\}$

找出集合d中的最小值$d_{\min}=\min(d)$,将$d_{\min}$对应的数据H_i划分到其对应类C的上界.

接着计算集合d中其余点与$d_{\min}$的差值d_{poor}

if $d_{\text{poor}}<p$

将样本H_i也添加到对应类C的上界

end if

end for

Step3. 最终得到三支聚类结果.

3　聚类结果评价指标

聚类的评价指标大致分为两类:外部聚类和内部聚类.外部聚类评价指标包括Entropy,F-measure,Purity,Rand Statistic等.内部聚类

评价指标包括轮廓系数(Silhouette Coefficient, S_i),DB_Index(Davies-Bouldin Index, DBI),Calinski-Harabasz(CH)指标,Krzanowski-Lai(KL)指标等.本文所用的评价指标为准确率(Accuracy, ACC),DBI, S_i和平均轮廓系数(Average Silhouette Coefficient, AS).

3.1 准确率 ACC是一种常见的评价聚类结果好坏的外部指标,根据预测的结果与真实值做对比,此值越高说明聚类结果越好.

定义1 ***ACC***[20]

$$ACC=\frac{1}{N}\sum_{i=1}^{k}C_i$$

其中,N表示总样本个数,C_i表示正确划分到类i的样本个数,k表示聚类数.本论文的三支聚类算法实验所计算的ACC是使用核心域的对象来计算的.

3.2 Davies-Bouldin Index DBI是Davies and Bouldin[21]于1979年提出的一种内部聚类评价指标,其主要思想是度量每个簇类最大相似度的均值.

定义2 ***DBI***[21]

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\bar{c}_i+\bar{c}_j}{\|w_i-w_j\|_2}\right)$$

其中,$\bar{c}_i$表示第i类中所有样本到聚类中心w_i的平均距离,$\|w_i-w_j\|_2$表示类i与类j聚类中心之间的欧式距离,k表示聚类数.

3.3 平均轮廓系数 S_i是一种评价聚类结果好坏的指标,最早由Rousseeuw[22]在1986年提出.它结合内聚度和分离度两种因素,可以用来在相同原始数据的基础上评价不同算法、或者算法不同运行方式对聚类结果所产生的影响.

定义3 单个样本$\boldsymbol{d}_i$的轮廓系数$\boldsymbol{S}_i$[22]

$$S_i=\frac{b_i-a_i}{\max(a_i,b_i)}$$

其中,a_i表示样本d_i与同类簇中其他所有样本的平均距离,称为类内相似度,a_i越大说明该样本属于该类簇的可能性越大.$b_i=\min\{D(d_i-c_j)\}$,表示样本d_i到类c_j中所有样本的最小平均距离,称为类间相异度,b_i越大说明该样本属于其他类簇的可能性越小.

3.4 平均轮廓系数

定义4 ***AS***[22]

$$AS=\frac{1}{N}\sum_{i=1}^{N}S_i$$

其中,N表示样本总数,S_i表示第i个样本的轮廓系数.平均轮廓系数是用所有样本的轮廓系数的均值表示,取值范围[-1,1],值越大表示样本属于该类簇的可能性越大,反之可能性就越小.

4 实验结果

UCI数据集的纯度高,噪音数据较少,因而被广泛认可.本文采用五组UCI数据集对算法进行验证,具体信息如表1所示.本文将基于稳定性的三支决策聚类与传统的聚类k-means进行ACC,DBI和AS等聚类指标的对比,得出了基于稳定性的三支决策聚类可以提高聚类精度、改善聚类性能的结论.

表1 实验中使用的数据集

Table 1 Datasets used in experiments

Datasets	Sample numbers	Sample dimensions	Categories
Bank	1372	4	2
Glass	214	9	6
Wine	178	13	3
Congressional	435	16	2
Breast	106	9	6

本实验先对每组数据进行100次聚类集成,最后取得ACC,AS,DBI的值作为实验结果,实验结果如表2所示.

从表2的实验结果可以看出,与k-means算法比较,本文提出的基于稳定性的三支聚类算法可以提高ACC和AS,并且可以降低DBI,使得聚类结果更好,质量更高.但是此算法因为先开始使用了聚类集成,导致算法的开销增大,效率有所降低,这是一个待解决的问题.

表2　UCI数据集上的实验结果

Table 2　Experimental results on UCI datasets

Datasets	Algorithm	*DBI*	*AS*	*ACC*
Bank	*k*-means	1.1913	0.5000	0.5758
	Three-*k*-means	1.1772	0.5079	0.5751
Glass	*k*-means	0.9625	0.5325	0.5981
	Three-*k*-means	0.9252	0.6129	0.6774
Wine	*k*-means	1.3053	0.4763	0.9550
	Three-*k*-means	1.2430	0.5121	0.9704
Congressional	*k*-means	1.4865	0.4407	0.8666
	Three-*k*-means	1.3889	0.4723	0.8812
Breast	*k*-means	0.8826	0.5644	0.7735
	Three-*k*-means	0.7288	0.6817	0.7945

5　结束语

本文利用样本的稳定性给出了一种基于稳定性的三支聚类算法.该算法首先通过聚类集成结果定义每个元素的稳定性,然后利用元素的稳定性将元素分为核心集合与边界集合.对核心集合中的元素采用硬聚类的方法聚类,而对边界集合中的元素,利用它们和核心集合的距离将它们分到相应的类别边界域中.实验也表明此方法可以提高聚类的精度.目前算法的不足之处在于:聚类集成的时候单一用*k*-means不是很好,可以尝试多种不同的聚类方法.另外,对于利用集成方法求样本的稳定性方面,尝试不同的集成算法,并且改进稳定点的求法使得此算法可以适应更多的数据.

参考文献

[1] Hoppner F, Klawonn F, Kruse R, et al. Fuzzy cluster analysis: methods for classification, data analysis and image recognition. New York: Wiley, 1999, 770.

[2] Yao Y Y, Lingras P, Wang R Z, et al. Interval set cluster analysis: A re-formulation// Sakai H, Chakraborty M K, Hassanien A E, et al. Rough sets, fuzzy sets, data mining and granular computing. Springer Berlin Heidelberg, 2009: 398－405.

[3] Yu H, Chu S S, Yang D C. Autonomous knowledge-oriented clustering using decision-theoretic rough set theory. Fundamenta Informaticae, 2012, 115: 141－156.

[4] Yu H, Liu Z G, Wang G Y. An automatic method to determine the number of clusters using decision-theoretic rough set. International Journal of Approximate Reasoning, 2014, 55(1): 101－115.

[5] Li F J, Qian Y H, Wang J T, et al. Clustering ensemble based on sample's stability. Artificial Intelligence, 2019, 273: 37－55.

[6] Yao Y Y. Three-way decisions with probabilistic rough sets. Information Sciences, 2010, 180(3): 341－353.

[7] Yao Y Y. The superiority of three-way decisions in probabilistic rough set models. Information Sciences, 2011, 181(6): 1086－1096.

[8] Yao Y Y. An outline of a theory of three-way decisions// Yao J. Rough sets and current trends in computing. Springer Berlin Heidelberg, 2012: 1－17.

[9] Yu H. A framework of three-way cluster analysis// Proceedings of International Joint Conference on Rough Sets. Springer Berlin Heidelberg, 2017: 300－312.

[10] Yu H, Jiao P, Yao Y Y, et al. Detecting and refining overlapping regions in complex networks

with three-way decisions. Information Sciences, 2016,373:21－41.

[11] Yu H, Zhang C, Wang G Y. A tree-based incremental overlapping clustering method using the three-way decision theory. Knowledge-Based Systems,2016,91:189－203.

[12] Zhang Q H, Xia D Y, Wang G Y. Three-way decision model with two types of classification errors. Information Sciences,2017,420:431－453.

[13] Li J H, Huang C C, Qi J J, et al. Three-way cognitive concept learning via multi-granularity. Information Sciences,2017,378:244－263.

[14] Hao C, Li J H, Fan M, et al. Optimal scale selection in dynamic multi-scale decision tables based on sequential three-way decisions. Information Sciences,2017,415－416:213－232.

[15] 李金海,邓硕. 概念格与三支决策及其研究展望. 西北大学学报(自然科学版),2017,47(3):321－329. (Li J H, Deng S. Concept lattice, three-way decisions and their research outlooks. Journal of Northwest University (Natural Science Edition), 2017,47(3):321－329.)

[16] Yu H, Chu S S, Yang D C. Autonomous knowledge-oriented clustering using decision-theoretic rough set theory// Yu J, Greco S, Lingras P, et al. Rough set and knowledge technology. Springer Berlin Heidelberg,2010:687－694.

[17] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research,2002,3:583－617.

[18] MacQueen J. Some methods for classification and analysis of multivariate observations// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA, USA: University of California Press,1967:281－297.

[19] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics,1979,9(1):62－66.

[20] Schölkopf B, Platt J, Hofmann T. A local learning approach for Clustering// International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2007: 1529－1536.

[21] 孙吉贵,刘杰,赵连宇. 聚类算法研究. 软件学报. 2008,19(1):48－61. (Sun J G, Liu J, Zhao L Y. Clustering algorithms research. Journal of Software,2008,19(1):48－61.)

[22] Fahad A, Alshatri N, Tari Z, et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing, 2014, 2(3): 267－279.

(责任编辑 杨可盛)