ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 4,Jul. 2020,pp. 613－629
DOI:10. 16337/j. 1004-9037. 2020. 04. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种面向不完备信息系统的集对k-means聚类算法

张春英[1,3]，高瑞艳[1]，刘凤春[2]，王佳昊[1]，陈　松[1]，冯晓泽[1]，任　静[1]

(1. 华北理工大学理学院,唐山，063210;2. 华北理工大学迁安学院,唐山，063210;3. 河北省数据科学与应用重点实验室,唐山,063210)

摘　要：针对不完备信息系统的数据聚类问题,将集对分析理论引入k-means聚类中,同时为了更好地表示样本与类簇的关系,构建了一种面向不完备信息系统的集对k-means (Set pair k-means,SPKM)聚类算法。首先,基于集对理论提出了一种集对距离度量方法,并将该度量方法运用到k-means算法中,得到初步聚类结果;随后,对于同时属于多个类的样本,将其分配到相应类的边界域,对于只属于一个类的样本,将其分配到相应类的正同域或边界域,其中聚类结果由肯定属于该类簇的正同域、可能属于该类簇的边界域以及肯定不属于该类簇的负反域3个部分共同表示;最后通过选取UCI数据库中的6个数据集与4种对比算法进行实验评价。实验结果表明,SPKM算法在准确率、F_1值、Jaccard系数、FMI和ARI等指标上均具有良好的聚类性能。

关键词：集对信息粒;不完备信息;k-means;集对距离度量;集对k-means聚类

中图分类号：TP391　　**文献标志码**:A

A Set Pair k-means Clustering Algorithm for Incomplete Information System

ZHANG Chunying[1,3], GAO Ruiyan[1], LIU Fengchun[2], WANG Jiahao[1], CHEN Song[1], FENG Xiaoze[1], REN Jing[1]

(1. College of Science, North China University of Science and Technology, Tangshan, 063210, China; 2. Qian'an College, North China University of Science and Technology, Tangshan, 063210, China; 3. Key Laboratory of Data Science and Application of Hebei Province, Tangshan, 063210, China)

Abstract: For the data clustering problem of incomplete information system, the set pair analysis theory is introduced into k-means clustering. At the same time, to better represent the relationship between the sample and the cluster, a set pair k-means(SPKM) clustering algorithm for incomplete information system is constructed. Firstly, a set pair distance measurement method is proposed according to set pair theory, and the measurement method is applied to the k-means algorithm to obtain the preliminary clustering results. Then, for samples belonging to multiple clusters at the same time, the samples are assigned into the boundary region of the corresponding clusters. And for samples belonging to only one cluster, it is assigned into the positive region or boundary region of the corresponding clusters. The clustering results are expressed by three parts, which are the positive region belonging to the cluster, the boundary region that may belong to the cluster and the negative region which does not belong to the cluster. Finally, six data sets in the UCI database and four contrast algorithms are selected for experimental evaluation.

基金项目:河北省自然科学基金(F2018209374,F2016209344)资助项目。
收稿日期:2020-04-30;**修订日期**:2020-07-10

Experimental results show that the SPKM algorithm has good clustering performance in accuracy, F_1 value, Jaccard coefficient, FMI and ARI.

Key words: set pair information granule; incomplete information; k-means; set pair distance measurement; set pair k-means clustering

引　言

作为一种强大的数据分析工具,聚类在数据挖掘中起着重要作用,广泛应用于异常数据检测[1-2]、生物信息学[3]和网络结构分析[4]等领域。聚类的目的是将相似的样本分组到同一簇中,将不同的样本分组到不同的簇中,从而可以从数据集中找到潜在的相似模式。

目前现有的大多数聚类方法是假设每个样本必须精确地分配给一个类簇,即一个样本只属于一个类簇。然而在实际应用中,一个样本可以同时分配给两个或多个类簇,其在信息不完整或不准确的情况下,很难给出清晰的划分结果。三支决策[5-6]理论认为,人们通常根据现有的信息和证据作决策,然而,如果信息不足或薄弱,则无法做出接受或拒绝的决策,因此,人们可以选择延迟决策来解决这一问题,待获取更多信息后,再给予进一步的决策。于洪[7]将其应用于聚类,并提出了三支聚类的概念,认为一个聚类不再由单一的具有清晰边界的集合表示,而是通过一对集合来表示类簇。随后学者们又对三支聚类的概念进行了进一步的研究。Wang等[8]采用重叠聚类来获得聚类的支持,再利用扰动分析将核心区域从聚类的支持中分离开来,形成对聚类的三支解释。随后,Wang等[9]又基于数学形态学的腐蚀和膨胀,提出了三支聚类(CE3)的总体框架。Yu等[10]提出了一种基于改进的DBSCAN(Density-based spatial clustering of application with noise)的三支聚类,对样本间相似性计算进行了改进,并用一对嵌套集来表示一个类簇。以上聚类算法,虽然考虑了样本间的不确定关系,但主要是针对完备数据集,其对不完备数据集可能并不完全适用。

当聚类算法应用于实际数据集,会出现一个不可避免的问题,就是样本中部分属性值缺失,然而传统的聚类算法不能直接用于不完备的数据集,其只能应用于完备的数据集。为了解决不完备数据聚类问题,国内外学者基于模糊C均值算法(Fuzzy C-means, FCM)算法提出了一系列改进方法。Aydilek等[11]提出了一种支持向量机和遗传算法的混合方法来估计FCM算法中的缺失值并对参数进行优化;Li等[12]为区间数据定义了新的距离函数,并扩展经典FCM以处理缺失数据;Zhang等[13]利用预先分类的聚类结果设计了一种改进的区间构造方法,并通过粒子群优化寻找最优聚类。

目前,对于不完备数据聚类的研究,大多数方法采取了对缺失值进行填充,然而,缺失值本身就具有不确定性,删除或者填充都会造成一定的误差,进而影响聚类效果。为了有效解决不完备数据聚类问题以及更好地表示样本与类簇的关系,本文提出了一种面向不完备信息系统的集对k-means(Set pair k-means, SPKM)聚类算法。SPKM算法的主要贡献体现在以下几方面:(1)对于缺失值的处理给出了相应方法,运用集对信息粒的粒化表达方法,将缺失值对应的粒度记为差异度,使得原本聚类过程中不同样本之间距离扩展成包含正同度、差异度和负反度3个维度的距离定义,可全面地反映聚类效果的正同度、差异度和负反度,比从单一角度衡量更具有系统性。由此根据提出的集对距离度量,获得距离各个样本最近的聚类中心,进而得到初步聚类结果;(2)针对一个样本可能不止和一个类有关系的情况,也给出了相应的聚类方法。对于同时属于多个类的样本,将其分配到相应类的边界域;对于只属于一个类的样本,根据建立的集对联系度公式,将其划分到相应类的正同域或边界域,进而形成由正同域,

边界域和负反域表示的集对聚类结果;(3) 通过6个UCI数据集与其他4种有代表性的算法进行了实验对比分析,结果表明,该算法可以有效处理具有缺失值的不完备数据集,并且得到较好的聚类效果。

1 基本理论

1.1 不完备信息系统

信息系统又称为知识表达系统,是一个四元组$S=(U,A,V,f)$,其中$U=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$是非空有限样本集,称为论域,$n$为论域中数据样本的个数;$A=\{a_1,a_2,\cdots,a_m\}$是非空有限属性集,$m$为属性值的个数;$V=\{V_1,V_2,\cdots,V_m\}$是$U$关于$A$的属性的值域集合,$V_s$是属性$a_s(1\leqslant s\leqslant m)$的值域;$f$是信息函数,$f:v_{is}=f(x_i,a_s)\in V_s$,表示样本$x_i$在属性$a_s$上的取值为$v_{is}$。

x_i是论域中的第i个样本,其具有$A=|m|$个属性值,当存在缺失属性值时,信息系统S是不完备的,本文的研究对象是具有缺失值的不完备信息系统。

1.2 集对分析

集对分析是以同、异、反来描述事物的一种理论,通过建立两个事物之间的集对联系度以期来描述确定-不确定性,集对联系度表达式为

$$\rho=\frac{S}{N}+\frac{F}{N}i+\frac{P}{N}j \tag{1}$$

式中:S代表属性值相同的数目,P代表属性值相反的数目,F代表属性值既不相同又不对立的数目,N表示属性值的总个数。

记$\frac{S}{N}=a,\frac{F}{N}=b,\frac{P}{N}=c$,则式(1)可以表示为

$$\rho=a+bi+cj \tag{2}$$

式中:a表示正同度,b表示差异度,c表示负反度。其中$i\in[-1,1]$和$j=-1$分别称为差异度和负反度标记符号。

定义1[14](确定粒集、不确定粒集、确定度和不确定度)设$W=(U,A,V)$,$W_0=(U_0,A_0,V_0)$,$A_0\subseteq A$,$V_0\subseteq V$,$R\subseteq A_0$,定义W上一对子集,确定粒集$X^C=\{X_1^C,X_2^C,\cdots,X_m^C\}$和不确定粒集(差异粒集)$X^U=\{X_1^U,X_2^U,\cdots,X_h^U\}$,则对于信息$x\in W_0$,存在一对映射

$$\tau X^C:W_0\to[0,1],x\to\tau X^C(x)=a_R+c_R \tag{3}$$

$$\tau X^U:W_0\to[0,1],x\to\tau X^U(x)=b_R \tag{4}$$

式中:a_R+b_R和c_R分别称为x关于X^C,X^U的确定度和不确定度;$X_i^C\in X^C(1\leqslant i\leqslant m)$为确定信息粒;$X_j^U\in X^U(1\leqslant j\leqslant h)$为不确定信息粒(差异信息粒)。

定义2[14](正同粒集、负反粒集、正同度和负反度)基于确定信息粒X^C,$R\subseteq A_0$,定义X^C上一对子集,正同信息粒集$X^{C_s}=\left\{X_1^{C_s},X_2^{C_s},\cdots,X_k^{C_s}\right\}$和负反信息粒集$X^{C_o}=\left\{X_1^{C_o},X_2^{C_o},\cdots,X_l^{C_o}\right\}$,则对于信息$x\in X^C$,存在一对映射

$$\tau X^{C_s}:X^C\to[0,1],x\to\tau X^{C_s}(x)=a_R \tag{5}$$

$$\tau X^{C_o}:X^C\to[0,1],x\to\tau X^{C_o}(x)=c_R \tag{6}$$

式中:a_R和c_R分别称为x关于X^{C_s},X^{C_o}的正同度和负反度;$X_i^{C_s}\in X^{C_s}(1\leqslant i\leqslant k)$为正同信息粒;

$X_j^{C_o} \in X^{C_o}(1 \leqslant j \leqslant l)$为负反信息粒。

1.3 传统聚类结果表示

针对图1给出的数据集，由传统聚类算法得到的结果如图2所示。图2中每个样本点被明确地划分到一个类簇，实际上位于中间部分的样本点x_1, x_2, x_3存在不确定信息，不论是将其划分到哪一类，都不能很好地体现聚类结构，导致得到的聚类结果存在误差。

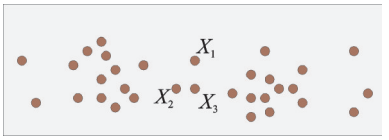


图1 数据集示意图

Fig.1 Schematic diagram of dataset

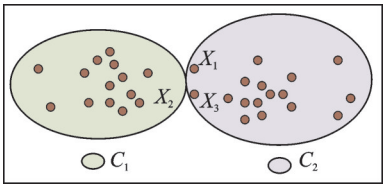


图2 传统聚类结果示意图

Fig.2 Schematic diagram of traditional clustering results

2 基于集对的k-means的聚类算法

2.1 集对距离度量

衡量样本点之间的距离是聚类过程中非常关键的一步，然而由于缺失值的存在，一些传统的距离计算方法不能直接用于计算不完备数据之间的距离。为此，本文基于集对分析的相关理论，提出了集对距离度量方法，将原本聚类过程中不同样本之间的距离扩展成包含正同度、差异度和负反度3个维度的距离定义，能有效地处理含有缺失值的不完备数据集。

定义3（集对联系度）假定样本集为$U=\{x_1, x_2, \cdots, x_n\}$，任意两个样本$x_p, x_q \in U$，每个样本$x_p=\{v_{p1}, v_{p2}, \cdots, v_{pm}\}$由$m$个属性描述，设定阈值为$\varepsilon_1, \varepsilon_2(\varepsilon_2 > \varepsilon_1)$，通过标准化后的数据，将样本$x_p$和样本$x_q$建立集对联系度，令满足$|v_{ps}-v_{qs}| \leqslant \varepsilon_1(1 \leqslant s \leqslant m)$的记为正同信息粒$S$；满足$|v_{ps}-v_{qs}| \geqslant \varepsilon_2$的记为负反信息粒$P$；由于缺失属性值本身就具有不确定性，删除或者填充都会造成一定的误差，然而差异信息粒可以表示模糊、不确定的信息，故将满足$\varepsilon_1 < |v_{ps}-v_{qs}| < \varepsilon_2$以及缺失的属性值均记为差异信息粒$F$，则样本$x_p$和$x_q$之间建立的集对联系度为

$$\rho_{pq}=\frac{S}{N}+\frac{F}{N}i+\frac{P}{N}j \tag{7}$$

式中：N代表属性值的总数目；S代表属性值数据差的绝对值小于等于ε_1的数目；P代表属性值数据差的绝对值大于等于ε_2的数目；F代表其他属性值数目，其包括缺失的属性值。

式(7)也可以简化为

$$\rho_{pq}=a+bi+cj \tag{8}$$

式中：$a=\frac{S}{N}$为正同度，$b=\frac{F}{N}$为差异度，$c=\frac{P}{N}$为负反度，$a+b+c=1$。

定义4（集对联系度矩阵）设待聚类样本集为U，A为属性集，在关系R下，任取U中的样本x_p和$x_q(p \neq q; p, q=1, 2, \cdots, n)$，令$x_p$与$x_q$建立集对联系度，从而获得集对联系度矩阵$T$为

$$T=\begin{bmatrix} a_{11}+b_{11}i+c_{11}j & a_{12}+b_{12}i+c_{12}j & \cdots & a_{1n}+b_{1n}i+c_{1n}j \\ a_{21}+b_{21}i+c_{21}j & a_{22}+b_{22}i+c_{22}j & \cdots & a_{2n}+b_{2n}i+c_{2n}j \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1}+b_{n1}i+c_{n1}j & a_{n2}+b_{n2}i+c_{n2}j & \cdots & a_{nn}+b_{nn}i+c_{nn}j \end{bmatrix} \tag{9}$$

式中：a_{pq}, b_{pq}和$c_{pq}(1 \leqslant p \leqslant n, 1 \leqslant q \leqslant n)$满足$a_{pq}+b_{pq}+c_{pq}=1$，其中$a_{pq}$为正同度，$b_{pq}$为差异度，$c_{pq}$为负反度。

定义5(集对距离度量)任意两个样本$x_p,x_q(1\leqslant p,q\leqslant n)$之间的集对联系度$\rho_{pq}=a+bi+cj$可以通过式(8)得到。根据集对联系度定义了一种集对距离度量方法,可通过正同度a和负反度c的大小来确定与每个样本x_p距离最近的样本,即

$$x_{p.\text{closet}}=x_{\lambda_q}\leftarrow\left\{\lambda_q=\arg\max_{q\neq p,q\in 1,2,\cdots,n}\{a_{pq}\}\right\} \tag{10}$$

$$x_{p.\text{closet}}=x_{\lambda_q}\leftarrow\begin{Bmatrix}\lambda_q=\arg\max\limits_{q\neq p,q\in 1,2,\cdots,n}\{a_{pq}\}\\ \lambda_q=\arg\min\limits_{q\neq p,q\in 1,2,\cdots,n}\{c_{pq}\}\end{Bmatrix} \tag{11}$$

对任意样本x_p,根据式(10)得到与之建立的集对联系度中正同度最大的样本,如果满足条件的只有一个样本,则将其确定为距离样本x_p最近的样本,如果满足条件的不止一个样本,则根据条件更为严格的式(11),选取与之建立的集对联系度同时满足正同度最大、负反度最小的样本,将其确定为距离样本x_p最近的样本。需要注意的是,对于每个样本x_p,找到距离其最近的样本,可能为一个,也可能为多个。

2.2 集对k-means聚类结果表示

本文将集对信息粒的相关理论引入k-means聚类中,基于集对信息粒中的正同粒集,差异粒集和负反粒集的定义,提出用正同域C_s,边界域C_u和负反域C_o三个域来表示聚类结果。其中,正同域表示属于这个类,边界域表示可能属于这个类,负反域表示不属于这个类。聚类结果如图3所示,设定了一个边界线用于更好地显示正同域和边界域。

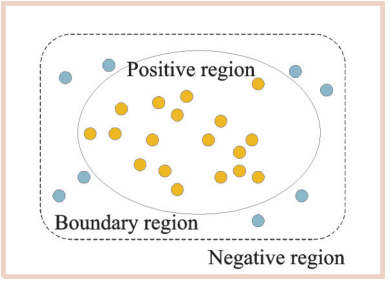


图3 聚类的可视化

Fig.3 Visualization of clustering

聚类的目的是将相似程度高的样本划分到正同域,使其位于类的中心,将相似程度较低的样本划分到边界域,用这两个域可以更好地表示一个类。这3个域满足如下性质

(1)$C_s(C_i)\neq\varnothing$

(2)$(C_s(C_i)\bigcup C_u(C_i))=U$

(3)$C_s(C_i)\bigcap C_s(C_j)=\varnothing,i\neq j$

式中:$C_s(C_i)$表示类簇C_i的正同域,$C_u(C_i)$表示类簇C_i的边界域。性质(1)说明,每个类的正同域不能为空;性质(2)说明U中的任何一个样本必须属于一个正同域或者至少属于一个边界域;性质(3)说明任何一个样本最多只能属于一个类的正同域。

2.3 算法描述与分析

集对k-means聚类算法可以用于处理存在缺失值的不完备数据集。缺失值处理的主要思想是将样本的缺失属性值在进行集对分析时,将其粒度记为相应的差异度。另外,其是用正同域、边界域共同来表示一个聚类,而不是一个单一的集合。然而传统k-means聚类算法是用具有清晰边界的集合来表示一个聚类,其思想是先初始化k个聚类中心,作为k个初始类簇,然后将每个样本依次分配到各个类簇。但在聚类过程中只考虑了两种关系,会降低对不确定点划分的准确性,而样本与类簇之间存在属于、可能属于、不属于这3种关系。针对这3种关系的划分,本文实验中的聚类任务分为两阶段,第1阶段构造包含正同域和边界域的集合,第2阶段使正同域和边界域分离。

第1阶段:假定样本集为$U=\{x_1,x_2,\cdots,x_n\}$,选取的k个初始聚类中心为$\{\mu_1,\mu_2,\cdots,\mu_k\}$。将样本集数据先进行标准化处理,再通过式(8)依次将样本$x_p(1\leqslant p\leqslant n)$与各聚类中心$\mu_j(1\leqslant j\leqslant k)$建立集对联

系度。根据集对距离式(10)和(11)得到距离每个样本最近的聚类中心,将样本x_p分配到与之距离最近的类簇$C_{\lambda_q}=C_{\lambda_q}\cup\{x_p\}$,由此确定每个样本的簇标记$\lambda_q$。在迭代过程中,新聚类中心的计算公式为

$$\mu_j'=\frac{1}{|C_j|}\sum_{x\in C_j}x \tag{12}$$

式中:$x\in C_j$,$x=\{v_{\cdot 1},v_{\cdot 2},\cdots,v_{\cdot m}\}$,$j=1,2,\cdots,k$,$|C_j|$表示类簇$C_j$的元素个数。

上述过程得到了聚类的初步结果,将类中样本分为两种类型,即

$$\{x_i\in C_i\,|\,\exists j=1,2,\cdots,k,j\neq i,x_i\in C_j\}$$

$$\{x_i\in C_i\,|\,\forall j=1,2,\cdots,k,j\neq i,x_i\notin C_j\}$$

第2阶段:对初步聚类结果进行细分,使得正同域和边界域进行分离。第1种类型的样本$\{x_i\in C_i\,|\,\exists j=1,2,\cdots,k,j\neq i,x_i\in C_j\}$,显然,其与多个类簇存在关系,则将其同时分配到类簇C_i的边界域$C_{\mathrm{u}}(C_i)$和类簇C_j的边界域$C_{\mathrm{u}}(C_j)$。对于第2种类型的样本$\{x_i\in C_i\,|\,\forall j=1,2,\cdots,k,j\neq i,x_i\notin C_j\}$,其只属于一个类簇$C_i$中,不再属于其他任何类簇$C_j(j\neq i)$,只需判断其位于正同域还是边界域。计算方法如下:设定正同度阈值为α,负反度阈值为β,计算该样本与所在类的聚类中心的集对联系度,比较其正同度、负反度与阈值之间的大小关系,将类中样本依次分配到相应类簇的正同域C_{s}或边界域C_{u},公式为

$$\begin{cases}C_{\mathrm{s}}=\{a\geqslant\alpha\ \text{or}\ c<\beta\}\\ C_{\mathrm{u}}=\{0<a<\alpha\ \text{and}\ c\geqslant\beta\}\end{cases} \tag{13}$$

基于以上讨论,集对k-means聚类的结果可表示为

$$C=\{C_{\mathrm{s}}(C_1)\cup C_{\mathrm{u}}(C_1)\},\{C_{\mathrm{s}}(C_2)\cup C_{\mathrm{u}}(C_2)\},\cdots,\{C_{\mathrm{s}}(C_k)\cup C_{\mathrm{u}}(C_k)\} \tag{14}$$

式中:$C_j=\{C_{\mathrm{s}}(C_j)\cup C_{\mathrm{u}}(C_j)\}(1\leqslant j\leqslant k)$表示一个类簇的聚类结果。对于图1中的样本点,本文所提SPKM算法得到的聚类结果如图4所示。

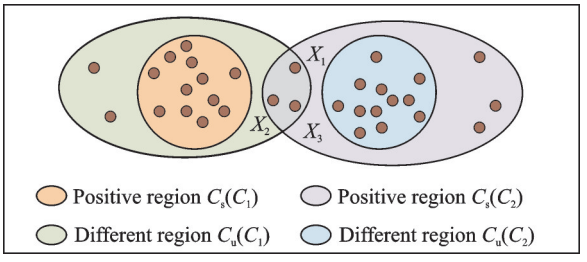


图4 集对k-means聚类示意图

Fig.4 Schematic diagram of set pair k-means clustering

2.4 算法流程

算法步骤如下:

输入:样本集$U=\{x_1,x_2,\cdots,x_n\}$,类簇数目k,参数$\varepsilon_1,\varepsilon_2,\alpha,\beta$

输出:聚类结果$C=\{\{C_{\mathrm{s}}(C_1)\cup C_{\mathrm{u}}(C_1)\},\{C_{\mathrm{s}}(C_2)\cup C_{\mathrm{u}}(C_2)\},\cdots,\{C_{\mathrm{s}}(C_k)\cup C_{\mathrm{u}}(C_k)\}\}$

(1)随机选取k个样本作为初始聚类中心$\{\mu_1,\mu_2,\cdots,\mu_k\}$

(2)Repeat

(3) 令 $C_j=\varnothing$ ($j=1,2,\cdots,k$)

(4) For $p=1,2,\cdots,n$ do

(5) 对于每个样本x_p($1\leqslant p\leqslant n$),根据式(10)和(11)计算得到与之距离最近的聚类中心μ_j

(6) 令λ_j为簇标记,$\lambda_j=\arg\max\limits_{j\in\{1,2,\cdots,k\}}\{a_{pj}\}$,如果满足$\max\{a_{pj}\}$的样本不唯一,则增加条件$\min\{c_{pj}\}$,选择$\lambda_j=\arg\min\limits_{j\in\{1,2,\cdots,k\}}\{c_{pj}\}\bigcap\arg\max\limits_{j\in\{1,2,\cdots,k\}}\{a_{pj}\}$

(7) 将样本x_p分配到类簇$C_{\lambda_j}=C_{\lambda_j}\bigcup\{x_p\}$中

(8) 计算新的聚类中心$\mu_j'=\dfrac{1}{\left|C_j\right|}\sum\limits_{x\in C_j}x$

(9) End for

(10)直到$\left|\mu_j'-\mu_j\right|<\delta$或者达到最大迭代次数

(11)For $j=1,2,\cdots,k$ do

(12) 对于类簇C_j中的每个样本x,依次计算$H=\left\{x:x\in C_j,\exists i,i\neq j,x\in C_i\right\}$;

(13) If $H\neq\varnothing$ then

(14) 将样本x同时分配到类簇C_i和类簇C_j的边界域C_{u}

(15) Else

(16) 根据公式$C_{\mathrm{s}}=\left\{a\geqslant\alpha\text{ or }c<\beta\right\}$,$C_{\mathrm{u}}=\left\{0<a<\alpha\text{ and }c\geqslant\beta\right\}$,将样本分配到类簇$C_j$的正同域$C_{\mathrm{s}}(C_j)$或者边界域$C_{\mathrm{u}}(C_j)$

(17) End if

(18)End for

(19)输出$C=\{\{C_{\mathrm{s}}(C_1)\bigcup C_{\mathrm{u}}(C_1)\},\{C_{\mathrm{s}}(C_2)\bigcup C_{\mathrm{u}}(C_2)\},\cdots,\{C_{\mathrm{s}}(C_k)\bigcup C_{\mathrm{u}}(C_k)\}\}$

上述算法主要分为两个阶段:第1阶段(步骤(1)~(10))是构造包含正同域和边界域的集合。在步骤(5)~(7)中,是对存在缺失值的样本,根据本文提出的集对距离度量方法,计算得到距离每个样本最近的聚类中心,由此将每个样本分配到距离最近的类簇中。为了减少初始聚类中心对聚类结果的影响,在每次分配一个样本后,根据步骤(8)对聚类中心进行更新,进而将所有样本分配到k个类簇中,得到初步聚类结果。第2阶段(步骤(11)~(18))是将正同域和边界域分离,在步骤(12)中判断每个样本是否只存在一个类中,将样本分为只属于一个类和属于多个类两种情况,分别对其采取不同划分方法。步骤(13)~(14)处理的是属于多个类的样本,将其同时分配到这些类簇的边界域,步骤(16)处理的是只属于一个类的样本,根据式(13)进行计算,进而将其分配到类簇的正同域或者边界域,由此得到两个域共同表示的集对k-means聚类结果。算法流程图如图5所示。

2.5 算法复杂度分析

针对时间复杂度,设论域中的样本数目为n,聚类数目为k,属性数目为m。SPKM算法的时间复杂度主要由预处理过程中的数据标准化、步骤(5)中依次对每个样本计算距离和步骤(12)~(16)中将样本划分到相应类簇的正同域或边界域产生。数据标准化过程中,需要对每个样本都进行处理,每个样本有m个属性,故产生的时间复杂度为$O(mn)$;每次迭代计算各个样本与聚类中心的距离产生的时间复杂度为knm,设迭代次数为I,则这部分的时间复杂度为$O(knmI)$;对于每个样本划分到相应类簇的正同域或边界域,须对类中每个元素x_p($p=1,2,\cdots,n$)与所在类簇C_j($j=1,2,\cdots,k$)的聚类中心依次进行计算,这部分产生的时间复杂度为$O(knm)$。所以,SPKM算法的时间复杂度为$O(n)=O(mn)+$

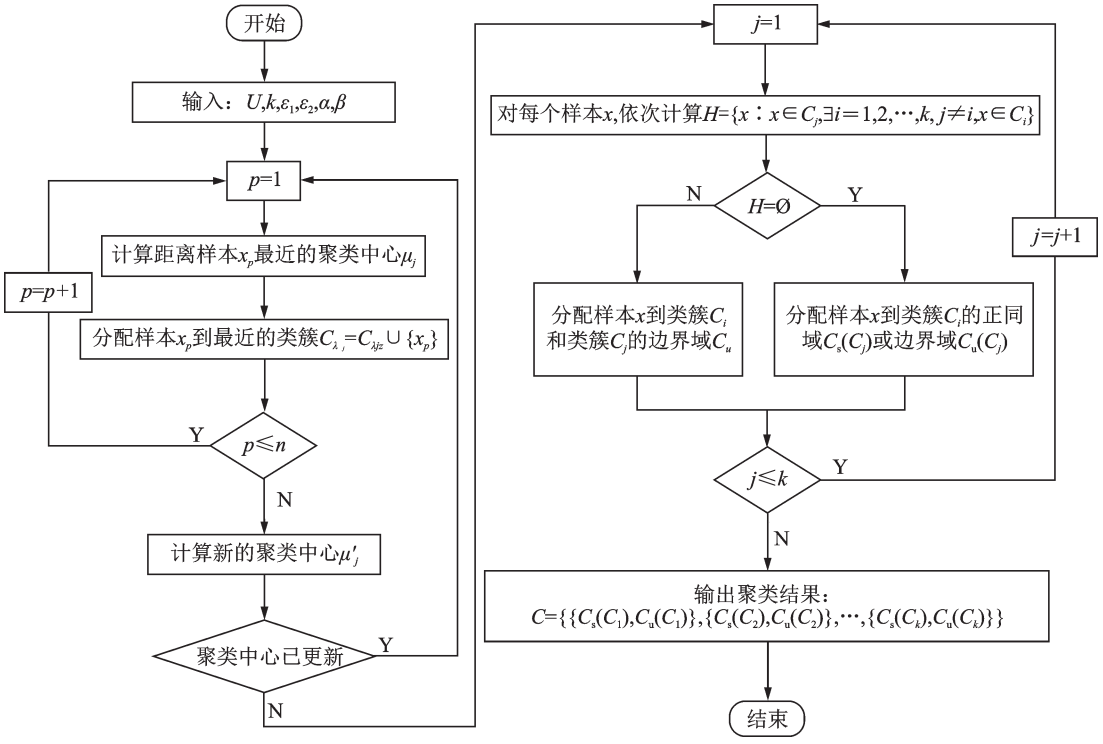


图5 集对k-means聚类算法流程图

Fig.5 Flow chart of set pair k-means clustering algorithm

$O(knmI)+O(knm)$。本文算法在没有增加时间复杂度的前提下，解决了存在缺失值的数据空白问题，提高了聚类效果。

针对空间复杂度，SPKM算法处理的样本数目是n，初始聚类中心的数目是k个，样本的属性值的维数是m，故SPKM算法的空间复杂度为$O((k+n)m)$。

3 模型验证

为了评估所提出的SPKM算法的性能，进行了一系列实验，主要分为两方面：一是对算法参数进行分析，二是选取了4种具有代表性的算法进行对比分析。本文所提算法和对比算法是在一台DELL（Windows 10，Intel(R) Core(TM) i5-8300H，CPU@ 2.30 GHz 2.30 GHz）计算机上使用Python3.7环境实现的。

3.1 数据集选取与评价

聚类的评价，又被称为聚类有效性，是评估学习方法在聚类方面表现的关键过程，度量方法将影响到几种聚类方法的性能比较。为了验证本文算法的性能，选取了UCI中的6个数据集Iris，Wine，Seeds，Liver disorders，Wave form以及Page blocks，表1给出了这些数据集的大小、属性个数和类簇个数。

对于数据集：$U=\{x_1,x_2,\cdots,x_n\}$，假定通过聚类得到的聚类结果为$C=\{C_1,C_2,\cdots,C_k\}$，数据集的真实聚类结果为$C^*=\{C_1^*,C_2^*,\cdots,C_k^*\}$，令$\lambda$和$\lambda^*$分别表示$C$和$C^*$对应的簇标记，将样本两两配对考虑，定义

$$a=|SS|,SS=\{(x_i,x_j)\,|\,\lambda_i=\lambda_j,\lambda_i^*=\lambda_j^*,i<j\}$$

$$b=|SD|,SD=\{(x_i,x_j)\,|\,\lambda_i=\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\}$$

$$c=|DS|,DS=\{(x_i,x_j)\,|\,\lambda_i\neq\lambda_j,\lambda_i^*=\lambda_j^*,i<j\}$$

$$d=|DD|,DD=\{(x_i,x_j)\,|\,\lambda_i\neq\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\}$$

本文选取几种常见的聚类性能度量指标。

(1)Macro F_1(宏平均)

作为一个多标签任务的衡量指标,Macro F_1可表达为

$$\text{Macro}\,F_1=2\cdot\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} \tag{15}$$

式中

$$\text{precision}=\frac{1}{n}\sum_{i=1}^{n}\frac{\text{TP}_{C_i}}{\text{TP}_{C_i}+\text{FP}_{C_i}} \tag{16}$$

$$\text{recall}=\frac{1}{n}\sum_{i=1}^{n}\frac{\text{TP}_{C_i}}{\text{TP}_{C_i}+\text{FN}_{C_i}} \tag{17}$$

式中:TP_{C_i}表示C_i中样本被正确聚为该类的数量;FP_{C_i}表示非C_i中样本被错误聚为该类的数量;FN_{C_i}表示C_i中样本被错误聚为其他类的数量。Macro F_1性能度量的结果在区间[0,1]。

(2)Accuracy(准确率)

$$\text{Accuracy}(C)=\frac{1}{n}\sum_{k=1}^{n}\varphi_k \tag{18}$$

式中:φ_k为簇C_k中正确划分的样本数目,n为样本的总数。准确率越高,聚类结果越好。

(3)JC(Jaccard系数)

$$\text{JC}=\frac{a}{a+b+\text{c}} \tag{19}$$

(4)FMI(FM指数)

$$\text{FMI}=\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}} \tag{20}$$

(5)ARI(兰德指数)

$$\text{ARI}=\frac{\sum_{ij}C^2_{|C_i\cap C_j^*|}\left(\sum_i C^2_{|C_i|}\sum_j C^2_{|C_j^*|}\right)\Big/C_n^2}{\frac{1}{2}\left(\sum_i C^2_{|C_i|}+\sum_j C^2_{|C_j^*|}\right)-\left(\sum_i C^2_{|C_i|}\sum_j C^2_{|C_j^*|}\right)\Big/C_n^2} \tag{21}$$

式中:$|C_i|$为类簇C_i的样本数目;$|C_j^*|$为真实类簇C_j^*的样本数目;$|C_i\cap C_j^*|$表示类簇C_i和真实类簇C_j^*共同拥有的样本数目。

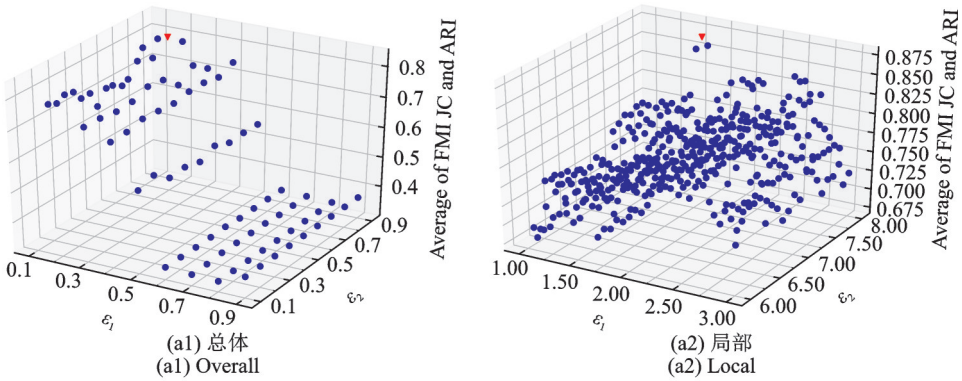
表1 实验数据集的描述

Table 1 Description of lab datasets

ID	数据集	样本数	特征数	聚类数
1	Iris	150	4	3
2	Wine	178	13	3
3	Seeds	210	7	3
4	Liver disorders	345	6	2
5	Wave form	5 000	21	3
6	Page blocks	5 473	10	5

3.2 SPKM算法参数分析

以Iris数据集为例,对算法在不同参数下的性能进行了详细分析。由于数据集是经典完备数据集,需要对其进行处理,随机生成带有缺失值的不完备数据集,本文选取的缺失率为5%,10%,15%和20%,通过使用评价指标JC,FMI和ARI对该聚类算法的性能进行评价。图6给出了3个评价指标平均值随着4个参数变化的波动情况。表2给出了Iris数据集在最优参数下的聚类结果。由于本文提出SP-

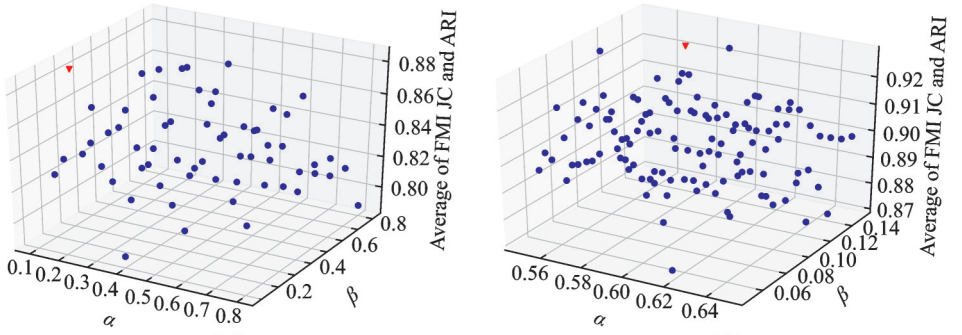


(a1) 总体
(a1) Overall

(a2) 局部
(a2) Local

(a) 正同域和边界域的聚类结果随参数ε_1,ε_2的变化

(a) Variation of clustering results with parameters ε_1,ε_2 in positive and boundary region

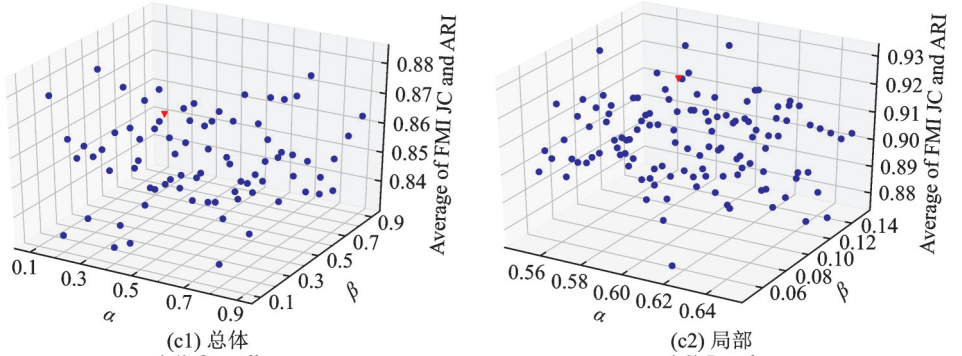


(b1) 总体
(b1) Overall

(b2) 局部
(b2) Local

(b) 正同域聚类结果随参数α, β的变化

(b) Variation of clustering results with parameters α, β in positive region



(c1) 总体
(c1) Overall

(c2) 局部
(c2) Local

(c) 正同域和边界域的聚类结果随参数α, β的变化

(c) Variation of clustering results with parameters α, β in positive and boundary regions

图6 数据集Iris的参数变化结果

Fig.6 Result of parameter changes in Iris dataset

KM算法是以正同域和边界域来表示聚类结果,所以在每个指标下都分别给出了单独的正同域C_s以及正同域和边界域$C_s \cup C_u$这两个评价结果。

首先讨论的是参数ε_1,ε_2,以运行100次的均值作为一次实验值,进行大量实验,图6(a2)是相对图6(a1)的一个局部图,是通过缩小步长进行的更为精细的处理,得到最佳参数为$\varepsilon_1=0.16$,$\varepsilon_2=0.79$,其对应的100次均值的最大实验值达到0.87(正同和边界域),最优值在图中已用红色三角进行区分。然

表2 Iris数据集在最优参数下的性能分析

Table 2 Performance analysis of Iris dataset under optimal parameters

缺失率/%	C_s			$C_s \cup C_u$		
	JC	FMI	ARI	JC	FMI	ARI
5	0.947	0.946	0.920	0.993	0.987	0.980
10	0.847	0.853	0.785	0.973	0.947	0.921
15	0.713	0.683	0.544	0.893	0.808	0.712
20	0.593	0.632	0.464	0.860	0.760	0.639

后保持$\varepsilon_1=0.16$,$\varepsilon_2=0.79$不变,讨论参数α和β。用同样的方法得到的最佳参数为$\alpha=0.6$,$\beta=0.09$,其对应的100次均值的最大实验值达到0.922(正同域),0.923(正同和边界域)。从图6可知数据集Iris的最优参数为$\varepsilon_1=0.16$,$\varepsilon_2=0.79$,$\alpha=0.6$,$\beta=0.09$。对于其他5个数据集也通过实验方式得出了最优参数,具体结果如表3所示。

表3 6个数据集的最优参数

Table 3 Optimal parameters for six datasets

数据集	ε_1	ε_2	α	β
Iris	0.16	0.79	0.60	0.09
Wine	0.18	0.28	0.69	0.24
Seeds	0.11	0.23	0.15	0.30
Liver disorders	0.10	0.31	0.70	0.38
Wave form	0.26	0.44	0.74	0.05
Page blocks	0.28	0.32	0.90	0.10

表2给出的是数据集Iris在最优参数下的聚类结果,可以看出在指标JC,FMI和ARI下均得到了较好的聚类效果。其中,针对正同域和边界域$C_s \cup C_u$,在缺失率为5%的情况下得到的JC指数为0.993,FMI为0.987,ARI为0.980;在缺失率为10%的情况下达到了JC为0.973,FMI为0.947,ARI为0.921。同时,从表中也可以看到,针对正同域C_s,得到的评价指标的值均低于在$C_s \cup C_u$下的。这是因为这两个集合是通过对二支聚类的收缩或扩展得到的,所以$C_s \cup C_u$的性能优于C_s是合理的。通过这3个指标的实验结果可以分析出,随着缺失率的增加,聚类质量下降,究其原因是缺失率越大,不确定信息越多,其对样本划分产生的影响越大。

在最优参数下,对Iris数据集的集对k-means聚类结果进行了可视化,结果如图7所示。从图7可知,数据集被聚成了3个类簇,每个类簇通过正同域和边界域两个域共同表示。从图7可以看出,同时属于多个类簇的样本位于多个类簇的交界处,将其分配到边界域是合理的;对于只属于一个类,但其位于类簇的边缘区域的样本,将其分配到边界域也是合理的。

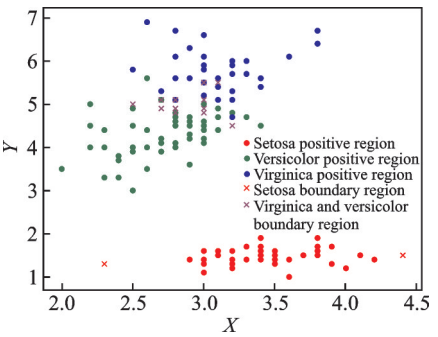


图7 Iris数据集最优参数下聚类结果

Fig.7 Clustering result under optimal parameters of Iris dataset

3.3 SPKM算法的实验对比分析

为了识别SPKM算法的聚类质量,通过评价指标Accuracy,Macro F_1,JC,FMI和ARI与选取的4种算法进行了对比分析。第1种是Wang等基于数学形态学的腐蚀和膨胀提出的三支聚类框架(CE3-k-means)[9];第2种是Yu等基于改进的DBSCAN提出的三支聚类算法(3W-DBSCAN)[10];第3种是Zhang基于三支权重和三支分配,提出的一种三支c-means聚类算法[15],依照原文将其记为TCM;第4种是Yang等基于密度峰值提出的不完备三支聚类算法,依照原文将其记为Adopted methods[16]。

将对所有正同域和边界域形成的聚类结果进行实验评价。由于本文的研究对象是不完备数据集，故将选取的6个数据集按照5%，10%，15%，20%的比例作随机缺失处理。本文所提算法和4种对比算法在5个评价指标下的聚类结果如表4—8所示。在不同算法之间的最优结果用粗体进行标记。

表4 5种算法在指标Accuracy下的对比结果

Table 4 Comparison results of five algorithms under index Accuracy

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
Iris	5	0.865	0.606	0.873	0.933	**0.993**
	10	0.823	0.560	0.853	0.913	**0.973**
	15	0.785	0.593	0.832	0.873	**0.893**
	20	0.738	0.520	0.793	0.820	**0.860**
Wine	5	0.702	0.348	0.743	0.725	**0.915**
	10	0.674	0.331	0.722	0.711	**0.859**
	15	0.646	0.298	0.713	0.698	**0.836**
	20	0.629	0.236	0.685	**0.689**	0.644
Seeds	5	0.861	0.524	0.858	0.743	**0.869**
	10	0.828	0.367	0.850	0.732	**0.851**
	15	0.805	0.205	**0.836**	0.719	0.812
	20	0.786	0.200	0.802	0.698	**0.803**
Liver disorders	5	0.575	0.420	0.548	0.432	**0.581**
	10	0.566	0.409	0.527	0.422	**0.569**
	15	0.559	0.420	0.516	0.419	**0.562**
	20	**0.560**	0.426	0.498	0.390	0.544
Wave form	5	0.560	0.329	0.481	0.512	**0.613**
	10	0.550	0.309	0.448	0.503	**0.602**
	15	0.536	0.298	0.429	0.510	**0.601**
	20	0.506	0.298	0.412	0.493	**0.599**
Page blocks	5	0.867	0.848	0.755	**0.880**	**0.880**
	10	0.857	0.837	0.742	**0.871**	0.852
	15	0.562	0.831	0.737	**0.851**	0.849
	20	0.596	0.808	0.721	0.840	**0.842**

表5 5种算法在指标Macro F_1下的对比结果

Table 5 Comparison results of five algorithms under index Macro F_1

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
Iris	5	0.865	0.773	0.913	0.929	**0.993**
	10	0.823	0.667	0.900	0.901	**0.973**
	15	0.785	0.633	0.887	0.867	**0.893**
	20	0.738	0.527	0.851	0.852	**0.860**
Wine	5	0.702	0.348	0.785	0.719	**0.881**
	10	0.674	0.331	0.712	0.702	**0.864**
	15	0.646	0.303	0.699	0.698	**0.842**

续表

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
	20	0.629	0.254	0.625	0.667	**0.833**
Seeds	5	0.861	0.509	0.775	0.893	**0.904**
	10	0.828	0.409	0.735	0.886	**0.900**
	15	0.805	0.357	0.727	**0.887**	0.857
	20	0.786	0.309	0.715	0.805	**0.833**
Liver disorders	5	0.575	0.428	0.550	0.432	**0.730**
	10	0.566	0.373	0.519	0.443	**0.718**
	15	0.559	0.379	0.502	0.402	**0.653**
	20	0.560	0.371	0.486	0.389	**0.631**
Wave form	5	0.562	0.329	0.401	0.391	**0.613**
	10	0.537	0.170	0.391	0.385	**0.602**
	15	0.520	0.124	0.401	0.386	**0.601**
	20	0.492	0.121	0.387	0.374	**0.599**
Page blocks	5	0.867	0.873	0.690	0.659	**0.880**
	10	0.851	0.748	0.681	0.642	**0.852**
	15	**0.857**	0.543	0.682	0.644	0.849
	20	0.823	0.370	0.669	0.626	**0.842**

表6 5种算法在指标JC下的对比结果

Table 6 Comparison results of five algorithms under index JC

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
Iris	5	0.865	0.630	0.933	0.721	**0.993**
	10	0.823	0.500	0.931	0.715	**0.973**
	15	0.785	0.463	0.890	0.704	**0.893**
	20	0.738	0.357	0.858	0.699	**0.860**
Wine	5	0.702	0.211	0.691	0.540	**0.915**
	10	0.674	0.199	0.687	0.523	**0.859**
	15	0.646	0.175	0.667	0.502	**0.836**
	20	0.629	0.134	0.641	0.498	**0.644**
Seeds	5	0.831	0.355	0.619	0.601	**0.850**
	10	0.828	0.224	0.603	0.598	**0.846**
	15	0.805	0.114	0.598	0.587	**0.812**
	20	**0.786**	0.111	0.577	0.575	0.603
Liver disorders	5	0.575	0.273	0.407	0.195	**0.581**
	10	0.566	0.271	0.487	0.185	**0.569**
	15	0.559	0.266	0.464	0.167	**0.563**
	20	**0.560**	0.271	0.459	0.159	0.544
Wave form	5	0.562	0.197	0.457	0.501	**0.598**
	10	0.537	0.093	0.426	0.500	**0.552**
	15	0.520	0.066	0.425	0.487	**0.575**

续表

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
	20	0.492	0.066	0.408	0.466	**0.573**
Page blocks	5	0.867	0.801	0.715	0.525	**0.880**
	10	0.851	0.703	0.709	0.520	**0.852**
	15	**0.857**	0.522	0.689	0.515	0.849
	20	0.823	0.351	0.671	0.501	**0.842**

表7 5种算法在指标FMI下的对比结果

Table 7 Comparison results of five algorithms under index FMI

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
Iris	5	0.785	0.719	0.875	0.753	**0.987**
	10	0.775	0.771	0.859	0.741	**0.947**
	15	0.687	0.729	**0.847**	0.726	0.808
	20	0.659	0.678	**0.834**	0.718	0.760
Wine	5	0.572	0.569	0.536	0.600	**0.838**
	10	0.543	0.581	0.515	0.595	**0.750**
	15	0.522	0.564	0.505	0.588	**0.726**
	20	0.494	**0.573**	0.499	0.560	0.426
Seeds	5	0.762	0.585	0.806	0.665	**0.812**
	10	0.717	0.534	**0.790**	0.651	0.783
	15	0.676	0.504	0.778	0.644	**0.780**
	20	0.649	0.536	0.669	0.623	**0.703**
Liver disorders	5	0.543	0.538	0.529	0.507	**0.608**
	10	0.548	0.515	0.520	0.482	**0.578**
	15	0.552	0.615	0.515	0.481	**0.696**
	20	0.549	0.456	0.503	0.467	**0.551**
Wave form	5	0.495	0.616	0.499	0.398	**0.675**
	10	0.229	0.577	0.488	0.412	**0.594**
	15	0.522	0.535	0.481	0.407	**0.654**
	20	0.478	0.535	0.476	0.381	**0.610**
Page blocks	5	0.881	**0.895**	0.789	0.548	0.852
	10	0.721	0.723	0.771	0.527	**0.774**
	15	0.681	0.667	**0.765**	0.531	0.700
	20	0.581	0.506	0.662	0.522	**0.673**

表8 5种算法在指标ARI下的对比结果

Table 8 Comparison results of five algorithms under index ARI

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
Iris	5	0.675	0.559	0.791	0.622	**0.980**
	10	0.662	0.568	0.781	0.615	**0.921**
	15	0.530	0.511	0.645	0.600	**0.712**

续表

数据集	缺失率/%	CE3-k-means	3W-DBSCAN	TCM	Adopted methods	SPKM
	20	0.476	0.443	0.621	0.590	**0.639**
Wine	5	0.354	0.493	0.290	0.387	**0.756**
	10	0.308	0.428	0.281	0.367	**0.624**
	15	0.268	0.445	0.266	0.360	**0.576**
	20	0.216	0.204	0.247	0.354	**0.426**
Seeds	5	0.644	0.244	**0.650**	0.459	0.481
	10	**0.575**	0.161	0.519	0.441	0.479
	15	0.434	0.319	0.474	0.429	**0.477**
	20	0.426	0.113	0.441	0.420	**0.457**
Liver disorders	5	−0.004	**0.537**	0.028	0.015	0.308
	10	−0.003	**0.465**	0.020	0.011	0.237
	15	−0.008	0.187	0.003	0.002	**0.203**
	20	−0.010	0.096	0.005	0.001	**0.103**
Wave form	5	0.267	**0.289**	0.049	0.005 0	0.212
	10	0.254	0.214	0.027	−0.001 0	**0.264**
	15	0.159	0.162	0.007	−0.001 3	**0.178**
	20	0.117	0.114	0.001	−0.014 6	**0.132**
Page blocks	5	0.027	0.238	0.128	0.098	**0.246**
	10	0.021	**0.189**	0.087	0.072	0.188
	15	0.021	0.123	0.043	0.040	**0.177**
	20	0.021	0.089	0.007	0.024	**0.098**

从表4中的实验结果可知,SPKM算法在评价指标Accuracy上得到的聚类结果要优于对比算法,特别是在数据集Iris,Liver disorders和Wave form,该算法的准确率均高于对比算法。对于数据集Wine和Seeds,本文算法在4个缺失率下,都能达到在3个缺失率下优于对比算法,其中Wine在缺失率为20%时,比最优算法低了0.045,Seeds在缺失率为15%时,能做到优于3个对比算法。针对数据集Page blocks,在缺失率为5%和20%时,准确率仍高于所有对比算法,但在缺失率为10%,15%时,与最优对比算法Adopted methods相比分别下降了0.019,0.002,但仍比其他3个对比算法的效果要优。

从表5的实验结果可以看出,对于数据集Iris,Wine,Liver disorders和Wave form,用评价指标Macro F_1得到的聚类结果要明显高于对比算法,也体现出了SPKM算法的优越性。对于其他两个数据集也得到了不错的聚类结果,比如数据集Seeds,在缺失率为15%的情况下,即使它不是最好的算法,但在缺失率为5%,10%和20%时,Macro F_1值要高于其他4个对比算法。从表6中的实验结果可以看出,对于数据集Iris,Wine和Wave form,用评价指标JC得到的聚类结果均优于对比算法。从表7中的实验结果可知,在4个缺失率下数据集Liver disorders和Wave form得到的聚类结果均优于对比算法。从表8的实验结果可知,对于数据集Iris和Wine,用评价指标ARI得到的聚类结果要明显高于对比算法。综上所述,本文提出的聚类算法可以有效处理含有缺失值的不完备数据集,而且具有良好的聚类性能。

4 结束语

针对不完备数据的聚类问题以及为了更好地表示样本与类簇的关系,本文将集对分析的相关理论

引入到k-means聚类中，提出了一种面向不完备信息系统的集对k-means聚类算法，将聚类结果划分为3部分，用正同域C_s，边界域C_u，负反域C_o表示。本文提出的集对k-means聚类算法，将原本聚类过程中不同样本之间距离扩展成包含正同度、差异度和负反度3个维度的距离定义，有效解决了不完备数据集的距离度量问题，并基于这个定义扩展了k-means聚类算法，很好地表示了样本与类之间的3种关系。同时，对于样本可能和多个类有关系的情况，也给出了相应的聚类方法，将其划分到相应类的边界域。实现了对聚类结果结构的改进以及聚类准确率的提高。通过对6个UCI数据集进行对比实验，结果表明该算法可以有效解决具有缺失值的不完备数据集，并且改善了聚类结构。然而，参数的变化对聚类结果的影响较大，如何获取更为合适的参数也是下一步的工作，其将对提高聚类质量有重要影响。

参 考 文 献：

[1] 贾凡，严妍，张家琪. 基于K-means聚类特征消减的网络异常检测[J]. 清华大学学报(自然科学版)，2018，58(2)：137-142.
JIA Fan，YAN Yan，ZHANG Jiaqi. K-means based feature reduction for network anomaly detection[J]. Journal of Tsinghua University(Science and Technology)，2018，58(2)：137-142.

[2] 张仁斌，许辅昊，刘飞，等. 基于K-均值聚类的工业异常数据检测[J]. 计算机应用研究，2018，35(7)：2180-2184.
ZHANG Renbin，XU Fuhao，LIU Fei，et al. Industrial anomaly data detection based on K-mean clustering[J] Computer Applied Research，2018，35(7)：2180-2184.

[3] MOSHFEGH S，ASHOURI A，MANDAVIFAR S，et al. Integrable-chaos crossover in the spin-1/2 XXZ chain with cluster interaction[J]. Physica A：Statistical Mechanics and its Applications，2019，516：502-508.

[4] JIAO Pengfei，YU Wei，WANG Wenjun，et al. Exploring temporal community structure and constant evolutionary pattern hiding in dynamic networks[J]. Neurocomputing，2018，314：224-233.

[5] YAO Yiyu. Three-way decision and granular computing[J]. International Journal of Approximate Reasoning，2018，103：107-123.

[6] YAO Yiyu. Three-way decisions and cognitive computing[J]. Cognitive Computation，2016，8(4)：543-554.

[7] 于洪. 三支聚类分析[J]. 数码设计，2016，5(1)：31-35.
YU Hong. Three-way cluster analysis[J]. Peak Data Science，2016，5(1)：31-35.

[8] WANG Pingxin，SHI Hong，YANG Xibei，et al. Three-way K-means：Integrating K-means and three-way decision[J]. International Journal of Machine Learning and Cybernetics，2019，10(10)：2767-2777.

[9] WANG Pingxin，YAO Yiyu. CE3：A three-way clustering method based on mathematical morphology[J]. Knowledge-Based Systems，2018，155：54-65.

[10] YU Hui，CHEN Luyuan，YAO Jingtao，et al. A three-way clustering method based on an improved DBSCAN algorithm[J]. Physica A：Statistical Mechanics and its Applications，2019，535：1-14.

[11] AYDILEK I B，ARSLAN A. A hybrid method for imputation of missing values using optimized fuzzy C-means with support vector regression and a genetic algorithm[J]. Information Sciences，2013，233：25-35.

[12] LI Dan，GU Hong，ZHANG Liyong. A fuzzy C-means clustering algorithm based on nearest-neighbor intervals for incomplete data[J]. Expert Systems with Applications，2010，37(10)：6942-6947.

[13] ZHANG Li，BING Zhaohong，ZHANG Liyong. A hybrid clustering algorithm based on missing attribute interval estimation for incomplete data[J]. Pattern Analysis and Applications，2015，18(2)：377-384.

[14] 张春英，王立亚，李明霞，等. 基于集对信息粒空间的三支决策模型及应用[J]. 通信学报，2016，37(S1)：19-28.
ZHANG Chunying，WANG Liya，LI Mingxia，et al. Model of three-way decision based on the space of set pair information granule and its application[J]. Journal of Communications，2016，37(S1)：19-28.

[15] ZHANG Kai. A three-way C-means algorithm[J]. Applied Soft Computing Journal，2019，82：1-7.

[16] YANG Lin, HOU Kaiyan. A method of incomplete data three-way clustering based on density peaks[C]//AIP Conference Proceedings. Busan, South Korea: [s.n.], 2018: 1-7.

作者简介:

张春英(1969-),女,教授,研究方向:数据挖掘、粗糙集,E-mail:hblg_zcy@126.com。

高瑞艳(1995-),女,硕士研究生,研究方向:数据挖掘、集对分析。

刘凤春(1976-),**通信作者**,男,副教授,研究方向:数据挖掘、机器学习,E-mail:18849778@qq.com。

王佳昊(2000-),男,本科生,研究方向:大数据和高性能计算。

陈松(2000-),男,本科生,研究方向:数据挖掘、聚类分析。

冯晓泽(1994-),女,硕士研究生,研究方向:数据挖掘、三支决策。

任静(1995-),女,硕士研究生,研究方向:机器学习、决策树。

(编辑:王静)