一种带权的混合数据聚类个数确定算法

李顺勇 张苗苗
(山西大学数学科学学院 山西 太原 030006)

摘 要 混合数据的聚类过程中通常面临一个不可回避的问题:聚类个数的确定。基于 Liang k-prototype 算法引入属性权重,重新定义混合数据缺失某类的类间熵和(SBAE_M)、有效性指标(CUM) 及相异性度量。提出一种带权的混合数据聚类个数确定算法。该算法的基本思想是:用 newk-prototype 算法将混合数据进行聚类,计算其聚类结果的 CUM 及 SBAE_M,将最坏的类剔除,并将该类中的对象用新的相异性度量进行重新分配,CUM 最大时包含的类别数即为聚类个数。在 5 个 UCI 数据集上验证了该算法的有效性。

关键词 聚类个数 混合数据 属性权重 有效性指标

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.050

A WEIGHTED CLUSTERING NUMBER DETERMINING ALGORITHM FOR MIXED DATA

Li Shunyong Zhang Miaomiao
(*School of Mathematical Sciences*, *Shanxi University*, *Taiyuan 030006*, *Shanxi*, *China*)

Abstract Determining the number of clusters is an unavoidable problem in the clustering process of mixed data. This paper introduced attribute weight on the basis of Liang k-prototype algorithm, redefined the sum of between-cluster entropies in absence of a cluster(SBAE_M), the validity index(CUM) and the dissimilarity measure of mixed data, and proposedaweighted algorithm for determining the number of mixed data clustering. New k-prototype algorithm was used to cluster the mixed data. CUM and SBAE_M of the clustering results were calculated and the worst class was eliminated. The objects in this class were reassigned with new dissimilarity measure. The number of categoriesincluding at the maximum of CUM was the number of clusters. The effectiveness of the improved k-prototype clustering algorithm was verified on five data sets from UCI.

Keywords The number of clustering Mixed data Attribute weight Validity index

0 引 言

聚类分析是一种无监督学习,聚类个数是提前不确定的。因此,确定聚类个数是聚类分析的一个重要课题。针对数值型数据,文献[1]是基于 FCM 算法[2]的自动确定聚类个数的算法,通过比较不同参数下的有效性指标的大小来确定类别个数。文献[3]提出一种基于层次划分的最佳聚类数确定方法,该算法通过寻找一条聚类质量曲线的极值点来估计最佳聚类个数,并提出一种新的有效性指标。文献[4]提出了一种用于数值数据的凝聚的模糊 k-means 聚类算法,是对标准模糊 k-means 算法的扩展,通过在目标函数中引入惩罚项来使聚类过程对数据不敏感。文献[5]提出了一种利用信息论和自上而下的层次聚类算法在数据集中查找聚类数的新方法。该算法从大量集群开始,在每次迭代中减少一个集群,然后将其数据点分配给其余集群。最后,通过测量信息势,检测所需数据集中的确切类别数。针对分类型数据,文献[6]通过研究分类型数据的熵特性提出一种确定类别数的层次聚类算法。实验结果表明,该方法能够有效地识别显著的聚类结构。文献[7-8]也是对分类型数据聚类个

收稿日期:2018-07-27。国家自然科学基金项目(61573229);山西省基础研究计划项目(201701D121004);山西省回国留学人员科研项目(2017-020);山西省高等学校教学改革创新项目(J2017002)。**李顺勇**,副教授,主研领域:统计机器学习。**张苗苗**,硕士生。

数的研究。

但日常生活中我们接触的数据大部分是混合数据,即数据集中既有数值型数据也有分类型数据。文献[9-13]对混合型数据进行了聚类,但都得预先设定聚类个数。而对混合型数据集聚类个数的研究就相对发展得慢了一些,文献[14]基于信息熵提出了确定聚类最佳个数的算法,该算法提出了改进的 k-prototype,并通过将类内熵[15]与类间熵[16]结合建立了一种广义机制,最后定义了有效的聚类有效性指标。但 Liang的算法中目标函数只考虑了数值型属性、分类型属性以及总的属性个数,而没有反映各个属性的重要度。文献[17]提出的是基于属性加权的分类数据聚类算法。而本文改进了 Liang k-prototype 聚类算法,在其基础上考虑属性重要度,通过加权方式提出了混合数据聚类个数算法,并在 4 个 UCI 数据集上验证了 k-prototype 聚类改进算法的有效性。

1　相关研究

本节是对 Huang[18] k-prototype 聚类算法及 Liang 等[19]改进的 k-prototype 算法的回顾。

Huang 较早提出了 k-prototype 算法来处理混合型数据集。该算法融合了 k-means 算法和 k-modes 算法,其中 k-modes 算法中对分类型数据的处理采用了 0、1 匹配的方式。该算法使得以下目标函数达到最小:

$$F = \sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{li}d(x_i,Q_l) \tag{1}$$

式中:$d(x_i,Q_l) = \sum_{j=1}^{p}(x_{ij}^r - q_{lj}^r)^2 + \mu_l\sum_{j=p+1}^{m}\delta(x_{ij}^c,q_{lj}^c)$。

为了方便研究,本文对 Huang 的目标函数进行重新改写:

$$D(x,y) = D_{A^{\text{num}}}(x,y) + \gamma D_{A^{\text{cat}}}(x,y) \tag{2}$$

式中:$D_{A^{\text{num}}}(x,y)$代表数值型属性的值,$D_{A^{\text{cat}}}(x,y)$代表分类型属性的值。

文献[19]实质上也是 γ 的变异。首先,他将 Renyi 熵[15]与互补熵[19]整合起来,提出了一个可以将最坏类剔除的广义性机制;其次,为了评价聚类结果的好坏,定义了一个聚类有效性指标 CUM;最后,在改进的 k-prototype 算法[14]基础上提出了混合数据聚类个数确定算法。其相关概念定义如下:

定义 1　设 $NDIS = <U, AT = C\cup D, V, f>$ 是一个数值型决策信息系统,若已经形成 k 类,即 $c_i(i=1,2,\cdots,k)$类,$\forall A_t \in AT, (t=1,2,\cdots,|C|)$,则对 $\forall C_r, C_r \in C^k$,缺失一类后的类间熵和 *SBAE_N* 定义为:

$$SBAE\_N(C_r) = \sum_{C_i\in C^k, i\neq r}\ \sum_{C_j\in C^k, j\neq r} BE\_N(C_i,C_j) \tag{3}$$

其中:

$$BE\_N(C_i,C_j) = -\log\frac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{y\in C_j} W_{2\sigma^2}(x,y) \tag{4}$$

$$W_{\sigma^2}(x,y) = \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\left(-\frac{(x-y)^{\mathrm{T}}(x-y)}{2\sigma^2}\right) \tag{5}$$

$BE_N(C_i,C_j)$表示不同类 C_i, C_j的类间熵,熵越大表示类间不确定性越大,分散度越大。因此,定义的 *SBAE_N* 值越大,表示缺失一类后的类间分散度越大,聚类效果越好。而类间熵又由 Renyi 熵[15]来定义,Renyi 熵可通过 Parzen 窗口估计法[16]得到的概率密度得到。这里,用高斯核函数 $W_{\sigma^2}(x,y)$作为 Parzen 窗函数,σ^2表示窗宽。

定义 2　设 $CDIS = <U, AT = C\cup D, V, f>$ 是一个分类型决策信息系统,若已经形成 k 类,即 $c_i(i=1,2,\cdots,k)$类,$\forall A_j \in AT, (j=1,2,\cdots,|C|)$,令 $c_i/IND(\{A_j\}) = \{X_1, X_2, \cdots, X_k\}$,缺失一类后的类间熵和 *SBAE_C* 定义为:

$$SBAE\_C(C_r) = \sum_{C_i\in C^k, i\neq r}\ \sum_{C_j\in C^k, j\neq r} BE\_C(C_i,C_j) \tag{6}$$

其中:

$$BE\_C(C_i,C_j) = \frac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{y\in C_j} D_{A^{\text{cat}}}(x,y) \tag{7}$$

这里从距离的角度定义分类型数据的类间熵,因为经验证类内熵与类内平均距离是等价的。其中:

$$D_{A^{\text{cat}}}(x,y) = \sum_{A\in A^{\text{cat}}} d_A(x,y)$$

$$d_A(x,y) = \begin{cases} 0 & f(x,A) = f(y,A) \\ 1 & f(x,A) \neq f(y,A) \end{cases}$$

定义 3　设 $MDIS = <U, C(=A^{\text{num}}\cup A^{\text{cat}})\cup D, V, f>$ 为一个混合决策信息系统,其中 $C = A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 $k(k\geqslant 2)$类,即 $C^k = \{C_1, C_2, \cdots, C_k\}$,则对 $\forall C_r, C_r \in C^k$,缺失一类后的类间熵和 *SBAE_M* 定义为:

$$SBAE\_M(C_r) = \frac{|A^{\text{num}}|}{|C|}\frac{SBAE\_N(C_r)}{\sum_{i=1}^{k} SBAE\_N(C_i)} + \frac{|A^{\text{cat}}|}{|C|}\frac{SBAE\_C(C_r)}{\sum_{i=1}^{k} SBAE\_C(C_i)} \tag{8}$$

$SBAE\_N(C_r)$、$SBAE_C(C_r)$分别为数值型数据及分类型数据集中形成的这个类 C_r是不好的类的数值。$SBAE_M(C_r)$越大表示缺失类 C_r后的类间熵和越大,类间分散度越大,聚类效果越好。

定义 4　设 $MDIS = <U, C\cup D, V, f>$ 为一个混合

决策信息系统,其中 $C = A^{num} \cup A^{cat}$,若系统已被分成 $k(k \geqslant 2)$类,即 $C^k = \{C_1, C_2, \cdots, C_k\}$,则有效性指标 CUM 定义为:

$$CU\_M(C^k) = \frac{|A^r|}{|A|} CU\_N(C^k) + \frac{|A^c|}{|A|} CU\_C(C^k) \quad (9)$$

式中:CU_N、CU_C 分别表示数值型及分类型数据的分类效用值。CU_M 值越大,得到的聚类效果越好。

定义 5 设 $MDIS = <U, C \cup D, V, f>$ 为一个混合决策信息系统,其中 $C = A^{num} \cup A^{cat}$,则相异性度量定义为:

$$D'(x,z) = \frac{|A^r|}{|A|} \frac{D_{A^r}(x,z)}{\sum_{i=1}^{k} D_{A^r}(x,z_i)} + \frac{|A^c|}{|A|} \frac{D_{A^c}(x,z)}{\sum_{i=1}^{k} D_{A^c}(x,z_i)} \quad (10)$$

式中:$|A^r|$、$|A^c|$和$|A|$分别表示数值型属性、分类型属性以及总的属性个数。

算法 1 Liang k-prototype 算法

Step 1:从混合数据决策传统 MDIS 中随机选取 k_{max} 个对象 $\{z_1, z_2, \cdots, z_{k_{max}}\}$ 作为初始类中心。

Step 2:i 从 k_{max} 到 k_{min}

Step 2.1:用改进的 k-prototype 对其进行聚类,得出聚类结果 $C^i = \{C_1, C_2, \cdots, C_i\}$;

Step 2.2:计算聚类结果 C^i 的有效性指标 $CUM(C^i)$ 和 $SBAE\_M(C_i)$,找出最差的类 $C_w, C_w \in C^i$;

Step 2.3:对任意的 $x \in C_w$,用相异性度量将其分配到最近的类中;

Step 2.4:更新类中心,将其作为下次迭代的初始中心。

Step 3:比较计算出的有效性指标,选出使 $CUM(C^i)$ 最大的聚类个数 k,$k = \arg\max_{i=k_{max},\cdots,k_{min}} CUM(C^i)$。

Liang k-prototype 算法能够简单、有效地确定混合数据聚类个数,但在缺失一类的类间熵和 SBAE、有效性指标 CUM 和相异性度量的定义上只考虑了数值型、分类型及总的属性个数,而没有考虑每个属性的权重。因此,本文提出了一种带权的混合数据聚类个数算法。

2 带权的混合数据聚类个数确定算法

聚类分析中数据集中的每一个对象都应划分到一个类别中去,这和一个对象的属性有直接的关系,而多个属性在一个对象中的权重不尽相同。为此,在聚类过程中为每个属性赋予一定的权重必将对聚类产生积极的影响。本节将讨论引入权重后如何确定混合数据最佳类别数问题。在提出新算法前,先定义数值型、分类型数据的属性重要度,并根据属性重要度定义混合型数据的属性权重。

2.1 聚类分析中属性权重

对于数值型数据,文献[21]在 Renyi 熵的基础上定义了在某一属性下的类内熵。在某一属性下的类内熵越小,则该属性在这类的不确定性越小,该属性重要度越大,权重越大。因此,本文将数值型属性的重要度定义为类内熵形式。

定义 6 设 $NDIS = <U, AT = C \cup D, V, f>$ 是一个数值型决策信息系统,若已经形成 k 类,即 $c_i(i = 1,2,\cdots,k)$类,$\forall A_t \in AT, t = 1,2,\cdots,|C|$,则数值属性 A_t 在 c_i 中的重要度定义为:

$$WE\_N(C_r, A_t) = -\log \frac{1}{|C_r|^2} \sum_{x_i \in C_r} \sum_{x_j \in C_r} W_{2\sigma^2}(x_{i,t}, x_{j,t}) \quad (11)$$

式中:$W_{\sigma^2}(x,y) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{(x-y)^{\mathrm{T}}(x-y)}{2\sigma^2}\right)$。

对于分类型数据,文献[21]基于文献提出的互补熵定义了分类型数据的类内熵。同数值型数据类似,某分类型属性在某类的类内熵越大,该属性在该类的重要度越大,权重越大。

定义 7 设 $CDIS = <U, AT = C \cup D, V, f>$ 是一个分类型决策信息系统,若已经形成 k 类,即 $c_i(i = 1,2,\cdots,k)$类,$\forall A_j \in AT, j = 1,2,\cdots,|C|$,令 $c_i/IND(\{A_j\}) = \{X_1, X_2, \cdots, X_k\}$,则分类型属性 A_j 在 c_i 中的重要度为:

$$Sig(c_i, A_j) = \sum_{l=1}^{k} \frac{|X_l|}{|c_i|} \frac{|X_l^c|}{|c_i|} = \sum_{l=1}^{t} \frac{|X_l|}{|c_i|} \left(1 - \frac{|X_l|}{|c_i|}\right) \quad (12)$$

由于 $Sig(c_i, A_j)$ 反映属性 A_j 在 c_i 中的重要度,而且应用了信息熵,故 $Sig(c_i, A_j)$ 的大小直接影响到属性 A_j 在 c_i 中的重要性。

对于混合型数据,它是数值型与分类型数据的结合,故将其分开处理。数值型属性的重要度是每个数值型属性重要度之和,分类型属性重要度是每个分类型属性重要度之和。属性重要度越大,属性权重越大,故结合指数函数的单调性并进行归一化处理,我们有以下定义:

定义 8 设 $MDIS = <U, C \cup D, V, f>$ 是一个混合型决策信息系统,其中 $C = A^{num} \cup A^{cat}$,若已经形成 k 类,即 $c_i(i = 1,2,\cdots,k)$类,则数值属性 A^{num} 及分类属性 A^{cat} 在 c_i 中的权重分别定义为:

$$\alpha_{C_r}^{num} = \frac{\sum_{i=1}^{p} \exp(-WEN(C_r, A_i))}{\sum_{i=1}^{p} \exp(-WE\_N(C_r, A_i)) + \sum_{j=p+1}^{n} \exp(-Sig(C_r, A_j))} \quad (13)$$

$$\beta_{C_r}^{\text{cat}}=\frac{\sum_{j=p+1}^{n}\exp(-Sig(C_r,A_j))}{\sum_{i=1}^{p}\exp(-WE\_N(C_r,A_i))+\sum_{j=p+1}^{n}\exp(-Sig(C_r,A_j))} \tag{14}$$

2.2　寻找最坏类的广义性机制

文献[14]已经提出一个找混合数据中最坏类的广义性机制,其目标函数仍只考虑了数值型属性、分类型属性和总的属性个数,而没有考虑各个属性的不同权重。因此,我们对此目标函数做了重新定义:

定义 9　设 $MDIS=<U,C\cup D,V,f>$ 为一个混合决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 k ($k\geqslant 2$)类,即 $C^k=\{C_1,C_2,\cdots,C_k\}$,则对 $\forall C_r,C_r\in C^k$,有:

$$\overset{\text{new}}{SBAE\_M}(C_r)=\alpha_{C_r}^{\text{num}}\frac{SBAE\_N(C_r)}{\sum_{i=1}^{k}SBAE\_N(C_i)}+\beta_{C_r}^{\text{cat}}\frac{SBAE\_C(C_r)}{\sum_{i=1}^{k}SBAE\_C(C_i)} \tag{15}$$

式中:$\alpha_{C_r}^{\text{num}}$表示数值型属性参数,$\beta_{C_r}^{\text{cat}}$表示分类型属性参数,且 $\eta=\alpha_{C_r}^{\text{num}}+\beta_{C_r}^{\text{cat}}=1$,$\alpha_{C_r}^{\text{num}}$与 $\beta_{C_r}^{\text{cat}}$的具体形式由式(13)、式(14)给出。

从定义可以看出,$\alpha_{C_r}^{\text{num}}+\beta_{C_r}^{\text{cat}}=1$ 实质上还是沿用了 k-prototype 的思想,但其中引入了熵。而且,当 $\forall A_s\in A^{\text{num}}\cup A^{\text{cat}}$,$X_{is}=X_{js}$,我们有 $\alpha_{C_r}^{\text{num}}=\frac{|A^{\text{num}}|}{|C|}$,$\beta_{C_r}^{\text{cat}}=\frac{|A^{\text{cat}}|}{|C|}$,这说明 Liang 定义的$SBAE_M(C_r)$是$\overset{\text{new}}{SBAE_M}$的特殊形式。

缺失一类后的类内熵和$\overset{\text{new}}{SBAE_M}(C_r)$越大,说明缺失该类后的分散度越大,聚类效果越好,该类为最坏的类。

定义 10　设 $MDIS=<U,C\cup D,V,f>$ 为一个混合决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 k ($k\geqslant 2$)类,即 $C^k=\{C_1,C_2,\cdots,C_k\}$,则对 $\forall C_r,C_r\in C^k$,有:

$$C_w=\arg\max_{C_r\in C^k}\overset{\text{new}}{SBAE\_M}(C_r) \tag{16}$$

式中:C_w表示混合数据集中不好的类。

2.3　混合数据聚类的有效性指标

与上文$\overset{\text{new}}{SBAE_M}$的定义相似,本文也在 Liang 定义的有效性指标基础上加上了对每个属性的权重,重新定义如下:

定义 11　设 $MDIS=<U,C\cup D,V,f>$ 为一个混合决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 k ($k\geqslant 2$)类,即 $C^k=\{C_1,C_2,\cdots,C_k\}$,则有:

$$\overset{\text{new}}{CU\_M}=\alpha^{\text{num}}CU\_N(C^k)+\beta^{\text{cat}}CU\_C(C^k) \tag{17}$$

式中:$\overset{\text{new}}{CU_M}$表示混合数据集中形成的稳定 k 类的数值。$\alpha^{\text{num}}=\frac{1}{k}\sum_{i=1}^{k}\alpha_{c_i}^{\text{num}}$,$\beta^{\text{cat}}=\frac{1}{k}\sum_{i=1}^{k}\beta_{c_i}^{\text{cat}}$。式(17)中的 $CU_N(C^k)$和 $CU_C(C^k)$分别定义为:

定义 12　设 $NDIS=<U,C\cup D,V,f>$ 为一个数值型决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 k($k\geqslant 2$)类,即 $C^k=\{C_1,C_2,\cdots,C_k\}$,则有:

$$CU\_N(C^k)=\frac{1}{k}\sum_{l=1}^{|A^{\text{num}}|}\left(\delta_l^2-\sum_{j=1}^{k}p_j\delta_{lj}^2\right) \tag{18}$$

式中:$\delta_l^2=\sum_{x\in U}(f(x,A_l)-m_l)^2/|U|$,$\delta_{lj}^2=\sum_{x\in C_j}(f(x,A_l)-m_{lj})^2/|C_j|$,$p_j=\frac{|C_j|}{|U|}$。

定义 13　设 $CDIS=<U,C\cup D,V,f>$ 为一个分类型决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,若系统已被分成 k($k\geqslant 2$)类,即 $C^k=\{C_1,C_2,\cdots,C_k\}$,则有:

$$CU\_C(C^k)=\frac{1}{k}\sum_{A\in A^{\text{cat}}}\sum_{X\in(U/IND(A))}\sum_{i=1}^{k}\frac{|C_i|}{|U|}\left(\frac{|X\cap C_i|}{|C_i|^2}-\frac{|X|^2}{|U|^2}\right) \tag{19}$$

2.4　混合数据相异性度量

从式(2)可以看出 γ 对结果有一定的影响。Huang 已对此进行了研究分别对 γ 取了不同的值进行了讨论。为此我们在混合数据集上重新定义,

$$D(x,y)=\alpha D_{A^{\text{num}}}(x,y)+\beta D_{A^{\text{cat}}}(x,y) \tag{20}$$

式中:α 和 β 分别表示数值型属性和分类型属性的参数,$D_{A^{\text{num}}}(x,y)$表示数值型属性的值,$D_{A^{\text{cat}}}(x,y)$表示分类型属性的值。

从以上定义不难看出,Huang 的文章中 $\alpha=1,\beta=\gamma,\eta=\alpha+\beta=1+\gamma$;Liang 的文章中由于使用了归一化的方法,从而有 $\eta=\alpha+\beta=1$,其中:

$$\alpha=\frac{|A^r|}{|A|},\beta=\frac{|A^c|}{|A|},\eta=\alpha+\beta=\frac{|A^r|+|A^c|}{|A|}=1,$$

$A^r\cup A^c=A$。

由于$|A^r|$、$|A^c|$和$|A|$只是反映了数值型属性、分类型属性以及总的属性个数,而没有反映各个属性的重要度,本文在其基础上加上了对每个属性考虑其权重,则有以下定义:

定义 14　设 $MDIS=<U,C\cup D,V,f>$ 为一个混合决策信息系统,其中 $C=A^{\text{num}}\cup A^{\text{cat}}$,则相异性度量定义为:

$$D^{\text{new}}(x,z)=\alpha^{\text{num}}\frac{D_{A^r}(x,z)}{\sum_{i=1}^{k}D_{A^r}(x,z_i)}+\beta^{\text{cat}}\frac{D_{A^c}(x,z)}{\sum_{i=1}^{k}D_{A^c}(x,z_i)} \tag{21}$$

式中:α^{num}和β^{cat}分别表示数值型属性和分类型属性的属性重要,$\alpha^{num}=\frac{1}{k}\sum_{i=1}^{k}\alpha_{c_i}^{num}$,$\beta^{cat}=\frac{1}{k}\sum_{i=1}^{k}\beta_{c_i}^{cat}$。

基于新的相异性度量,new k-prototype 算法描述如下:

算法 2 new k-prototype 聚类算法

输入:混合数据决策系统 MDIS,α^{num},β^{cat},聚类个数 k。

输出:聚类结果 $C^i=\{C_1,C_2,\cdots,C_i\}$

Step 1:从混合数据决策传统 MDIS 中随机选取 k 个对象作为初始类中心;

Step 2:用式(21)计算每个对象到每个类中心的距离,将其分配到距离最小的类中;

Step 3:更新类中心(数值型属性取均值作为中心,分类型属性取频率最大的最为中心),将其作为下次迭代的初始中心;

Step 4:重复 Step 2,Step 3,当类中心不变时,算法结束。

2.5 混合数据聚类个数确定的算法

基于上文对混合数据剔除最坏类的广义性机制、有效性指标和相异性度量的重新定义,本文提出一种带权的混合数据聚类个数确定的算法,其具体描述如下:

算法 3 k-prototype 聚类改进算法

输入:混合数据决策系统 MDIS,k_{max},k_{min},窗宽 σ

输出:最佳聚类个数 k

Step 1:从混合数据决策传统 MDIS 中随机选取 k_{max} 个对象 $\{z_1,z_2,\cdots,z_{k_{max}}\}$ 作为初始类中心.

Step 2:i 从 k_{max} 到 k_{min}

Step 2.1:给定初始类中心 Z^i,当 $k=k_{max}$ 时,用改进 k-prototype[14] 对其进行聚类;否则,用 new k-prototype 算法对其进行聚类,得出聚类结果 $C^i=\{C_1,C_2,\cdots,C_i\}$;

Step 2.2:计算 $\alpha_{C_r}^{num}$、$\beta_{C_r}^{cat}$和$SBAE_\overset{new}{M}(C_r)$,找出最差的类 C_w,$C_w\in C^i$;

Step 2.3:计算聚类结果 C^i的有效性指标$\overset{new}{CU_M}$;

Step 2.4:对任意的 $x\in C_w$,用式(21)将其分配到最近的类中;

Step 2.5:更新类中心(数值型属性取均值作为中心,分类型属性取频率最大的最为中心),将其作为下次迭代的初始中心;

Step 3:比较计算出的有效性指标,选出使 $CUM(C^i)$最大的聚类个数 k,即 $k=\arg\max_{i=k_{max},\cdots,k_{min}}\overset{new}{CU\_M}(C^i)$。

3 实验结果与分析

为了检验算法的有效性,我们采集 UCI 的 5 个混合数据集在 Liang k-prototype 算法和新提出的 k-prototype 聚类改进算法上进行内外评价指标及聚类结果的比较性分析。对于 k 的取值,我们采取 Bezdek[22] 的建议,设置 $k_{max}=\sqrt{n}$,$k_{min}=2$,n 表示对象个数。数据集的信息描述如表 1 所示。

表 1 数据集信息描述

数据集	对象数	数值型属性	分类型属性	类别数
Statlog	270	7	6	2
CHeart	303	5	8	2
Credit	690	6	8	2
GC	1 000	7	13	2
CMC	1 473	2	7	3

3.1 评价标准

为了对提出算法进行有效评价,本文采用了五个外部评价指标:精度(AC)、纯度(PR)、召回率(RE)、兰德指数(ARI)[23] 和归一化互信息(NMI)[24],一个内部有效性指标:CUM。AC,PR,RE,ARI,NMI 和 CUM 的值越大,聚类结果越接近于数据集的真实类划分,聚类效果越好。

设 X 是一矩阵对象数据集,$C=\{C_1,C_2,\cdots,C_{k'}\}$ 是 X 的聚类结果,$P=\{P_1,P_2,\cdots,P_k\}$ 是真实标签,聚类个数为 k',真实类别数为 k。假定 $k'=k$,其他 5 种评价指标定义如下:

$$AC=\frac{1}{n}\max_{j_1j_2..j_k\in S}\sum_{i=1}^{k}n_{ij_i} \tag{22}$$

$$PR=\frac{1}{k}\sum_{i=1}^{k}\frac{n_{ij_i^*}}{p_i} \tag{23}$$

$$RE=\frac{1}{k'}\sum_{i=1}^{k'}\frac{n_{ij_i^*}}{c_i} \tag{24}$$

$$ARI=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{p_i}{2}\sum_j\binom{c_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{p_i}{2}+\sum_j\binom{c_j}{2}\right]-\left[\sum_i\binom{p_i}{2}\sum_j\binom{c_j}{2}\right]\Big/\binom{n}{2}} \tag{25}$$

$$NMI=\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}n_{ij}\log(n_{ij}n/p_ic_j)}{\sqrt{\sum_{i=1}^{k}p_i\log(p_i/n)\sum_{j=1}^{k'}c_j\log(c_j/n)}} \tag{26}$$

式中:p_i 和 c_j 分别是 P_i 和 C_j 中对象的个数,$n_{ij}=|P_i\cap C_j|$,$n_{ij_i^*}$ 表示正确分到第 i 类的对象个数。

3.2 算法比较

本文将 k-prototype 聚类改进算法与 Liang k-prototype 算法进行外部评价指标(AC、PR、RE、ARI、NMI)、内部评价指标(CUM)及聚类结果的比较(见表 2)。

表 2　数据集在五种外部评价指标下的比较

数据集	二种算法	AC	PR	RE	ARI	NMI
Statlog	Liang k-prototype	0.555 6	0.545 7	0.500 0	0.008 3	0.005 9
	Proposed k-prototype	**0.588 9**	**0.579 7**	**0.575 0**	**0.027 4**	**0.017 8**
CHeart	Liang k-prototype	0.541 3	0.543 8	0.500 0	0.002 8	0.000 9
	Proposed k-prototype	**0.594 1**	**0.589 3**	**0.586 1**	**0.032 1**	**0.022 6**
Credit	Liang k-prototype	0.560 9	0.779 2	0.506 5	0.003 0	0.030 1
	Proposed k-prototype	**0.562 3**	**0.779 6**	**0.508 1**	**0.003 8**	**0.034 4**
GC	Liang k-prototype	0.700 0	0.656 0	0.500 0	0.052 6	0.012 5
	Proposed k-prototype	**0.700 0**	**0.652 8**	**0.500 0**	**0.053 5**	**0.012 9**
CMC	Liang k-prototype	0.429 7	0.471 5	0.357 6	0.008 4	0.038 3
	Proposed k-prototype	**0.438 6**	**0.471 4**	**0.370 1**	**0.021 7**	**0.041 1**

表 2 显示 Proposed k-prototype 聚类算法普遍比 Liang k-prototype 聚类算法在五种外部评价指标下的值高。AC、PR、RE、ARI 和 NMI 的值越大,聚类结果越接近于数据集的真实类划分,聚类结果越好。尤其在 Statlog 和 CHeart 数据集上,表现得更为明显。在 GC 和 Credit 数据集上,虽然有个别指标与后者相等或减少,但总体效果还好。这是因为 Liang k-prototype 算法只引入了信息熵,而我们新提出 k-prototype 聚类改进算法除了利用信息熵的特性还充分考虑了各个属性的权重。

由图 1 – 图 5 可知,5 个数据集上的有效性指标都随类个数的增加呈先增后减趋势,且极大值所对应的横坐标相同,即聚类个数相同。换句话说,用 proposedk-prototype 算法与 Liang k-prototype 算法确定的聚类个数相同,但 proposed k-prototype 算法比 Liang k-prototype 算法的 CUM 值高,值越高类内相似度越大,进一步说明 k-prototype 聚类改进算法的聚类结果更好。

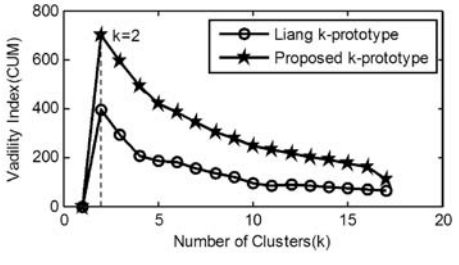

图 1　Statlog 数据集上有效性指标随类个数变化曲线图

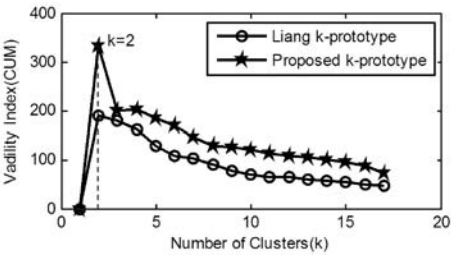

图 2　CHeart 数据集上有效性指标随类个数变化曲线图

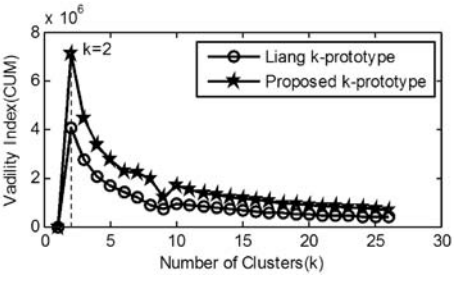

图 3　Credit 数据集上有效性指标随类个数变化曲线图

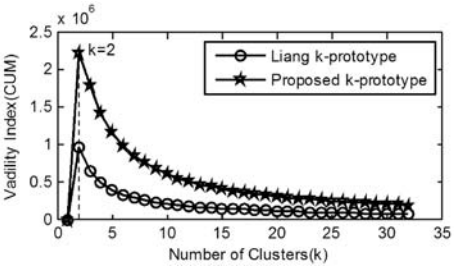

图 4　GC 数据集上有效性指标随类个数变化曲线图

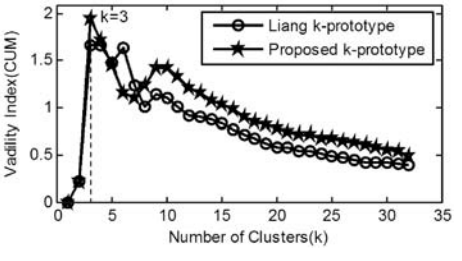

图 5　CMC 数据集上有效性指标随类个数变化曲线图

表 3 显示,5 个数据集在两种聚类算法上的聚类个数相同,且都与真实结果相同,说明新设计的算法正确。而且,k-prototype 聚类改进算法的外部有效性指标 CUM 明显比 Liang k-prototype 算法的 CUM 值高,几乎是后者的 2 倍,更验证了 k-prototype 聚类改进算法的有效性。

表 3　五个数据集在二种聚类算法的聚类结果比较

数据集	二种算法	CUM	聚类结果	真实结果
Statlog	Liang k-prototype	396.402 2	2	2
	proposed k-prototype	**698.586 5**	2	
CHeart	Liang k-prototype	190.449 0	2	2
	proposed k-prototype	**333.439 1**	2	
Credit	Liang k-prototype	4.070 8	2	2
	proposed k-prototype	**7.138 4**	2	

续表 3

数据集	二种算法	CUM	聚类结果	真实结果
GC	Liang k-prototype	0.962 4	2	2
	proposed k-prototype	**2.219 4**	2	
CMC	Liang k-prototype	1.665 9	3	3
	proposed k-prototype	**1.945 1**	3	

4 结 语

本文从信息熵和属性权重两方面对 Liang k-prototype 算法进行了改进,重新定义了相异性度量、缺失某类后的类内熵和(SBAE)和有效性指标(CUM),提出了带权的混合数据聚类个数确定算法,并用 5 个 UCI 数据集对该算法进行验证。实验表明 k-prototype 聚类改进算法在聚类结果相同并且正确的情况下(即确定了相同的聚类个数),6 种评价指标都有较高的值,值越高聚类效果越好,充分验证了该算法的有效性。

参 考 文 献

[1] Sun H, Wang S, Jiang Q. FCM-based model selection algorithms for determining the number of clusters[J]. Pattern Recognition, 2004, 37(10):2027 - 2037.

[2] Bezdek J C, Ehrlich R, Full W. FCM: the fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2):191 - 203.

[3] 陈黎飞, 姜青山, 王声瑞. 基于层次划分的最佳聚类数确定方法[J]. 软件学报, 2008, 19(1):62 - 72.

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

[5] Aghagolzadeh M, Soltanian-Zadeh H, Araabi B N, et al. Finding the number of clusters in a dataset using an information theoretic hierarchical algorithm[C]//IEEE International Conference on Electronics, Circuits and Systems. IEEE, 2006:1336 - 1339.

[6] Chen K, Liu L. The "Best K" for entropy-based categorical data clustering[C]//International Conference on Scientific and Statistical Database Management, SSDBM 2005, 27 - 29 June 2005, University of California, Santa Barbara, Ca, Usa, Proceedings. DBLP, 2005:253 - 262.

[7] Bai L, Liang J, Dang C. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data[J]. Knowledge-Based Systems, 2011, 24(6):785 - 795.

[8] Yan H, Chen K, Liu L, et al. Determining the best K for clustering transactional datasets: A coverage density-based approach[J]. Data & Knowledge Engineering, 2009, 68(1):28 - 48.

[9] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data[J]. Data & Knowledge Engineering, 2007, 63(2):503 - 527.

[10] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining & Knowledge Discovery, 1998, 2(3):283 - 304.

[11] Li C, Biswas G. Unsupervised learning with mixed numeric and nominal data[J]. Knowledge & Data Engineering IEEE Transactions on, 2002, 14(4):673 - 690.

[12] Hsu C C, Chen Y C. Mining of mixed data with application to catalog marketing[J]. Expert Systems with Applications, 2007, 32(1):12 - 23.

[13] 常茜茜, 张月琴. 一种基于划分的混合数据聚类算法[J]. 计算机应用与软件, 2014, 31(6):154 - 157.

[14] Liang J, Zhao X, Li D, et al. Determining the number of clusters using information entropy for mixed data[J]. Pattern Recognition, 2012, 45(6):2251 - 2265.

[15] Renyi A. On measures of information and entropy[J]. Maximum-Entropy and Bayesian Methods in Science and Engineering, 1961, 1(2):547 - 561.

[16] Parzen E. On estimation of a probability density function and mode[J]. Annals of Mathematical Statistics, 1962, 33(3):1065 - 1076.

[17] 丁祥武,谭佳,王梅. 一种分类数据聚类算法及其高效并行实现[J]. 计算机应用与软件,2017,34(7):249 - 256.

[18] Huang Z. Clusterin large data sets with mixed numeric and categorical values[C]//Proceeding of the First Pacific Asia Knowledge Discovery and Data Mining Conference. 1997: 21 - 34.

[19] Liang J, Chin K S, Dang C, et al. A new method for measuring uncertainty and fuzziness in rough set theory[J]. International Journal of General Systems,2002,31(4):331 - 342.

[20] Jenssen R, Eltoft T, Erdogmus D, et al. Some equivalences between kernel methods and information theoretic methods[J]. Journal of Vlsi Signal Processing Systems for Signal Image & Video Technology, 2006, 45(1 - 2):49 - 65.

[21] 赵兴旺,梁吉业. 一种基于信息熵的混合数据属性加权聚类算法[J]. 计算机研究与发展,2016,53(5):1018 - 1028.

[22] Bezdek J C. Pattern Recognition in Handbook of Fuzzy Computation[M]. Boston:IOP Publishing Ltd., 1998.

[23] Liang J, Bai L, Dang C, et al. The k-means-type algorithms versus imbalanced data distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4):728 - 745.

[24] Strehl A, Ghosh J. Cluster ensembles: A knowledge reuse framework for combining partitionings[J]. Journal of Machine Learning Research, 2002, 3(3):583 - 617.