结合缺失模式的不完整数据模糊聚类

郑奇斌[1] 刁兴春[2] 曹建军[2]
(解放军理工大学指挥信息系统学院 南京 210007)[1] (南京电讯技术研究所 南京 210007)[2]

摘 要 数据的完整性是数据可用性的重要维度。由于数据采集等过程中存在的问题,现实中的数据往往存在缺失。现有的聚类算法在面对不完整数据时一般采用忽略缺失或填补缺失的策略,但是当数据缺失属于非随机缺失时,这样的处理策略会导致聚类精度严重下降。当数据缺失属于非随机缺失时,数据缺失模式与缺失属性的取值相关,因此在不完整对象的相似度量中加入缺失模式相似的度量,提出了两种结合缺失模式的 PCM(Possibilistic c-means)模糊聚类算法:最小化缺失模式距离之和的 PatDistPCM 算法和基于缺失模式聚类的 PatCluPCM 算法。在两个公开数据集上的实验证明,考虑缺失模式的模糊聚类 PatDistPCM 和 PatCluPCM 算法,在对存在非随机缺失的数据进行聚类时,能有效提高聚类结果的准确性。

关键词 数据完整性,模糊聚类,非随机缺失,缺失模式,可能性 c-均值算法
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.011

Fuzzy Clustering Algorithm for Incomplete Data Considering Missing Pattern

ZHENG Qi-bin[1] DIAO Xing-chun[2] CAO Jian-jun[2]
(College of Command Information System, PLA University of Science and Technology, Nanjing 210007, China)[1]
(Nanjing Telecommunication Technology Institute, Nanjing 210007, China)[2]

Abstract Data integrality is an important metric for data availability. For the problems in data acquisition, datasets in real world are always incomplete. Missing data are usually ignored or imputed in common clustering algorithm. When data missing is missing not at random, ignorance or imputation will result poor clustering accuracy. Considering the relationship of the data missing pattern and the missing value, two PCM (Possibilistic c-means) clustering algorithms were proposed: PatDistPCM based on minimizing the sum of missing pattern distance and PatCluPCM based on missing pattern clustering. The experiments on public datasets show that the two proposed fuzzy clustering algorithms PatDistPCM and PatCluPCM can improve clustering precision and recall when clustering data are of missing not at random.

Keywords Data integrality, Fuzzy clustering, MNAR, Missing pattern, Possibilistic c-means

1 引言

聚类分析是一种重要的数据挖掘工具,在传感器数据分析、图像模式识别、安全、商务智能等领域有着广泛应用[1]。由于数据获取限制、数据理解错误或数据遗漏等原因,聚类分析中经常出现数据缺失的问题[2],例如设备故障、网络延迟等原因会造成传感器网络数据发生缺失,本文称这种含有缺失数据的数据集为不完整数据集。数据的不完整性使得聚类分析变得更加困难,传统的聚类方法难以应用于不完整数据的聚类。现有的不完整数据聚类算法主要基于数据填补或者局部距离,当数据缺失属于完全随机缺失(Missing Completely At Random, MCAR)和随机缺失(Missing At Random, MAR)时,这些算法的聚类结果较为准确;但是当数据缺失属于非随机缺失(Missing Not At Random, MNAR)时,这些算法会产生严重的偏差。

为提高对不完整数据进行聚类的精度,在 PCM 算法的基础上增加对缺失模式的考虑能够更加充分地利用不完整数据中的信息。在真实数据集上的实验证明,该方法在面对具有非随机缺失特征的不完整数据集上具有更高的精度。

本文第2节回顾了不完整数据聚类问题的相关研究;第3节介绍了基于局部距离的 PCM 算法;第4节提出了两种结合数据缺失模式的 PCM 算法,即 PatDist-PCM 和 PatClu-PCM;第5节通过实验对提出的算法进行检验;最后总结全文。

2 相关工作

对不完整数据进行聚类并非是一个新的问题,Dixon 早在1979年就提出了3种方法来处理模式识别中的缺失数据[3]:1)简单地删除含有缺失值的向量或特征;2)使用 k 近邻方法来填补缺失值;3)计算含有缺失值的向量之间的局部距离。多数情况下对不完整数据进行聚类均是基于上述3种方法。

Bezdek 提出的 FCM(Fuzzy c-Means)是经典的模糊聚类算法,它使用取值为0~1的隶属度来表示样本点属于一个类

到稿日期:2016-10-11 返修日期:2016-11-12 本文受国家自然科学基金(61371196)资助。
郑奇斌(1990-),男,博士生,主要研究方向为数据挖掘、数据质量,E-mail:zqb1990@hotmail.com;刁兴春(1967-),男,硕士,研究员,主要研究方向为数据工程;曹建军(1975-),男,博士,高级工程师,主要研究方向为数据工程。

别的程度，相对硬聚类算法而言可以更好地刻画现实中复杂事物的模糊边界[4]。FCM 算法本身并不能对不完整数据进行聚类，为了利用它对不完整数据进行聚类，Hathaway 提出了 4 种策略来处理 FCM 聚类算法中的不完整数据：1）全数据策略（Whole Data Strategy）：当数据的缺失比例较小时只对未完整的样本进行聚类，然后计算不完整记录与各簇中心的距离，并选择其中距离最小的簇作为该记录所属的簇，依此策略提出了 WDSFCM 算法；2）部分距离策略（Partial Distance Strategy）：使用局部距离而不是欧氏距离来度量样本间的距离，并修改了每次迭代中簇中心的更新方法，依此提出 PDSFCM 算法；3）最优完整策略（Optimal Complete Strategy）：在 FCM 的每次迭代中增加了对缺失数据的估计，在一次迭代中使用当前簇中心属性值的加权均值作为估计值来填充不完整样本的缺失属性，在下一次迭代中使用填充完整的样本进行 FCM 聚类，依此提出了 OCSFCM 算法；4）最近原型策略（Nearest Prototype Strategy）：在 FCM 的每次迭代中增加对缺失数据的估计，但不使用各簇中心的属性均值而是使用距离缺失样本最近簇的中心属性值来对缺失样本进行填充，依此提出 NPSFCM 算法[5]。Balkis 在Hathaway的 OCSFCM 算法的基础上提出了 OCS-FSOM 和 Multi-OCSFSOM 算法，使用模糊自组织图（Fuzzy Self Organising Map）来处理不完整数据，并利用多层 OCS-FSOM 确定簇数目[6]。

鉴于 FCM 算法对异常值的敏感性，Krishnapuram 于 1993 年在 FCM 的基础上提出了 PCM 算法，该算法放松了 FCM 样本点对各类隶属度之和的约束，并引入惩罚项来避免无意义的平凡解[7]。类似于 FCM 算法，PCM 算法也不支持对不完整数据的聚类。Zhang 等人为了解决不完整分布式传感器网络数据的聚类问题，基于 PCM 算法提出了 WPCM 算法，该算法使用局部距离来衡量不完整数据和簇中心的距离，在聚类过程中对不完整的样本点分配较低的权重来降低其对聚类的影响，并基于 Map-Reduce 架构实现了其分布式版本的算法 DWPCM[8]。

包括 Hathaway 提出的 4 种策略、Zhang 提出的 WPCM 以及 Balkis 提出的 OCS_FCM 等在内的不完整数据聚类算法，都只利用了不完整数据中的完整部分信息，对关系型数据而言就是记录中可以观测的部分属性，而忽略了蕴含在数据缺失模式中关于缺失如何产生的信息。Rubin 提出对缺失数据的分析处理应该结合两个过程：数据产生过程以及缺失产生过程，而很多对不完整数据的分析中只关注了数据自身而忽略了缺失的产生过程[9]。为了描述缺失与数据集中真实值的关系，Rubin 提出数据缺失机制来描述缺失的产生过程，并将缺失机制分为完全随机缺失、随机缺失以及非随机缺失[10]。设不完整的数据集 $X=\{X_o,X_m\}$，其中 X_o 为数据集中可观测的数据，X_m 为缺失的不可观测数据。

$$P(M|X,\Phi)=P(M|\Phi) \tag{1}$$

其中，M 为缺失指示矩阵，$M_{ij}=0$ 当且仅当 x_{ij} 缺失，否则 $M_{ij}=1$，Φ 为其他参数。由式(1)可知，若数据出现缺失的概率与数据集中的任何值都无关，则数据缺失属于完全随机缺失。

$$P(M|X,\Phi)=P(M|X_o,\Phi) \tag{2}$$

由式(2)可知，若数据出现缺失的概率与数据集中的可观测相关而与缺失值自身无关，则数据缺失属于随机缺失。

$$P(M|X,\Phi)=P(M|X_m,\Phi) \tag{3}$$

$$P(M|X,\Phi)=P(M|X_o,X_m,\Phi) \tag{4}$$

由式(3)或式(4)可知，若数据出现缺失的概率与缺失值自身相关，则数据缺失属于随机缺失。

3 种缺失中完全随机缺失的条件最为苛刻，且大都是由于系统缺陷造成的，其随着软/硬件基础设施的日趋完善已经较为少见，现实科研和工程中出现更多的是随机缺失和非随机缺失。完全随机缺失和随机缺失属于可忽略缺失，而非随机缺失属于不可忽略缺失[9,11]。常见的不完整数据聚类（如 Hathaway 提出的 4 种策略）都基于数据缺失机制，从而可忽略假设，当假设不满足时会导致聚类精度下降。如果数据缺失机制不可忽略，应当对数据缺失过程进行建模，然后结合模型对存在缺失的数据进行分析。Marlin 在基于协同过滤的推荐系统中考虑到了评分数据中存在的不可忽略缺失，基于对评分数据的先验知识提出了结合数据缺失机制的协同过滤推荐模型[12-13]。

但在很多情况下，由于缺乏相关缺失过程的知识，建立正确的缺失过程模型十分困难，而数据缺失机制的直接体现——数据缺失模式却能被较容易获得。由于具有相同缺失模式的数据更有可能来自相同群体，具有不同缺失模式的数据更有可能来自不同群体[14]，为了提高不完整数据的聚类精度，本文提出了一种结合数据缺失模式的 PCM 聚类。

3　基于局部距离的模糊聚类算法

3.1　模糊聚类算法——PCM 及 ExtendedPCM

模糊聚类方法是一类基于划分的聚类方法，但是不同于一般的划分聚类方法，其在模糊聚类中采用非互斥的划分，即同一个对象可以同时属于多个簇，并使用隶属度来描述对象属于一个簇的强度。在不完整数据的聚类分析中，簇间的边界更加模糊，模糊聚类方法相对硬聚类方法能够更为准确地刻画数据中的群组结构。

Bezdek 提出的 FCM 算法是经典的模糊聚类算法，该算法约定每个聚类对象的隶属度之和为 1，因此可将 FCM 算法中的隶属度理解为对象属于簇的概率。该约定是为了避免产生隶属度都为 0 的平凡解，但是也带来了 FCM 的异常值敏感特性。

为了避免 FCM 算法对异常值的敏感问题，Krishnapuram 在 FCM 的基础上提出了 PCM（Possibilistic c-Means）算法，该算法弱化了 FCM 对聚类对象的隶属度约束，并加入了惩罚项来避免平凡解[7]。

设需要把 n 个对象 $X=\{x_1,x_2,\cdots,x_n\}$ 通过聚类划分为 l 个簇 $C=\{c_1,c_2,\cdots,c_l\}$，FCM 算法定义一个 $l\times n$ 隶属度矩阵 $U=\{u_{ij}\}$，其中 u_{ij} 表示对象 x_j 属于簇 c_i 的隶属度；d_{ij} 表示对象 x_j 到簇 c_i 的中心 v_i 的距离。PCM 算法将聚类转换为一个有约束的最优化问题：

$$\min\{J_m(U,V)=\sum_{i=1}^{l}\sum_{j=1}^{n}u_{ij}^m d_{ij}^2+\sum_{i=1}^{l}\eta_i\sum_{j=1}^{n}(1-u_{ij})^m\} \tag{5}$$

其中，所有 u_{ij} 需要满足式(5)中的约束条件为：

$$\begin{aligned}&u_{ij}\in[0,1]\\&0<\sum_{j=1}^{n}u_{ij}\leqslant n,\forall i\\&\max(u_{ij})>0,\forall j\end{aligned} \tag{6}$$

其中，$m>1$ 为模糊常数，η_i 为簇 c_i 内隶属度为 0.5 的点（即 3dB 点）距簇中心的距离。一般而言，η_i 的取值与簇的形状和大小相关，计算公式如式(7)所示：

$$\eta_i=\frac{\sum_{j=1}^{n}u_{ij}^m\cdot d_{ij}^2}{\sum_{j=1}^{n}u_{ij}^m} \tag{7}$$

使式(5)中的 J 对 U 求偏微分并令其等于 0，可以得到 J 取得极值的必要条件式(8)，显然式(8)是满足条件式(6)的。而且任意对象的隶属度只与其距簇中心的距离相关，因此在每次迭代中可以使用式(8)来更新隶属度矩阵。

$$u_{ij}=\frac{1}{1+(d_{ij}^2/\eta_i)^{1/(m-1)}} \tag{8}$$

簇中心可以通过对象在当前隶属度下的期望值来计算，在每次迭代中按照式(9)来更新每个簇中心 v_i。

$$\vec{v_i}=\frac{\sum_{j=1}^{n}u_{ij}^m\cdot\vec{x_j}}{\sum_{j=1}^{n}u_{ij}^m} \tag{9}$$

给定初始的隶属度矩阵、初始簇中心等参数，使用类似 EM 算法的迭代方法：通过原来的隶属度计算新的簇中心，然后使用新的簇中心计算新的隶属度，不断重复该过程直到目标函数值基本不变（即变化小于用户给定的阈值 ε）。PCM 聚类算法如算法 1 所示。

算法 1 PCM 算法

Step1 选择合适的参数 m,C,ε，并初始化隶属度矩阵 U；

Step2 根据式(9)计算簇中心；

Step3 根据式(7)估计 η_i；

Step4 根据式(8)更新隶属度矩阵 U；

Step5 若 $\|u-u^{old}\|^2\leqslant\varepsilon$ 则结束，否则返回 Step2。

PCM 算法存在一致性聚类的问题，即当簇间边界不明显时，容易在迭代过程中使得簇中心不断靠近，从而最终将所有对象聚为同一个簇。Pal 在 PCM 算法的基础上提出改进的 PCM 算法，这里称其为 ExtendedPCM，即在目标函数式(1)中加入另外一个惩罚项，在两个聚类中心靠近时该项的取值迅速变大，因此可以避免一致性聚类问题。加入了惩罚项之后的新目标函数为[15]：

$$\min\{J_m(U,V)=\sum_{i=1}^{l}\sum_{j=1}^{n}u_{ij}^m d_{ij}^2+\sum_{i=1}^{l}\eta_i\sum_{j=1}^{n}(1-u_{ij})^m+\sum_{i=1}^{l}\gamma_i\sum_{k=1,k\neq i}^{l}\frac{1}{d^2(v_i,v_k)}\} \tag{10}$$

其中，γ_i 为簇 c_i 中惩罚项的权重。由于新加入的项与隶属度 u_{ij} 无关，因此隶属度的计算公式同式(8)，而簇中心的计算公式如式(11)所示：

$$\vec{v_i}=\frac{\sum_{j=1}^{n}u_{ij}^m\vec{x_j}-\gamma_i\sum_{k=1,k\neq i}^{l}d(i,k)\vec{v_k}}{\sum_{j=1}^{n}u_{ij}^m-\gamma_i\sum_{k=1,k\neq i}^{l}d(i,k)} \tag{11}$$

ExtendedPCM 聚类算法的其他部分与原 PCM 算法相同，该算法在加入新的惩罚项之后可以对 iris 和 wine 等簇边界模糊的数据集进行聚类，从而得到良好的聚类结果。

3.2 基于局部距离的 PCM 聚类算法——PDS-ExtendedPCM

包括 FCM 和 PCM 算法在内的大部分基于划分的模糊聚类算法都必须首先确定每个簇的中心，然后根据对象 x_j 到簇中心的距离来计算对象的隶属度。常见的欧氏距离、马氏距离等距离度量方法都需要对象的每个维度取值参与计算。而不完整数据中对象的维度不全，从而无法计算隶属度，因此 FCM 和 PCM 算法都不能直接用于不完整数据的聚类。Hathaway 在 PDSFCM 算法中提出使用局部距离来代替一般的距离度量。以欧氏距离为例，维度为 D 的聚类对象 x_k 到簇 i 中心 v_i 的距离为：

$$d_{ij}=PD_{ij}=\frac{D}{I_j}\sqrt{\sum_{l=1}^{D}(x_{kl}-v_{il})^2 I_{jl}}$$
$$I_{jl}=\begin{cases}0, & \text{if } x_{kl} \text{ is missing}\\ 1, & \text{else}\end{cases} \tag{12}$$
$$I_j=\sum_{l=1}^{D}I_{jl}$$

其中，D 表示聚类对象总的属性数量，I_{jl} 表示聚类对象 x_k 的第 l 个属性是否缺失。Hathaway 提出的局部距离充分利用了不完整对象中的可观测信息，可以较为客观地反映聚类对象到簇中心的距离。引入局部距离后，簇中心的更新方式为式(13)，其中 v_{il} 为簇中心 v_i 第 l 个维度的取值。

$$v_{il}=\frac{\sum_{j=1}^{N}u_{ij}^m I_{jl}x_{jl}}{\sum_{j=1}^{N}u_{ij}^m I_{jl}} \tag{13}$$

类似于式(13)，将局部距离运用在 ExtendedPCM 中，簇中心的更新公式如式(14)所示：

$$\vec{v_i}=\frac{\sum_{j=1}^{n}u_{ij}^m I_{jl}\vec{x_j}-\gamma_i\sum_{k=1,k\neq i}^{l}d(i,k)\vec{v_k}}{\sum_{j=1}^{n}u_{ij}^m I_{jl}-\gamma_i\sum_{k=1,k\neq i}^{l}d(i,k)} \tag{14}$$

保持 PCM 算法的其他部分不变，基于局部距离的不完整数据聚类算法 PDS-ExtendedPCM 如算法 2 所示。

算法 2 PDS-ExtendedPCM 算法

Step1 选择合适的参数 m,C,ε，并初始化隶属度矩阵 U；

Step2 根据式(14)计算簇中心；

Step3 根据式(7)估计 η_i；

Step4 根据式(8)更新隶属度矩阵 U；

Step5 若 $\|u-u^{old}\|^2\leqslant\varepsilon$ 则结束，否则返回 Step2。

4 结合数据缺失模式的 PCM 聚类算法

引入局部距离可以使 PCM 聚类支持不完整数据，但当数据缺失属于非随机缺失时，局部距离会忽略缺失数据的特殊性从而得到有偏的距离度量。基于有偏的距离计算隶属度所得到的聚类结果是不准确的。

表 1 不完整数据

ID	Age	Male	*Income*/W	*Weight*/kg
1	28	Man	9	Miss(80)
2	35	Woman	Miss(15)	65
3	29	Man	Miss(20)	70
4	33	Woman	7	Miss(90)

表 1 所列的数据存在非随机缺失：$P(Income=\text{Miss}\mid \text{Income}>10\text{W})=0.9$，$P(Weight=\text{Miss}\mid Weight>75\text{kg})=0.8$，表中 Miss 表示属性值缺失，括号中的数据为其真实值。

设置簇数量为 2，使用基于局部距离的聚类算法对表中的不完整数据进行聚类，再使用同样的聚类算法对完整数据进行聚类，原本应该被划分为同一簇的对象 1 和 4 以及对象 2 和 3 被基于局部距离的聚类算法划分到了不同簇中。局部距离策略的失效，是由于非随机缺失的存在导致了缺失数据的有偏性，相同属性出现缺失的对象在该属性上的真实值具有一

定的相同特性（如示例中的 *Income*＞10W 以及 *Weight*＞75kg）。

4.1 数据缺失模式

为了提高对非随机缺失数据的聚类精度，在聚类算法中应结合数据的缺失机制。但是数据缺失机制通常是难以获取或检验的，一般只能基于先验知识提出假设缺失机制，且这种先验知识的获取也不易，因此在大部分情况下都难以利用缺失机制来辅助聚类分析。

数据缺失模式描述了数据集中哪些数据是缺失的，哪些数据是可观测的，更深层的意义在于缺失模式反映数据缺失机制[9]。表 1 中数据的缺失模式如表 2 所列，其中 0 表示数据值可观测，1 表示数据值缺失。

表 2　缺失模式

ID	Age	Male	*Income*/W	*Weight*/kg
1	1	1	1	0
2	1	1	0	1
3	1	1	0	1
4	1	1	1	0

由表 2 可知，属于同一类别的 1 和 4 以及 2 和 3 具有相同的缺失模式。这也印证了 Wang[14] 的观点：具有相同缺失模式的数据更有可能来自相同群体，具有不同缺失模式的数据更有可能来自不同群体。在聚类分析中加入缺失模式作为辅助信息，可以提高不完整数据的聚类准确性。为了进一步分析缺失模式对不完整数据聚类的影响，下文讨论完全随机缺失、随机缺失、非随机缺失 3 种缺失情况下的缺失模式。

1）当数据的缺失机制属于完全随机缺失时，由于数据的缺失与其真实取值无任何关系，因此缺失模式不能体现任何关于数据本身的信息。

2）当数据的缺失机制属于随机缺失时，缺失属性与非缺失属性的取值相关，缺失模式的相似性表明了非缺失属性取值的相似性。而局部距离本身就是对未缺失属性相似性的度量，因此缺失模式体现的信息与局部距离体现的信息一致。

3）当数据缺失机制属于非随机缺失时，缺失属性与缺失属性自身的取值相关，缺失模式的相似性表明了缺失属性取值的相似性。而局部距离不具备对这种相似性的表达能力，因此缺失模式的相似性是对局部距离的补充。

在随机缺失和非随机缺失的数据中，缺失模式可以体现某些数据本身的特征，可以用来辅助不完整数据的聚类；在完全随机缺失中缺失模式与数据自身无关系，缺失机制会对聚类分析产生干扰。

缺失模式作为辅助信息的基本思想：使每个簇内对象的缺失模式尽量相似。因此利用方式分为两种：1）直接计算聚类对象两两之间的模式距离，使得每个簇内的模式距离最短；2）在每个簇中确立一个缺失模式的概貌，以使每个簇中对象的缺失模式与概貌相似。基于这两种方式提出以下两种结合缺失模式的基于局部距离的 PCM 聚类算法，即 PatDistPCM 和 PatCluPCM。

4.2 最小化簇内缺失模式距离的 PCM 聚类算法——PatDistPCM

如 4.1 节所述，具有相似缺失模式的对象更有可能属于同一类别，在聚类时应该使簇内的模式差距尽可能小，因此考虑在经典 PCM 算法的优化目标函数中加入能够代表簇内对象缺失模式差异的项。最为直观的方法是计算簇 i 内的对象之间的模式差异之和。由于缺失模式是一个 0-1 向量，因此使用海明距离（Hamming distance）来度量模式间的差异。两种 d 维的缺失模式 p_1，p_2 间的海明距离定义为两个向量中不同分量的个数：

$$h(p_1,p_2)=\sum_{c}^{d}|p_{1l}-p_{2l}| \tag{15}$$

簇 i 内的所有模式差异之和为：

$$P_i=\sum_{j=1}^{N}\sum_{k=1}^{N}u_{ij}^{m}u_{ik}^{m}h_{jk} \tag{16}$$

其中，h_{jk} 为对象 x_i 和 x_j 缺失模式的海明距离。将式（16）加入到 PCM 的优化目标函数式（10）中，可以得到新的目标函数：

$$\min\{\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}d_{ij}^{2}+\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N}u_{ij}^{m}u_{ik}^{m}h_{jk}+\sum_{i=1}^{C}\eta_i\cdot\sum_{j=1}^{N}(1-u_{ij})^{m}\} \tag{17}$$

在约束条件式（6）下求解目标函数（17）是一个约束非线性规划问题。使式（17）对 u_{ij} 求偏微分并令其为 0，可得：

$$u_{ij}=\frac{1}{1+\left[\dfrac{d_{ij}^{2}+\sum_{k=1}^{N}u_{ik}^{m}(h_{jk}+h_{kj})}{\eta_i}\right]^{\frac{1}{m-1}}} \tag{18}$$

显然式（18）满足约束条件式（6），并且每个对象对簇 i 的隶属度 u_{ij} 仅与簇 i 内的其他对象相关，将簇 i 中所有隶属度联立起来得到一个 N 元 m 次方程组。使用 Newton-Raphson 迭代法可以得到此方程组的解，即使得目标函数（17）最小的解，而其他参数如 η_i、簇中心等的计算方法不变。综上，基于缺失模式差异之和的聚类算法——PatDistPCM 如算法 3 所示。

算法 3　PatDistPCM 算法

Step1　选择合适的参数 m，C，ε，并初始化隶属度矩阵 U；

Step2　根据式（14）计算簇中心；

Step3　根据式（7）估计 η_i；

Step4　根据式（18）更新隶属度矩阵 U；

Step5　若 $\|u-u^{old}\|^2\leqslant\varepsilon$ 则结束，否则返回 Step2。

4.3 基于数据缺失模式聚类的 PCM 聚类算法——PatCluPCM

4.2 节中计算簇内所有对象间缺失模式差异之和的方式虽然能够准确计量簇内模式差异，但是计算过于复杂，其时间复杂度随着数据集中对象数量的增长而急剧增长，因此难以应用于实际。

为了提高计算效率，在每个簇中建立另外一个簇中心来代表该簇中缺失模式的概貌（称为模式中心，记为 O），以所有簇中对象缺失模式到该中心的距离表示簇中缺失模式的差异。由于缺失模式是一组 0-1 向量，因此同样使用 0-1 向量来表示模式中心，用海明距离表示对象缺失模式到模式中心的距离。这种做法相当于对缺失模式进行聚类，因此称该算法是基于模式聚类的。通过模式聚类，簇中对象到簇的模式中心的距离之和即可反映簇内对象的缺失模式差异。

设簇 i 的模式中心为 o_i，则对象 x_j 的缺失模式 p_j 与模式中心 o_i 的距离 d'_{ij} 为：

$$d'_{ij}=h(o_i,p_j)=\sum_{l}^{d}|o_{il}-p_{jl}| \tag{19}$$

新的目标函数为：

$$\min\{\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}(d_{ij}^{2}+w_j d_{ij}'^{2})+\sum_{i=1}^{C}\eta_i\sum_{j=1}^{N}(1-u_{ij})^{m}+\sum_{i=1}^{L}\gamma_i\sum_{k=1,k\neq i}^{L}\frac{1}{d^{2}(v_i,v_k)}\} \tag{20}$$

为了避免对象缺失模式到模式中心的距离过度影响聚类过程，为其分配动态权重 w_j：对象 x_j 缺失的程度越高，则权重 w_j 越大；对象 x_j 缺失的程度越低，则权重 w_j 越小。图 1 给出了具有此特性的函数，因此使用式(21)来计算权重。

$$w_j=\frac{w}{e^{\psi(x_j)\psi(x_j)}} \tag{21}$$

其中，w 为全局权重，为对象 x_j 的完整度，可以通过完整属性占所有属性的比值(即式(22))来计算。

$$\psi(x_j)=\frac{\sum_{i=1}^{d}(1-p_{ji})}{|D|} \tag{22}$$

其中，D 为数据集属性集。

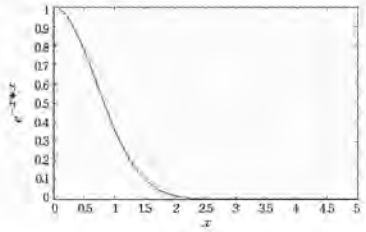

图 1 权重函数特性

为了使目标函数式(20)最小，对其求 u_{ij} 的偏微分并令其为 0，解得：

$$u_{ij}=\frac{1}{1+[(d_{ij}^{2}+w_j d_{ij}'^{2})/\eta_i]^{1/(m-1)}} \tag{23}$$

簇中心的计算方法同式(14)。计算簇内所有对象缺失模式的加权均值作为模式中心。由于需要保持其为 0-1 向量，因此按照四舍五入的原则对其取整，则簇的模式中心计算公式如式(24)所示：

$$o_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}p_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{24}$$

由于 η_i 会影响模式聚类，因此按照式(25)估计 η_i：

$$\eta_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}(d_{ij}^{2}+w_i d_{ij}'^{2})}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{25}$$

综上，基于模式聚类的 PCM 聚类算法——PatCluPCM 如算法 4 所示。

算法 4 PatCluPCM 算法

Step1 选择合适的参数 m，C，ε，并初始化隶属度矩阵 U；

Step2 根据式(14)计算簇中心；

Step3 根据式(24)计算簇模式中心；

Step4 根据式(25)估计 η_i；

Step5 根据式(23)更新隶属度矩阵 U；

Step6 若 $\|u-u^{old}\|^2\leq\varepsilon$，则结束，否则返回 Step2。

5 实验结果与分析

5.1 实验设置与数据集

为了检验上述方法的可行性与有效性，使用公开数据进行实验。

实验数据集一为 Iris 数据集，该数据集有 4 个特征：sepal length 和 sepal width，petal length 和 petal width，第 5 个属性为类标。数据集中共有 3 类数据：Iris Setosa，Iris Versicolour 和Iris Virginica。数据集本身不存在缺失，人为地在其中置入非随机缺失。Iris 数据集缺失机制如表 3 所列。

表 3 Iris 数据集缺失机制

序号	缺失机制
1	$p(sepal\ length=\text{Miss}\mid sepal\ length>6)=0.95$
2	$p(sepal\ width=\text{Miss}\mid sepal\ width>3.2)=0.9$
3	$p(petal\ length=\text{Miss}\mid petal\ length<3.5)=0.95$
4	$p(petal\ width=\text{Miss}\mid petal\ width>1.5)=0.95$

实验数据集二为 WineQuality 数据集，该数据集有 12 个特征：fixed acidity，volatile acidity，citric acid，residual sugar，chlorides，free sulfur dioxide，total sulfur dioxide，density，pH，sulphates，alcohol，quality。其中 quality 为输出变量，表示葡萄酒的质量，取 0～10 之间的实数。人为地对其置入非随机缺失。Wine Quality 数据集缺失机制如表 4 所列。

表 4 Wine Quality 数据集缺失机制

序号	缺失机制
1	$p(sepal\ length=\text{Miss}\mid sepal\ length>6)=0.95$
2	$p(sepal\ width=\text{Miss}\mid sepal\ width>3.2)=0.9$
3	$p(petal\ length=\text{Miss}\mid petal\ length<3.5)=0.95$
4	$p(petal\ width=\text{Miss}\mid petal\ width>1.5)=0.95$

5.2 实验方法及评价指标

为了验证算法对不完整数据的聚类效果，实验分别使用基于局部距离策略的 FCM 算法——PDS-FCM、基于局部距离策略的 PCM 算法——PDS-PCM，以及本文提出的PatDist-PCM 和 PatCluPCM 算法对数据集进行聚类，并对比 4 种算法的准确率、召回率以及 F1 指标。

实验主要评估聚类的质量。根据是否有基准可用，可将评估方法分为外在方法(extrinsic method)和内在方法(intrinsic method)，外在方法通过比较聚类结果和基准来评估聚类质量，内在方法一般通过考虑簇的分离情况等指标来评估聚类的质量[1]。实验中使用的数据集自身是带有类标的，因此采用外在方法来评估聚类的质量。Bagga 和 Baldwin 提出的 BCubed 准确率、BCubed 召回率是常用的聚类外部评价方法[15]，一个对象的 BCubed 准确率表示同一簇中有多少个其他对象与该对象同属一个类别，一个对象的 BCubed 召回率反映了同一类别的对象被分配到相同的簇中的数量。这两个指标值越高，说明聚类的质量越高。

5.3 实验结果

实验结果如表 5 所列。

表 5 实验结果

数据集	算法	准确率	召回率	F1
Iris	PDS_FCM	0.6472	0.6729	0.6598
	PDS-ExtendedPCM	0.6582	0.4389	0.5266
	PatDistPCM	0.7850	0.7955	0.7902
	PatCluPCM	0.7804	0.7907	0.7855
WineQuality	PDS-FCM	0.7565	0.4027	0.5256
	PDS-ExtendedPCM	0.7531	0.8032	0.7773
	PatDistPCM	0.7740	0.8701	0.8129
	PatCluPCM	0.7595	0.8724	0.8120

由表 5 可知，当数据缺失属于非随机缺失时，结合缺失模

式的聚类算法 PatDistPCM 和 PatCluPCM 可以有效减小数据的偏离性对聚类结果的影响，提高聚类的准确程度。而在准确性指标几乎相同的情况下，PatCluPCM 算法的效率远高于 PatDistPCM 算法，因此在数据量较大的情况下 PatClu-PCM 算法更适合于非随机缺失数据的聚类。

结束语　数据缺失模式是数据缺失机制的直接反映。本文提出的两种模糊聚类算法 PatDistPCM 和 PatCluPCM 分别通过最小化簇内缺失模式距离之和以及对缺失模式聚类两种方式，利用缺失模式中隐藏的缺失机制信息对不完整数据进行聚类，当数据缺失属于非随机缺失时这部分信息对聚类的结果会产生重大影响。实验结果表明，通过引入缺失模式能够提高 PCM 算法对非随机缺失数据的聚类精度。

参 考 文 献

[1] HAN J W, KAMBER M, PEI J. Data Mining: Concepts and Techniques(3rd ed)[M]. Morgan Kaufmann Publishers, 2011: 288-293.

[2] GU Y, YU G, LI X J, et al. RFID data interpolation algorithm based on dynamic probabilistic path-event model[J]. Journal of Software, 2010, 21(3): 438-451.

[3] DIXON J K. Pattern recognition with partly missing data[J]. IEEE Transactions on Systems, Man and Cybernetics, 1979, 9(10): 617-621.

[4] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. Plenum Press, 1981.

[5] HATHAWAY R J, BEZDEK J C. Fuzzy c-Means Clustering of Incomplete Data[J]. IEEE Transactions on System, Man, and Cybernetics, 2001, 31(5): 735-744.

[6] BALKIS A, YAHIA S B. A new algorithm for fuzzy clustering handling incomplete dataset[J]. International Journal on Artificial Intelligence Tools, 2014, 23(4): 1460012.

[7] KRISHNAPURAM R, KELLER J M. A possibilistic Approach to clustering[J]. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98-110.

[8] ZHANG Q, CHEN Z. A distributed weighted Possibilistic c-Means algorithm for clustering incomplete big sensor data[J]. International Journal of Distributed Sensor Networks, 2014, 2014(2): 4.

[9] LITEEL R J A, RUBIN D B. Statistical Analysis with Missing Data[M]. John Wiley & Sons, Inc. New Jersey, 2002.

[10] Donald D B. Inference and Missing Data[J]. Biometrika, 1976, 63(3): 581-592.

[11] ALLISON P D. 数据缺失[M]. 林毓玲，译. 上海：格致出版社，2012.

[12] MARLIN B M. Missing Data Problems in Machine Learning [D]. Toronto: University of Toronto, 2008.

[13] MARLIN B M, ZEMEL R S. Collaborative Prediction and Ranking with Non-Random Missing Data [C] // RecSys '09. New York, USA, 2009: 23-25.

[14] WANG H, WANG S. Discovering patterns of missing data in servey databases: An application of rough sets[J]. Expert Systems with Applications, 2009, 36(3): 6256-6260.

[15] TIMM H, BORGELT C, KRUSE R. An Extension of Possibilistic Fuzzy Cluster Analysis[J]. Fuzzy Sets and Systems, 2004, 147(1): 3-16.

[16] BAGGA A, BALDWIN B. Entity-based cross-document coreferencing using the vector space model[C] // Proc. 1998 Annual Meeting of the Association for Computational Linguistics and Int. Conf. Computational Linguistics (COLING-ACL '98). Montreal, Quebec, Canada, 1998.

(上接第 51 页)

[6] ZHOU Z H, ZHANG M L. A Review on Multi-Label Learning Algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[7] GIBAJA E. A Tutorial on Multi-Label Learning [J]. ACM Computing Surveys, 2015, 47(3): 1-38.

[8] ZHANG M L, ZHOU Z H. M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning [J]. Eighth IEEE International Conference on Data Mining, 2008, 176(1): 688-697.

[9] ZHOU Z H, ZHANG M L, HUANG S J. Multi-instance Multi-label Learning [J]. Artificial Intelligence, 2008, 176(1): 2291-2320.

[10] NGUYEN C T, WANG X L, LIU J. Labeling complicated objects: Multi-view multi-instance multi-label learning[C] // The 28th AAAI Conference on Artificial Intelligence Quebec City. Canada: AAAI Press, 2014: 2013-2019.

[11] NGUYEN C T, ZHAN D C, ZHOU Z H. Multi-modal image annotation with Multi-instance Multi-label LDA[C] // International Joint Conference on Artificial Intelligence. Beijing: IJCAI, 2013: 1558-1564.

[12] ZHOU Z H, ZHANG M L, HUANG S J, et al. MIML: A Framework for Learning with Ambiguous Objects [J]. Corr Abs, 2008, 3231(1): 2012.

[13] BOUTELL M R, LUO J B, SHEN X P. Learning Multi-label Scene Classification [J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[14] KNAUER C, LÖFFLER M, SCHERFENBERG M. The directed Hausdorff distance between imprecise point sets [J]. Theoretical Computer Science, 2011, 412(32): 4173-4186.

[15] LIBERTI L, LAVOR C, MACULAN N. Euclidean distance geometry and applications [J]. Siam Review, 2012, 56(1): 3-69.

[16] KIM S, CHOI J. An SVM-based high-quality article classifier for systematic reviews [J]. Journal of Biomedical Informatics, 2014, 47(2): 153-159.

[17] ŽALIK K R, ŽALIK B. Validity Index for Clusters of Different Sizes and Densities [J]. Pattern Recognition Letters, 2011, 32(2): 221-234.

[18] HONG T P, LIN C W, YANG K T, et al. Using TF-IDF to Hide Sensitive Itemsets [J]. Applied Intelligence, 2013, 38(4): 502-510.