维普资讯 http://www.cqvip.com

基于数据场改进的PAM聚类算法*)

余建桥 张 帆
(西南农业大学信息学院 重庆400716)

摘 要 PAM是基于k-中心值聚类的一种算法,在处理数据集的聚类问题时,具有良好的准确性和伸缩性。但PAM算法在随机选取初始中心点时存在不足,而且在处理存在孤立点或噪声的数据时算法不是很健壮。本文针对这两点不足,使用了数据场的概念对PAM聚类算法进行了有益的改进,提高了算法的准确性和处理孤立点或噪声的能力,使其更适合于对数据集的处理,提高了挖掘结果的质量。

关键词 PAM,数据场,聚类

An PAM Algorithm Based on Data Field

YU Jian-Qiao ZHANG Fan
(Information College, Southwest Agricultural University, Chongqing 400716)

Abstract PAM is a k-mediod algorithm. It is efficient and flexible to handle data sets. However, there are inaccurate and insufficient in random choosing initial data center in algorithm PAM, and PAM is not robust to treat data sets including outlier or noise. Based on data field conception, an improved PAM is designed to solve these problems. This improved algorithm is more efficient and improves quality of result of data mining.

Keywords PAM, Data field, Clustering

1 引言

聚类分析是将物理或抽象对象的集合分组成为由类似的对象组成的多个簇的过程[1]。由聚类所生成的簇(Cluster)是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象最大程度的相异。聚类分析中常用的一种算法是基于k-中心点(k-Medoid)划分方法的PAM[2](Partitioning around Medoid)算法。然而,PAM算法中的开始时随机选取初始中心点影响到PAM算法的复杂性和有效性。一般情况下,在随机选取初始中心点时很难保证在大量的、具有不确定性的数据中选出的是合理的中心点,如果选取的初始中心点与实际情况偏离很大,那就会增加算法的复杂性,降低算法的有效性。而且如果当被挖掘的数据中存在孤立点数据和"噪声"时,PAM算法对处理这类数据就显得不是那么健壮。

空间信息模型通常可以分两大类:场(Filed)模型和对象(Object),而场模型通常可用于具有连续的空间变化趋势的情况的建模,这样便把场的概念推广到空间数据挖掘领域。认为数据对象集中的每个对象都相当于一个辐射源,其周围存在一个辐射的作用场,在整个数域空间中就形成了以数据为辐射中心的叠加的数据场。将数据场的观念引入到PAM算法中,来对数据集进行有效的初始中心点的选取和进行孤立点处理。

2 PAM算法的分析

PAM算法是一种基于划分方法的聚类算法,也是最早提出的较为经典的k-中心算法之一。

2.1 PAM聚类算法

PAM算法的主要思想是试图对n个对象给出k个划分,最初随机选择k个初始中心点后,该算法反复地试图找出更好的中心点。所有可能的对象对被分析,每个对中的一个对象被看作是中心点,而另一个不是。对可能的各种组合,估算聚类结果的质量。一个对象O_j被可能产生最大平方-误差值减少的对象代替。在一次迭代中产生的最佳对象的集合成为下次迭代的中心点。当n和k的值较大时,这样的计算代价相当高。

PAM随机抽取数据集合的k个初始中心点,迭代复杂度是$O(nkt)$,其中n是对象的数目,k是簇的数目,而t是迭代的次数。PAM算法的复杂性和有效性与k个初始中心点的选取有很大关系。

2.2 PAM算法的过程

1. 随机选取k个对象作为初始的中心点对象O_j;

2. 指派每个剩余的对象给离它最近的中心点所代表的簇;

3. 随机选择一个非中心点对象O_{random};

4. 计算用O_{random}代替当前代表对象O_j的总代价,如果此代价值小于当前值,那就用O_{random}迭代O_j,形成新的k个中心点的集合。

5. 返回第二步再次重复,直到不再发生变化。

2.3 随机初始中心点选择的不足

PAM是在给定的数据集合中随机选取k个初始中心点后进行聚类,如果选取的初始中心点不属于数据本身聚类的最佳中心点,PAM迭代的次数就会增加,计算的复杂性也随之增加,算法的有效性会下降。

而且PAM算法是基于k-中心点的聚类算法,如果数据集中存在孤立点和"噪声"时,这些离群点容易对聚类过程产生影响,使得到的聚类结果的准确性降低,所以PAM算法在处理含有孤立点数据时就存在不足。

*)基金项目:重庆市教委资助项目(编号:030201)。余建桥 教授,硕士生导师,研究方向为数据库技术、人工智能。张 帆 硕士研究生,研究方向为数据挖掘。

维普资讯 http://www.cqvip.com

3 数域空间中的场概念

空间领域的数据都是相互联系的，而且有些数据还具有不确定性，空间数据之间的相互联系和不确定性[3]更是如此。由于数据之间不是相互独立的，因此每个数据对数域空间中的其它数据都具有影响力，数据的这种影响力是随距离衰减的，这种影响力就形成了数据辐射和数据场。数据场[4]的引入把数据的能量推广到数域空间，使用场来处理数据不仅考虑了数据之间相互联系、相互影响的情况，同时也考虑了数据的不确定性[5]，使挖掘数据的手段更加符合现实、更加有效和更易于操作。

3.1 势场及其属性

数据通过辐射将其数据能量辐射到整个数域空间，被数据辐射所覆盖的空间就成为一个充满数据能量的数据场[6](图1)。数据通过自己的数据场对场中的另一数据发射能量，数据场的概念体现了数据之间不是相互独立的，而是相互联系的。

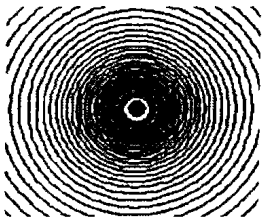

图1 数据场

数据场具有独立性，每个数据以自己为中心向外独立地辐射能量，这种数据的性质不会因其他数据的存在而受影响或有所改变。数据场除了具有独立性质外，其辐射的数据能量具有叠加的性质。当几个不同数据同时向一个观测点产生数据辐射时，这个观测点的数据能量就等于这几个数据独立辐射的数据能量在该点的数据能量之总和(图2)。

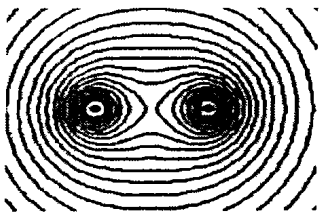

图2 数据场叠加

3.2 数据场的势函数

和物理上场的概念一样，数据场可用场强函数来表达数据场的分布规律，计算出数据场中各位置的能量大小。由于正态分布广泛存在于现实中，以及考虑到概率密度的分布函数和数据辐射大小随距离衰减的性质，数域空间中某个数据的场强分布函数形如：

$$p=e^{-\frac{r^2}{2k^2C_T(x)}} \tag{1}$$

根据式(1)的场强函数公式，通过对某点所受其周围数据辐射强度的累加合成，从而得到数域空间中数据场的势分布函数：

$$p=\sum_{i=1}^{n}p_i=\sum_{i=1}^{n}e^{\frac{r_i^2}{2k^2\cdot C_T(x_i)}} \tag{2}$$

式(2)中的 p 为某点所接受的全部数据辐射过来的数据场的能量强度之总和，称之为数据场的势；n 为数据的数量；r_i 为该点和数据 x_i 的距离；$C_T(x)$ 是数据 x_i 的辐射亮度，在很大程度上决定了数据辐射的能量强弱；k 为数据辐射因子。数据场的势是根据场强函数计算得到的全部单个数据场的强度之总和。

4 基于数据场的 PAM 算法

在数据场的概念和场强函数的理论基础上，进一步介绍势场所具有的等势线(面)、势心和自然聚类的概念，并分析势函数中辐射因子 k 对整个数据场的影响作用，进而将数据场的这些概念和性质用于改进传统的 PAM 算法。

4.1 势心及自然聚类概念

和传统描述等势线相似，在数域空间中把数据场中具有相同势值的点用平滑的曲线连接在一起就形成数据场的等势线。根据给定的势值，可以得到等势面。

由等势线(面)围绕的中心，称为势心。势心是空间实体在其属性数据值中所体现出来的特征值，势场中的势心构成空间实体的特征空间。数域空间内，一个数据形成的势场中，其势心就是该数据本身所在的位置；多个数据叠加而形成的势场中，其势心会靠近于辐射亮度较大的数据点，因为辐射亮度较大的数据点在叠加过程中起到的作用比辐射亮度小的数据点起到的作用大。一般情况下势心可能位于同类数据簇的中心位置，数据簇的势心是该类数据对数域空间中的某个概念的隶属中心，也即是数据在该概念特征聚类的中心点(图3)。

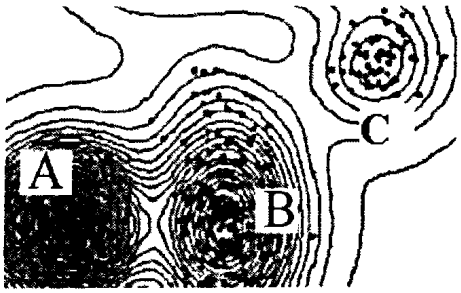

图3 等势线

图3中形成的自然拓扑结构显示了各点在空间中各自的聚集程度，而且还容易找出孤立点或“噪声”，离平均势最远的势最低点就可以认为是空间中的孤立点。如图3中 C 就可以看成为一个独立点。

4.2 k 因子对势场的影响

势函数公式中，数据辐射因子 k 是一个可改变的参数。它的改变与数据量的多少和数据分布有着联系，辐射因子 k 对数据场的分布规律起着调节作用，用户可以在处理数据过程中根据需要来改变 k，从而利用交互性的手段观察不同 k 值下，数据场的变化规律和分布情况。这增强了数据处理的过程中人机交互的灵活性。图4便是当数据的数量不变或变化时，k 对数据场的影响规律。

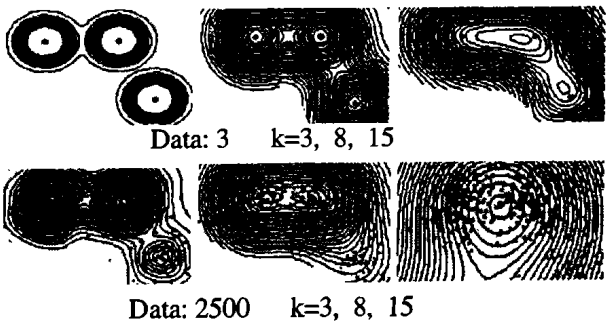

图4 辐射因子对数据场的影响

维普资讯 http://www.cqvip.com

4.3 消除数据中的噪声孤立点

前面我们已经提到，PAM 算法首先是要随机选择 k 个初始中心点。然后用迭代算法找出数据聚类的中心点。但如果数据集中存在孤立点和“噪声”时，这些离群点就会被考虑到算法中，从而影响到算法找出 k 个聚类中心点的准确性。

根据势分布函数和数据辐射因子对数据场的场强函数和势函数的调节作用，根据情况选取不同的数据辐射因子，找出那些势较小的点，越是孤立点，其势越小(图5)。这就能找出并清除孤立点和“噪声”，如图5中的 C 点可看成是孤立点和“噪声”。

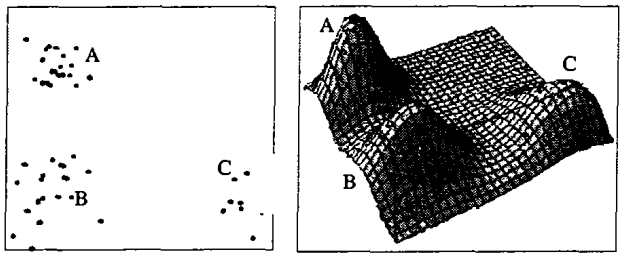

图5 调节辐射因子找出孤立点

4.4 利用势分布函数确定首选的中心点

当选择的 k 个初始中心点接近数据中最佳的中心，那就能用较少的迭代次数很快找出聚类的 k 个中心，这样就减少了算法的复杂性。但事实上，一般情况下随机选取初始中心是难以接近数据本身聚类的实际中心点。

前面通过调节数据场势函数中的数据辐射因子，找出并清除孤立点和“噪声”后，就能得到数据集的一个自然聚类(图3)，也就是数据的特征聚类中心。我们首先找到具有 k 个自然聚类的势心，找到的势心本质上就代表了同类簇的重心位置，然后我们选取在这 k 个势心或其附近的点作为 PAM 算法的 k 个初始中心点，这就能确保能最大限度使得选取的 k 个初始中心点尽可能代表或接近数据集实际聚类的中心点。这就在 PAM 算法中避免或减少了为找出 k 个聚类中心所进行的迭代重复次数，减小了算法的复杂性并提高了算法的有效性。

4.5 灵活的交互性聚类

把数据场引入 PAM 算法的聚类过程中，可以在聚类过程中根据实际情况或需要临时修改数据辐射因子 k。根据 k 值的不同，我们可以得到不同数目的势心，而这些势心一般都很可能是数据簇的中心位置，这样我们就能在聚类的过程中进行灵活的聚类挖掘，即能根据用户需要临时对聚类的类型数量进行调整。

由于数据场具有能处理不确定性数据的能力，所以改进后的算法也能用于空间数据或处理带不确定性数据的情况，而且可推广到多维空间。

4.6 效果评测

对同一数据集中的数据采用传统的 PAM 算法和使用改进后 PAM 算法进行聚类后进行比较，分别以不同数量的数据进行实验(图6)。由于改进后的 PAM 算法在选择初始的 k 个中心点时利用了每个自然聚类的势心点或其邻近的点，所以选择的 k 个初始中心点几乎代表或接近数据集中实际的聚类中心点，所以除了能很快准确找出数据中心点的位置，还能减少挖掘出高质量聚类所需要花费的时间。

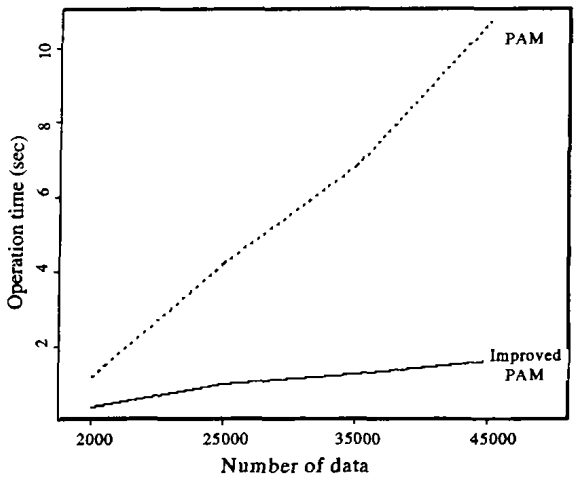

图6 改进前和改进后 PAM

结束语 引进数据场技术对现有的 PAM 聚类算法进行了有益的改进，使其消除孤立点和“噪声”，避免了中心点不受极端数据的影响；而且还能正确、有效果地选取初始中心点，实验证明了这种基于数据场改进的 PAM 聚类算法的有效性。而且这种基于数据场的 PAM 算法能在聚类分析中灵活地调整聚类的种类，这是传统 PAM 算法所不及的。

能否减少在处理数据场势函数时的计算开销，能否将其运用到大型数据集中，以及能否把它作为一个基础来建立更加复杂的聚类数据挖掘算法等，都有待进一步的研究。

参 考 文 献

1 Han J, Kanber M. Data Mining Concepts and Techniques. Morgan Kaufmann, 2001

2 Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley&Sons, Inc., New York, 1990

3 Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Mining. In: Proc. of the 20th VLDB Conf., 1994. 144～155

4 Shekhar S, Chawla S. Spatial Databases. Person Education, Inc, 2003

5 李德毅，王晔，吕辉军. 知识发现机理研究. 中国人工智能进展，2001

6 王树良. 基于数据场与云模型的空间数据挖掘和知识发现：[武汉大学博士学位论文]. 2002