第 33 卷第 4 期 河北科技大学学报 Vol. 33, No. 4
2012 年 8 月 Journal of Hebei University of Science and Technology Aug. 2012

文章编号:1008-1542(2012)04-0338-05

随机种子最近邻居搜索聚类算法研究

苏亚然[1,2],陈军霞[1],牛习现[3]

(1.河北科技大学经济管理学院,河北石家庄 050018;2.华北电力大学经济与管理学院,北京 102206;3.河北青年管理干部学院信息技术与传播系,河北石家庄 050031)

摘 要:提出了随机种子最近邻居搜索(RS-NNS)聚类算法,该算法从随机确定的种子开始沿着它最近邻居的方向搜索具有最大相似特征的邻居对象,形成局部最大聚类集合,并在搜索过程中动态调整数据对象的归属,以实现局部的最优分配,直到所有的数据对象完成聚类标识。经过验证,该算法可以适应数据集合的密度、形状、噪音、聚类个数等问题,并且相对于同类算法可以实现较快地优化搜索。

关键词:最近邻居搜索;随机种子;聚类分析;数据挖掘

中图分类号:TP301 **文献标志码**:A

Study on random seed nearest neighbour search clustering algorithm

SU Ya-ran[1,2], CHEN Jun-xia[1], NIU Xi-xian[3]

(1. College of Economics and Management, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. College of Economics and Management, North China Electric Power University, Beijing 102206, China; 3. Faculty of Information Technology and Propagation, Hebei Youth Administrative Cadres College, Shijiazhuang Hebei 050031, China)

Abstract: This paper presents a random seed nearest neighbour search clustering algorithm (RS-NNS). The method is to follow the nearest neighbours' direction of a random selected seed, search and find its neighbours which have the greatest similar features, form the local maximum cluster, adjust dynamically the data objects' belongingness to realize the local optimization, and end the clustering procedure until all the data objects are identified. Experiments verify that the new algorithm fits the problems such as different density, shape, noise, cluster number and so on, and can realize fast optimization searching.

Key words: nearest neighbour search; random seed; clustering analysis; data mining

聚类作为一种重要的数据分析方法,已经在模式识别、机器学习、数据挖掘、图像处理、文档分析、信息检索、医疗图像分析、市场研究、质量控制和欺诈检测等方面得到了广泛应用。数据聚类是一个探索和描述性数据分析的过程,并在这个过程中将相似的对象或数据进行标识和分组。聚类的任务是在没有训练样本的情况下,仅利用样本间的相似性寻找样本集针对某个评判准则的最佳类别划分[1-2]。聚类分析属于无监督学习问题,用来在无标识的数据集合中发现其内在结构和联系,将对象按照某方面的相似性进行组织分组的过程,因此每个聚类都是对象的集合,并且具有很强的相似性,而不同聚类之间的对象则具有相对较弱的相似性或者不具有相似性[3]。在实现数据集合聚类分析的研究过程中,针对不同的数据类型和应用目的,相关的研究人员对问题的关注方向也不尽相同,如对密度的适应能力、形状的适应能力、噪音的检测、边界对象的识别、聚类个数的确定、聚类结果的准确度、算法的优化速度、高维数据的聚类问题,因此研究具有最大的适应

收稿日期:2011-12-30;责任编辑:李 穆
基金项目:河北省社会科学基金资助项目(HB12YJ064)
作者简介:苏亚然(1972-),女,河北灵寿人,讲师,主要从事技术经济方面的研究。

能力的聚类算法就成为聚类研究的重要方向之一。早期的聚类算法，如 K-均值、K-中心点、FCM 等，它们只对空间分布为球形或者超球体的数据具有较好的性能，而对空间分布复杂的数据聚类效果较差[1]，而且它们都需要预先设定聚类的个数。基于密度的 DBSCAN 算法，考虑到数据分布的密度特性，对具有不同形状特征的数据集合有很好的适应性，但是它的参数设置比较困难且无法适应密度变化比较大的数据集合的数据分析任务。为了改进聚类的结果和算法的适应性，有些研究人员提出了基于 SNN(shared nearest neighbours)相似性聚类分析算法如 JP(jarvis-patrick)聚类算法，该类算法对数据对象集合的大密度变化以及任意形状均具有较好的适应性，但是该类算法需要完成较大量数据对象之间的相似性计算，算法的性能还有提高的空间，另外，对于计算对象相似性时的共享邻居个数的设定需要相关领域的知识和经验。经过对相关领域的聚类算法的综合研究和验证，笔者提出了随机种子最近邻居搜索(RS-NNS：random seed nearest neighbour search algorithm)聚类算法，该算法从随机确定的种子开始沿径向搜索具有最大相似特征的邻居对象，并形成局部最大聚类集合，并在搜索过程中动态调整数据对象的归属，以实现局部的最优分配，直到所有的数据对象完成聚类标识。经过验证，该算法可以有效地适应数据集合的密度、形状、噪音、聚类个数等问题，并且相对于同类算法可以实现较快的优化搜索。

1　相关领域问题的研究与分析

1.1　DBSCAN

DBSCAN(density-based spatial clustering of application with noise)是一种基于密度的聚类算法。该算法将具有足够高的密度的区域划分为簇，并在具有噪音的空间数据库中发现任意形状的簇，它将簇定义成密度相连的点的最大集合[2]。在给定了核心点、边界点和噪音点的定义以后，DBSCAN 算法中 Eps 和 MinPts 2 个参数的设置对聚类结果的影响非常大，并且它们的选择比较困难。DBSCAN 算法的描述性定义如下：

1）把所有的数据对象标记为核心点、边界点或者噪音点；

2）忽略所有的噪音点对象；

3）按照 Eps 为每个核心点对象建立边界；

4）把所有相互连接的核心点对象标记为同一簇；

5）把每个边界点分配到与它相关联的核心点对象相同的簇。

DBSCAN 的基于密度的簇的定义方式，使得该算法具有较强的抗噪音能力，并且能够分析处理具有任意形状和大小的簇的数据对象集合，因此能够发现很多 K-means 算法不能处理的簇。然而 DBSCAN 算法也存在相应的缺陷和不足：

1）当分析处理的数据对象集合的分布密度有较大变化时，Eps 和 MinPts 参数的选择将变得非常困难；

2）对高维数据对象集合分析时，密度的定义很难实现；

3）用于获得近邻对象的计算代价较大，跟高维数据集合的计算代价一样[1]。

1.2　基于 SNN 相似性的 JP 聚类算法

基于 SNN 相似性算法的基本思想是：如果 2 个数据对象与其他许多相同的数据对象具有相似性，尽管直接的相似性度量方法可能确定不了这种相似性，但是这 2 个数据对象之间的相似性是成立的。SNN 相似性定义为具有低密度或者密度变化较大特征的数据集合的聚类分析提供了可行的思路。JP 聚类算法通过最近共享邻居方法进行对象聚类，该算法的执行需要确定数据对象之间距离的度量方法以及 2 个参数 J 和 K，J 是最近邻居列表的大小，K 是共享邻居的个数。该算法的描述如下。

1)对聚类分析的数据集合中的每个对象确定它的 J 个最近的邻居；

2)把符合条件的对象分配到同一个簇中，它们相互包含在对方的最近邻居列表中，并且至少拥有 K 个共享的邻居对象。

因为 JP 聚类算法是基于 SNN 相似性概念的，所以它能够处理带有噪音、边界的数据集合的数据分析任务并且能够发现不同大小、形状和密度的数据对象簇。该算法对于高维数据的分析处理，特别是对于具有强关联性的结合紧密的簇的发现也是非常有效的。然而，JP 聚类算法把簇定义为在 SNN 相似图中相连接的对象集合，通过对单连接的判定来决定是否对一个对象集合进行分割或保留为一个簇，因此 JP 聚类算法

在一定意义上是脆弱的，它可能分割真正的簇或者合并应该保持分离的簇。另外，JP聚类算法不能实现对象的完全聚类并且最佳参数的选择也比较困难[2-6]。

2　随机种子最近邻居搜索聚类算法

数据分布密度、SNN相似性等测度方法的引入，使得复杂分布数据的结构特点得到了较好的体现，对于大多数的数据聚类分析问题能达到较好的效果。但算法的性能，如自然密度变化的数据分布的识别、不规则形状的数据分布的识别等仍然存在改进和提升的空间，因此随机径向搜索算法试图通过数据对象本身自然存在局部集聚特性以及数据对象的局部关联特性，来实现对数据对象的最佳聚类和算法性能的提升。

2.1　算法的实现原理

数据对象的分布是有局部集聚特性的，否则就失去了聚类分析的意义。对于任意一个数据对象而言，不管它周围的数据分布密度如何，总可以发现一些相对来说比较近的邻居对象，基于数据分布局部集聚特性，可以认为当前数据对象以及它的最近的邻居就可以构成一个具有局部集聚特征的聚类。在现实世界中，各种对象之间的关系是可以传递的，所以可以认为邻居的邻居之间也存在必然的关联。基于以上基本思想，可以从一个任意的种子对象开始，沿着它的四周邻居的径向方向进行关联搜索，并在搜索过程中对参加扩展搜索的邻居对象加以限制来实现局部数据对象的聚类发现，算法搜索的具体实现原理见图1。在图1中，从随机选定的对象点A开始，沿着它的每个最近邻居所在的方向开始搜索，符合条件的每个邻居都会成为新的搜索种子，直到完成局部集合的最优搜索过程，然后再随机选取没有标识的数据对象，如数据对象B，展开另一个聚类的搜索过程，直到实现整个数据空间的搜索工作。这样的搜索过程是通过邻居对象逐渐展开的，并且包括了各种可能方向的搜索，因此它可以达到对任意形状和密度变化的数据集合的有效适应[7-8]。

2.2　算法的实现步骤

为了提高算法的适应能力和聚类结果的准确性，针对一些关键的技术细节问题，提出了以下解决方法。

1）弱关联数据对象间的归属评判问题

在聚类搜索的过程中，对于任意分布的数据对象，可能存在一些数据对象虽然属于当前对象的最近邻居集合，但是该对象并不一定适合作为进一步搜索的种子对象，因此需要采取一定的方法避免让该类数据对象参加进一步的聚类搜索。由于每个对象在其局部范围内都会存在相对较近的邻居，这就为人们进行判定提供了依据，可以认为一个对象如果它的最近的邻居多数都属于当前种子对象的邻居集合或者不属于当前已经标识的数据集合，则可以认为它不适合作为种子进行进一步的搜索。

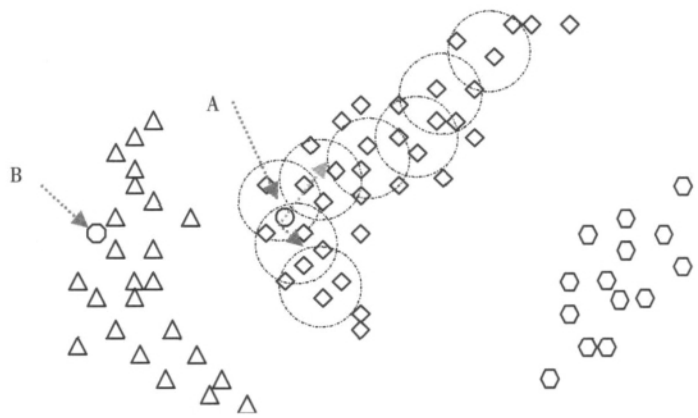

图1　RS-NNS算法实现原理图

Fig.1　Principle of RS-NNS clustering algorithm

2）已经标识的对象的最优判定问题

在搜索过程中如果遇到已经标识的数据对象，应该对它的归属做出最佳的判定。对于一个已经被标识的数据对象，如果在搜索过程中发现它的邻居对象中归属于其他类型的聚类标识占多数，则认为它应该标识为其他聚类类别，以实现数据集合的局部最优搜索。

3）初始种子对象选择问题

在整个聚类过程中总会存在多次初始种子对象的选择问题，如果它的选择正好落在数据对象比较集中的区域，则可以完成高质量的局部搜索，得到一个较好集聚的数据聚类，但是如果初始种子的选择落在了边界或噪音点上，则有可能得不到所预期的聚类结果，因此必须对种子对象的选取和搜索加以控制。如果当前种子数据对象到它最近的2个邻居的距离远远超出它的最近2个邻居的距离，则认为该种子对象为噪音数据或者停止对该种子的继续搜索。

2.3　随机种子最近邻居搜索算法一般性描述

通过对相关聚类算法的研究，总结和分析了不同类型的数据集合的分布特点，在此基础上设计了 RS-NNS 聚类算法，算法的具体实现如下：

1)读取数据集合并计算各个数据对象之间的距离，得到距离矩阵。

2)根据设定的最近邻居数 K，获得每个对象的最近邻居序列。

3)随机选取 1 个未标识对象作为种子，设置它的聚类标识，并把它压入堆栈。

While(堆栈非空){

获取栈顶数据对象；

取得当前数据对象的最近邻居列表；

把符合条件的最近邻居压入堆栈，并设置它的聚类标识；

}

4) 如果数据集合中还存在未标识的数据对象，则重复第 3)步，否则完成聚类过程。

3　试验与结果分析

为了对 RS-NNS 进行验证，用 Java 语言完成了相关的算法编程实现，将 RS-NNS，DBSCAN，JP 算法在大量的不同特性的人工合成数据集合上进行了验证工作，并对试验的结果进行了比较分析，试验结果表明，笔者所提出的 RS-NNS 算法是有效的，能够较好地适用各种类型的数据集合的聚类分析任务。该算法的设计、试验运行的环境均为 CPU Intel Pentium Dual Core 1.7 GHz、内存 2 GB 的微型计算机。

3.1　算法性能的验证与分析

为了能够直观地考察 RS-NNS 算法的性能，通过软件方法生成了分别有 100，250，500 个数据对象的多组模拟数据集合，让 3 个算法分别运行在不同大小的数据集合上，并取得它们运行的平均时间，它们对数据的处理性能如图 2 所示，可以明显看出 RS-NNS 算法相对于其他算法的优越性。

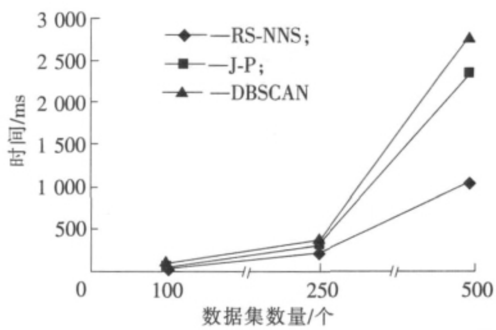

图 2　3 种算法性能比较

Fig. 2　Performance comparison of three algorithms

3.2　算法对数据集合的适应能力的验证与分析

为了更好地说明该算法对于具有不同分布密度特点的数据集合的适应能力，笔者选取了在试验中具有代表性的数据集合进行了测试。测试结果显示，该算法对于不同密度分布的数据对象的适应优于 DBSCAN，克服了 DBSCAN 必须由用户指定静态的密度参数的缺点，能够正确地识别有较大差别的不同密度的数据集合中的自然形状的聚类，如图 3 所示。试验结果表明该算法对于噪音数据的处理非常有效。

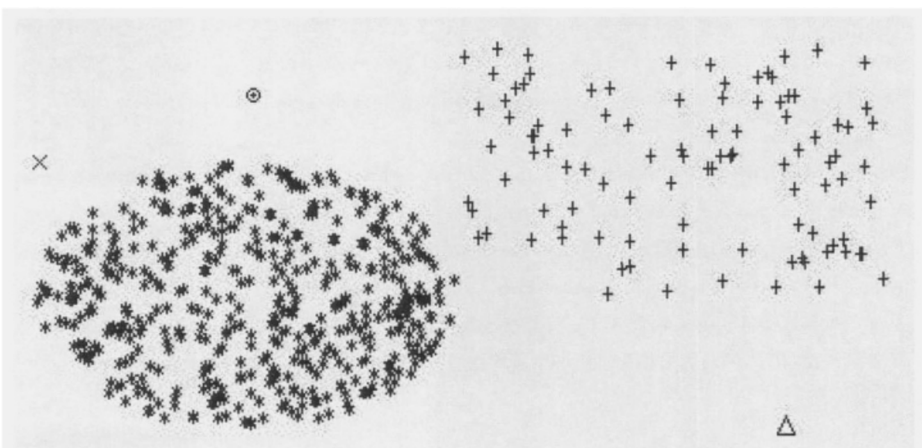

图 3　RS-NNS 算法的密度噪音适应性分析图

Fig. 3　RS-NNS adaptability of different density and noise

在聚类分析中,数据集合分布密度的变化,不一定是绝对的高密度和低密度,有可能会存在逐渐过渡的密度变化情形,这样的数据对象的局部集聚往往也是合理的,因此在进行聚类算法的设计时应考虑对此类聚类数据的处理和分析。另外,聚类数据的分布不一定是球形或者其他规则的形状,任意形状的数据聚类的存在也非常广泛。RS-NNS算法的设计充分考虑了以上数据分析的情况,使得它能够很好地适应密度逐渐变化以及任意形状的聚类的分析任务,它的适应性测试结果见图4。另外通过大量的试验表明,该算法同样可以准确地自动识别出数据集合中自然存在的聚类个数。

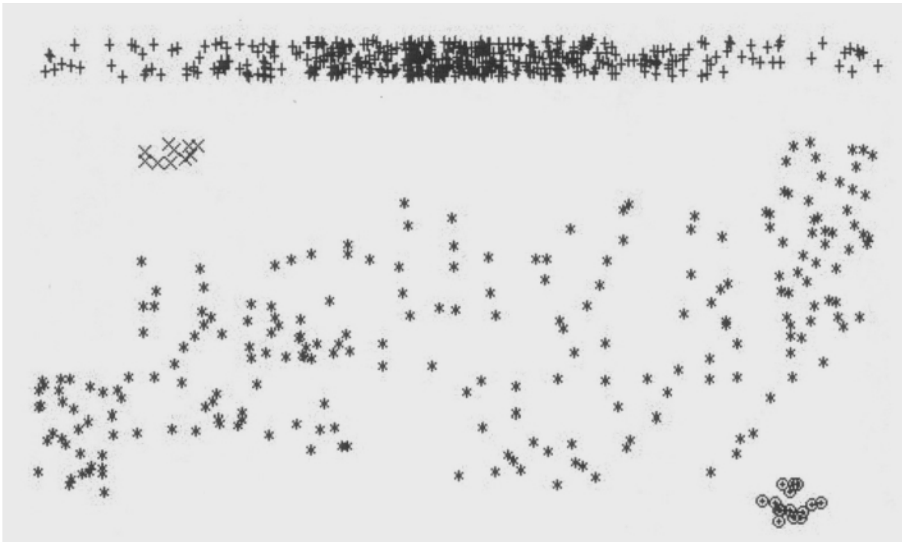

图4 RS-NNS算法对逐渐变化密度、任意形状、聚类个数适应性分析图

Fig.4 RS-NNS adaptability of gradually changed density, arbitrary shape and cluster number

4 结 语

笔者对当前流行的聚类算法(如DBSCAN和JP等)以及不同算法所针对的数据分布特点进行了综合分析,针对这些算法的局限性提出了RS-NNS算法。测试了不同算法在相同数据集合上的运行,RS-NNS算法通过简化数据处理过程,取得了明显的运算优势。另外,RS-NNS算法针对不同类型的数据集合也表现出了良好的适应性,它可以准确地识别不同密度、逐渐过渡的密度变化以及任意形状的数据对象的聚类集合。

参考文献:

[1] 公茂果,王 爽,马 萌,等. 复杂分布数据的二阶段聚类算法[J]. 软件学报(Journal of Software),2011,22(11):2 760-2 771.

[2] LEE J S. Sigurdur olafsson data clustering by minimizing disconnectivity[J]. Information Sciences, 2011,181:732-746.

[3] 牛习现,赵立川. 利用局部集聚特性的聚类算法的研究[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology),2011,32(5):466-470.

[4] JIM Z C, LAI A, HUANG T J. An agglomerative clustering algorithm using a dynamic *k*-nearest-neighbor list[J]. Information Sciences, 2011,181:1 722-1 734.

[5] GONZALEZ-BARRIOS J M, QUIROZ A J. A clustering procedure based on the comparison between the *k* nearest neighbors graph and the minimal spanning tree[J]. Statistics & Probability Letters, 2003,62(1):23-34.

[6] NOHA A Y, MOHAMED S K, MOHAMED A I. A distance-relatedness dynamic model for clustering high dimensional data of arbitrary shapes and densities[J]. Pattern Recognition,2009, 42:1 193-1 209.

[7] 储岳中,徐 波. 动态最近邻聚类算法的优化研究[J]. 计算机工程与设计(Computer Engineering and Design),2011,32(5):1 687-1 690.

[8] 王 茜,杨正宽. 一种基于加权KNN的大数据集下离群检测算法[J]. 计算机科学(Journal of Computer Science & Technology),2011,38(10):177-180.