文章编号:1672 - 058X(2013)06 - 0039 - 05

一种基于参考点的快速 k-均值算法

李 有 明
(重庆师范大学 数学学院,重庆 401331)

摘 要:聚类分析是模式识别的一个重要分支,以核心点和 k-均值算法为基础,提出了一种基于参考点的快速 k-均值算法;本算法以参考点作为第一个初始聚类中心,剩余初始聚类中心在核心点中选取,使得初始聚类中心能更好的反映模式样本集的几何特征,并且能减少迭代次数。

关键词:参考点;密度;k-均值
中图分类号:TP275 **文献标志码**:A

1 概 述

聚类分析是数据挖掘的一个重要分支,是将给定数据集中的数据划分成为不同组的过程。将给定的数据集称为模式样本集,给定数据集中的元素称为模式样本,将划分的组称为类。聚类分析要求两个模式样本如果属于同一个类,则这两个元素具有较高的相似度,反之如果两个元素分别属于不用的类,则这两个元素具有较低的相似度。它不仅可以作为一种工具来分析数据在空间中的分布情况,还可以作为其他算法的预处理工具,是一种高效常用的的信息组织手段。聚类分析经过多年的发展,出现的很多的算法,它们大致可以分为 5 类[1]:基于划分的聚类算法,基于层次的聚类算法,基于统计模型的聚类算法,基于密度的聚类算法,基于网格的聚类算法。

目前关于聚类的研究主要集中在寻找能对大型数据库进行快速有效聚类的算法。将不同的聚类方法结合是聚类分析现在研究的热点,给聚类分析这个分支带来了新的活力。

本文涉及的主要概念介绍[2]:模式样本集重心 G_N:$G_N = \overline{X}$为模式样本集的均值。

$N_{Eps}(p)$:表示以模式样本 p 为中心以 Eps 为半径的超球状区域内模式样本的集合。故 $N_{Eps}(p) = \{q \in G \mid dist(p,q) \leqslant Eps\}$ 示为,其中 $dist(p,q)$ 表示模式样本 p 和模式样本 q 之间的距离,G 表示给定的模式样本集。

密度 $density(p)$:$N_{Eps}(p)$ 中元素的个数,记为 $|N_{Eps}(p)|$。

核心点:对于给定的密度阈值 Min Pts,如果模式样本 p 的密度 $|N_{Eps}(p)|$ 大于 Min Pts,则 p 称之为核心点,反之,称之为非核心点。

直接密度可达:如果模式样本 p 与模式样本 q 之间条件 p 在 q 的领域中,即有 $p \in N_{Eps}(p)$;q 是核心点;则称模式样本 p 与模式样本 q 直接密度可达。

密度可达:存在一连串模式样本 $p(1),p(2),\cdots,p(n)$,如果其中模式样本 $p(i)$ 与模式样本 $p(i+1)$ 直接

收稿日期:2012 - 11 - 21;**修回日期**:2012 - 12 - 05.

作者简介:李有明(1988-),男,重庆巫山人,硕士研究生,从事聚类分析理论研究.

密度可达($i=1,2,\cdots,n-1$),则称 $p(n)$ 从 $p(1)$ 密度可达。

密度相连:若存在模式样本 o,使得模式样本 p 和模式样本 q 都从 o 密度可达,则称模式样本 p 与模式样本 q 之间是密度连接的。

参考点:密度最大的模式样本为参考点,可知参考点的个数可能不止一个。

聚类:模式样本集$\overline{D}$为待分类模式样本集 G 中的一个聚类,当且仅当$\overline{D}$满足模式样本集中任意两个模式样本 p 和 q,如果 $p\in\overline{D}$,且 q 从 p 密度可达,则有 $q\in\overline{D}$;如果空间中任意两点 p 和 q,有 $p\in\overline{D}$,则有$q\in\overline{D}$,则 p 和 q 是密度连接的。

2 相关工作介绍

2.1 关于 k-均值算法的介绍[4]

2.1.1 k-均值算法步骤

k-均值算法是一种经典的聚类算法,其步骤如下:

(1) 选择 k 个初始聚类中心;

(2) 将模式样本集中的每个模式样本根据一定的规则指派到某个聚类中心之中去,形成 k 个聚类;

(3) 重新计算每个聚类的中心;

(4) 重复步骤(2)、(3),直到聚类中心不在发生变化。

2.1.2 k-均值算法讨论

作为一种典型快速的聚类算法,k-均值算法有快速、理解简单等优点。但其本身也有缺陷。例如需要指定出使聚类个数以及需要选择 k 个初始聚类中心。聚类中心的选择不同,可能会得到不同的聚类结果。虽然提出了不少的改进办法,但是无论怎么改进,在最开始一般都很难选择出正确的聚类中心,要得到正确的聚类中心还需要不断的迭代,所以初始聚类中心的选择不仅影响到了聚类的结果,还影响到了迭代的次数,从而影响到了算法的效率。

2.2 关于基于密度的聚类的介绍[2]

(1) 基于密度聚类的介绍。对于一个聚类中的某一个模式样本,在其给定半径的领域中包含的对象不能少某一给定的最小数目,然后对具有密度连接特性的对象进行聚类。其基本步骤是按照某种顺序考察模式样本集 D 中的某个模式样本 p,如果 p 是核心点,则找到 $N_{Eps}(p)$,如果领域中的点和 p 属于同一个类,那么它们将作为下一轮考察的种子点。

(2) 关于基于密度聚类算法的讨论。基于密度的聚类算法可以挖掘任意形状的聚类,对数据的输入顺序不敏感,并且还具有处理异常数据点的功能,在这个方面比 k-均值算法要提高很多。对于有 N 个模式样本的模式样本集,该算法的时间复杂度为 $O(N^2)$。基于密度的聚类算法可以克服基于划分的聚类算法的部分缺陷,例如能自动发现聚类的数目、能够发现任意形状的类、对于噪声的处理较好的能力,但是基于密度的聚类算法是从任意点开始搜索,忽略了数据集本身具有的密度属性。

2.3 关于最大最小聚类算法的介绍

2.3.1 最大最小聚类算法的步骤[5]

任意选取一个模式样本为第一个聚类中心 $Z_{(1)}$;选择离 $Z_{(1)}$ 最远的模式样本作为第二个聚类中心 $Z_{(2)}$;逐个计算每个模式样本与已经确定的所有聚类中心之间的距离,并且选出其中最小距离。因为共有 N 个模式样本,故有 N 个最小距离;在所有最小距离中选出最大的一个距离,如果该最大距离大于给定的值,则这个值对应的模式样本为新的聚类中心,并且返回上一步,否则,聚类中心的计算步骤结束;寻找聚类中心的运算结束后,将模式样本按照距离最小的划分原则将它们指派到相应聚类中心所代表的聚类之中。

2.3.2　最大最小聚类算法的讨论

最大最小聚类算法具有快速简单的优点。可是最大最小聚类算法同样需要选择初始聚类中心,需要确定一个阈值,并且聚类中心在确定之后没有在改变,且没有迭代的过程,这样就可能不会选择出真正的聚类中心,而是聚类中心的替代点。

3　k－均值算法分析与实现

3.1　初始参考点的选取

在开始讨论初始参考点选取原则之前,在此先引入以下几个概念。最小 $R-M$ 距离(最小参考点——模式样本集重心距离):由于参考点有可能不止一个,则参考点到模式样本集重心的距离也可能存在多个,称这些距离中最小的那个距离为最小 $R-M$ 距离。最小距离对应的参考点称为最小 $R-M$ 距离参考点(最小参考点——模式样本集重心距离参考点)。

可能存在多个最小 $R-M$ 距离参考点,将这些参考点构成的集合称为最小 $R-M$ 距离参考点集合,记为最小 $R-M$ 集合。最小 $R-M$ 集合不会是空集,至少有一个元素。

根据定义,参考点为密度最大的模式样本,本来不会出现什么问题,但是当出现了两个或者两个以上的参考点的时候,选择什么样的参考点来开始聚类? 本文给出了以下准则:选择最小 $R-M$ 距离参考点作为初始参考点,如果存在多个最小 $R-M$ 距离参考点(最小 $R-M$ 集合中的元素个数大于 1),则随机从最小 $R-M$ 集合中选取一个作为初始参考点。

3.2　算法步骤分析

3.2.1　核心点的确定

要确定核心点,需要确定领域半径 *Eps* 和密度阈值 Min *Pts*。在选取 *Eps* 和 Min *Pts* 的时候要保证核心点集合中的元素足够的多而使得核心点组成的集合能够大致代表原模式样本集的几何性质,同时也要保证核心点组成的集合中的元素不要太多而不能达到简化计算的效果。根据给定的领域半径 *Eps* 和密度阈值 Min *Pts*可以求出由核心点组成的集合$\overline{G}$,在本文提使用的算例中使用 4 作为密度阈值,而 *Eps* 需要根据数据集的特征观察得出[6]。

3.2.2　初始聚类中心的选取

因为初始聚类中心需要在一定程度上反映模式样本集的几何性质,故假设非核心点不能成为初始聚类中心,则初始聚类中心的选择仅在$\overline{G}$中考虑。第一个聚类中心的选择通常会关系到其他初始聚类中心的选择,也会影响到后面迭代的次数,故在选择第一个初始聚类中心的时候要特别的注意,以最小 $R-M$ 距离参考点为第一个初始聚类中心,这是因为最小 $R-M$ 距离参考点离模式样本集的重心的距离最近,与其他的点相比更能反映模式样本集的几何性质。

其他聚类的中心的选择需要根据实际情况来确定。如果给定了聚类的数目 k,则可以按照最大最小聚类算法找出 k 个初始聚类中心。其具体步骤如下:以最小 $R-M$ 距离参考点作为第一个初始聚类中心。在$\overline{G}$中选择离最小 $R-M$ 距离参考点最远的模式样本作为第二个初始聚类中心。求出$\overline{G}$中模式样本到每个初始聚类中的聚类的距离,取最小的距离,然后在这些最小距离中取最大的距离相对应的点作为初始聚类中心,直到找到 k 个初始聚类中心。

如果没有给定聚类的个数,可以根据基于密度的聚类算法,先将$\overline{G}$中的元素进行基于密度的聚类,假设聚类之后得到 k 个类,选取每个类中密度最大的模式样本为初始聚类中心,这样就可以得到 k 个初始聚类中心。

3.2.3　非核心点的处理

对于模式样本集 G 中非核心点由于没有核心点重要,故在选取初始聚类中心时不予考虑,在迭代的时

候可以使用不同的处理方法。由于非核心点不像核心点那样能够反映模式样本集的几何位置,故对于大型的数据库,对于非核心点的处理可以借鉴判别分析的思想,例如:将核心点组成的集合看做一个待分类模式样本集,先对核心点聚类。完成对核心点的聚类后,可以根据最近邻法或者 k - 近邻法将非核心点划入相对应的聚类之中去。也可以在迭代的过程中认为核心点与非核心点没有差异,在迭代的过程中将非核心点考虑在内。本文对于非核心点的处理是在迭代的过程中将其考虑在内。

3.3 快速 k - 均值算法设计

(1) 核心点的求解:根据给定的 *Eps* 和 Min *Pts*,求出由模式样本集的核心点组成的集合,将其记为$\overline{G}$,并且求出最小 $R-M$ 距离参考点,将其记为 $p_{\max}$,则 $p_{\max}$ 具有性质 $p_{\max}=\{p\in\overline{G};\forall q\in G,\exists density(p)\geqslant density(q)\}$,根据$\overline{G}$中所有元素到最小 $R-M$ 距离参考点 $p_{\max}$ 的距离将集合$\overline{G}$中的元素进行排序。得到有序集$\overline{G}$,此时$\overline{G}=\{p_1,p_2,\cdots,p_M\}$,为一个有序集合,其中 $p_{\max}=p_1$,对于$\overline{G}$中的任意两个元素,如果 $i<j$,则有 $dist(p_{\max},p_i)>dist(p_{\max},p_j)$;

(2) 按照 3.2.2 的讨论,选择出 k 个初始聚类中心;

(3) 根据一定的规则,将 G 中的元素指派到相应的类中;

(4) 重新计算聚类中心;

(5) 重复步骤(3)、(4),直到聚类中心不再发生变化。

4 算法性能分析

为了验证本为提出算法的有效性,将使用 2 个二维数据集 DS1、DS2 和一个四维数据集 iris 作为待分类模式样本集。DS1、DS2 如图 1(a)和 1(b)所示。其中 DS1 包含 3 个聚类,每个聚类中有 100 个模式样本,DS2 中包含 4 个聚类和若干噪声点。Iris 数据集是一个四维数据集,包含 3 个聚类,每个聚类包含 50 个模式样本。实验结果如表 1 所示。

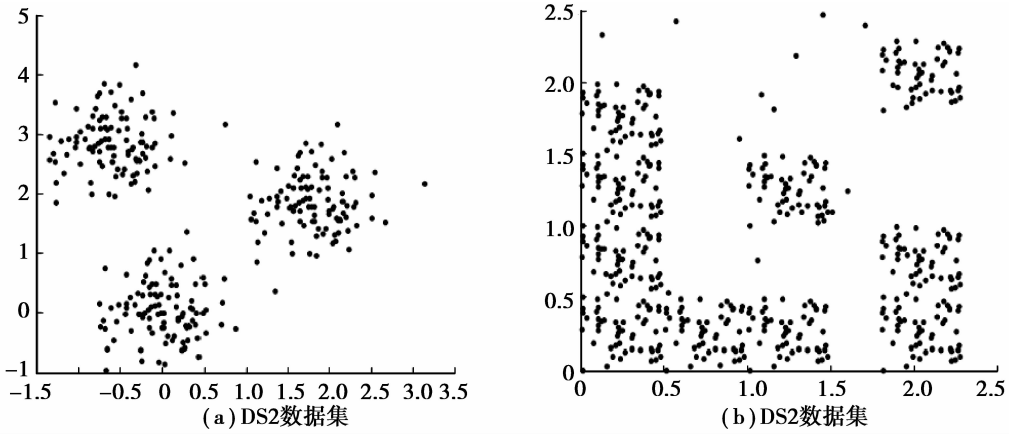

图 1 DS1 数据集和 DS2 数据集

k - 均值算法的聚类个数都需要手工设定;在验证本文算法时,DS1 与 DS2 的聚类个数是采用基于密度的聚类算法自己发现的,iris 数据集是手工设定;对 DBSCAN 进行验算时其聚类个数是算法自己发现的。

由表 1 可以看到,对于类似于 DS1 这样具有明显的分离状态的数据集,k - 均值算法、DBSCAN 算法和本文算法都有较高的聚类准确度;对于类似于 DS2 的数据集,有较多的孤立点,k - 均值算法的聚类效果较差,但是 DBSCAN 算法和本文算法具有较高的聚类准确度;对于类似于 iris 这样的高密度的球状数据集,k - 均值算法、DBSCAN 算法和本文提出的算法有较高的聚类准确度。

表 1　聚类实验结果

数据集	聚类算法	聚类个数	聚类准确度/%
DS1	k – 均值算法	3	100
	DBSCAN	3	94.28
	本文算法	3	97.72
DS2	k – 均值算法	4	69.33
	DBSCAN	4	97.64
	本文算法	4	94.32
iris	k – 均值算法	3	92.27
	DBSCAN	3	91.01
	本文算法	3	93.89

参考文献:

[1] 许虎寅,王治和. 一种改进的基于密度的聚类算法[J]. 微电子学与计算机,2012.02(26):44-46

[2] 马帅,王腾蛟,唐世渭,等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报 2003,14(6):1089-1095

[3] 王晶,夏鲁宁,荆继武. 一种基于密度最大的聚类算法.[J]. 中国科学院研究生学报. 2009.26(4):539-548

[4] 边肇祺,张学工. 模式识别[M]. 2 版. 北京:清华大学出版社,2000

[5] 齐敏,李大建,郝重阳,等. 模式识别导论[M]. 北京:清华大学出版社,2009

[6] ESTER M, KRIEGELl H, SANDER J . A density-based algorithm for discovering clusters in large spatial databases with noise [c]//Usama M Fayyad, Padhraic Smyth, Gregory Piatetsky Shapiro, Eds. Proc of 2nd International Conference on konwledge Discovery and Data Mining(KDD'96). Porland:ACM press,1996:226-231

[7] 孙凌燕,杨明,任建斌. 一种基于相对密度的快速聚类算法[J]. 微电子学与计算机,2009,12 (26):109-111

A Fast K-mean Algorithm Based on Reference and Density

LI You-ming

(School of Mathematics, Chongqing Normal University, Chongqing 401331, China)

Abstract: Clustering analysis is an important branch of pattern recognition, in this paper, based on the core-point and *k*-mean algorithm, a fast *k*-mean algorithm based on reference and density is proposed, this algorithm regards reference point as the first initial clustering center, the remaining initial clustering center is selected from the core points so that the initial clustering center can better respond geometric features of pattern sample set and can reduce the number of iterations.

Key words: reference-point; density; *k*- mean

责任编辑:代小红