

提前确定,数据分析困难时,比较难确定;2)聚类中心点的随机选取会影响聚类结果;3)采用欧式距离的方法适合球状簇的聚类分析.

### 3 算法改进

针对传统  $K$ -means 聚类算法的缺点聚类中心随机选取,提出改进办法如下:增加新变量密度的方法确立初始的聚类中心点,密度变量定义是以数据值为圆心  $r$  为半径数据对象的总个数<sup>[3]</sup>.

#### 3.1 定义

**定义 1** 领域半径  $Eps = Avg(X) = 1/(c\_n^2) \times \sum [d(x\_i, x\_j)]$ ,  $Avg(X)$  是数据集  $x$  中所有数据距离的平均值,  $c\_n^2$  是  $n$  中任意取两个的数目.

**定义 2** 点  $x\_i$  密度  $p(x\_i) = |x\_i| / d(x\_i, x\_j) \leq Eps$ , 公式含义是与  $x\_i$  之间的距离小于  $Eps$  的数据数.

**定义 3** 平均密度  $Ap(n) = (\sum_{i=1}^n p(x\_i)) / n$

选择  $K$  方法: 初始两点: 选距离最远的两数据.

确定第三个点: 与初始两点的最小距离中选择大的.

确定第四个点: 与已确定的三个点的最小距离中选择最大的. 重复一直选出  $K$  个点.

#### 3.2 改进后选择 $K$ 的算法步骤:

第一步: 输入  $n$  个数据.

第二步: 计算数据两两之间的距离  $d(x\_i, x\_j)$ , 组成一个  $n \times n$  矩阵.

第三步: 求出每个数据的点密度  $p(x\_i)$  和平均密度  $Ap(n)$ , 把所有点密度大于平均密度的数据归到新集合  $Y$  中, 作为聚类中心的参考点.

第四步: 在集合  $Y$  中选择密度最大的对象, 标记为  $y\_1$ , 放入集合  $Z$  中,  $Y$  中删去  $y\_1$ .

第五步: 在矩阵中找与  $y\_1$  距离最大的数据对象.

第六步: 判断此对象是否在集合  $Y$  中, 是则标记为  $y\_2$ , 放入集合  $Z$  中, 执行第七步, 不在, 将对应的值从矩阵中删去, 重复执行第五步.

第七步: 在矩阵中找与  $Z$  中数据对象最小距离中的最大的.

第八步: 根据上一步求出的数据对象判断是否在  $Y$  中, 是标记为  $y\_j$ , 放入集合  $Z$  中, 否则循环到第七步, 直到  $Z$  中个数为  $K$ .

第九步: 将集合  $Z$  中的数据作为初始聚类中心.

第十步: 聚类分析, 输出结果<sup>[4]</sup>.

### 4 实验分析

通过实验来对比传统  $K$ -means 算法和改进  $K$ -means 算法的聚类效果, 并对结果进行统计与分析, 下面的实验是先将学校校园网上的热点话题进行聚类然后对学校舆情进行预测和分析.

首先要进行数据采集, 借助 API 文本抓取工具, 在校园网上的论坛、贴吧、微博等系统上抓取数据, 接着对抓取的数据进行文本去噪后存入数据库, 然后利用 NLPIR 分词系统进行分词, 停用词过滤、词性过滤, 删去介词、连词、语气词之类的虚词只留下名词和动词, 预处理结束后, 为了将文本变为计算机可识别的格式, 用于输入算法中. 采用向量空间模型 vsm 的方法表示文本.

实验评价指标采用 TDP 评价标准:

准确率( $P$ ): 指正确归类的样本数据与全部样本数据的比例,  $P = TP / (TP + FP)$ .

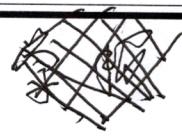
召回率( $R$ )指正确归类的样本数据在全部归类样本数据中所占比重,  $R = TP / (TP + FN)$ .

准确率和召回率加权调和平均( $F1$ ):  $F1 = 2PR / (P + R)$ .

$TP$ : 事实相关, 结果相关的文档.

$FP$ : 事实不相关, 结果为相关的文档.

$FN$ : 事实相关, 结果为不相关的文档.



# K-means聚类算法的改进与应用

刘建花

(晋中师范高等专科学校 数理科学系,山西 晋中 030600)

[摘要] K-means聚类算法具有实现简单、普及性强的优势,但存在聚类中心选取随意性强的劣势。文章提出增加一个密度变量的方式来选出合理的初始聚类中心,通过对校园网上热点话题聚类分析的实验,检验出改进K-means聚类算法聚类效果好。

[关键词] K-means 算法;密度;聚类中心

[文章编号] 1672-2027(2020)01-0081-03 [中图分类号] TP301 [文献标识码] A

## 1 问题提出背景

随着计算机技术和网络技术的不断发展,我们身边大量的数据不断生成,如何利用这些数据显得尤为重要,数据挖掘就是我们要利用的工具。数据挖掘也称知识发现,采用科学的方法或手段,为人们从数据库中找出对自己有益信息或感兴趣的信息提供帮助。聚类分析属于数据挖掘技术的一种,在模式识别、图像处理等领域都有广泛应用,当中的聚类算法也有很多,如 BRICCH, ROCK, DASCAN 算法等。

基于 K-means 聚类算法属于聚类分析中基于划分的一种。它具有简洁高效和收敛性好的特点。K-means 聚类算法中对初始聚类中心选择是非常重要的,选择不同的中心会造成完全不一样的聚类结果。此外在已有条件上分析确定聚类数目也很重要,事先给定的聚类数目同预期多多少少会有偏差,传统的 K-means 算法的选取中心点是随机的,如果选的不合适,有可能产生的是局部最优解,影响聚类正确性。为了弥补此缺陷,迫切需要对传统 K-means 聚类算法通过分析研究进行优化。

## 2 传统 K-means 聚类算法

### 2.1 算法思想

该算法的作用是解决聚类问题,把数据集的数据对象按照同一簇中的数据具有高相似度和与其他簇中的数据对象是低相似度划分条件将数据对象归类到不同的簇中<sup>[1]</sup>。

算法分以下几个步骤:

第一步:首先输入  $n$  个数据对象。

第二步:先确定  $K$ ,即聚类中簇的个数,然后从数据对象中随机选择聚类中心点  $K$  个。

第三步:计算其余数据对象与  $K$  个聚类中心点之间的距离,比较距离并将其归类到与之距离最近的聚类中心的簇当中。

第四步:通过计算,把  $K$  个聚类簇中所有数据对象的均值作为新的聚类中心。

第五步:判断是否满足收敛条件,满足则结束,输出结果,否则回到第三步循环<sup>[2]</sup>。

对 K-means 聚类算法而言,初始聚类中心的选择是非常重要的,选择不同的初始聚类中心会造成完全不一样的聚类结果。此外在已有条件上分析确定聚类数目也很重要,事先给定聚类数目同预期会有偏差,传统的 K-means 算法的选取初始聚类中心点是随机的,如果选的不合适,不仅会降低算法效率,而且会导致错误的结果。

### 2.2 算法优缺点

该算法优点:1)实现简单;2)效率高;3)输入数据的顺序不会对结果造成影响。缺点:1)聚类集群数目需

\* 收稿日期:2019-10-24

作者简介:刘建花(1982-),女,山西晋中人,晋中师范高等专科学校数理科学系讲师,主要从事数据库,数据挖掘研究。