文章编号：2096-1472(2019)-05-32-03
DOI:10.19644/j.cnki.issn2096-1472.2019.05.007

改进的k-means聚类算法在公交IC卡数据分析中的应用研究

杨健兵
(南通科技职业学院，江苏 南通 226007)

摘　要：针对传统k−means算法中初始聚类中心随机确定的问题，提出k−means改进算法。首先，定义变量权值，权值的大小等于样本密度乘以簇间距离除以簇内样本平均距离，通过最大权值来确定聚类中心，克服了随机确定聚类中心的不稳定性。然后在Hadoop平台上用Map−Reduce框架下实现算法的并行化。最后以南通公交IC刷卡记录为例，通过改进的k−means聚类算法进行IC卡刷卡记录的分析。实验表明，在Hadoop平台下改进k−means算法运行稳定、可靠，具有很好的聚类效果。

关键词：MapReduce；改进k−means算法；k−means；聚类
中图分类号：TP301　**文献标识码**：A

Study on the Application of Improved K-means Clustering Algorithm in the Data Analysis of Bus IC Cards

YANG Jianbing
(*Nantong Science and Technology College*,*Nantong* 226007,*China*)

Abstract: Aiming at the problem of random determination of initial clustering centers in traditional k-means algorithm,an improved k-means algorithm is proposed in this paper.First,the weight value of the variable is defined.The weight value is equal to the sample density multiplied by the distance between clusters and then divided by the average distance within the cluster.The clustering center is determined by the maximum weight,and the instability of the cluster center is determined randomly.Then the parallelization of the algorithm is implemented under the Map-Reduce framework on the Hadoop platform.Finally,taking the Nantong bus IC card record as an example,an improved k-means clustering algorithm is used to analyze the IC card record.Experiments show that the improved k-means algorithm is stable and reliable under the Hadoop platform,with a good clustering effect.

Keywords: MapReduce;improved k-means algorithm;k-means;clustering

1 引言(Introduction)

传统的公交客流调查大多数通过问卷调查获得，这种调查方法原始、相对落后、耗费大量的人力、物理和财力，并且最终获得的数据也不精确，往往为最终决策带来一定误差。而伴随着智能公共交通系统的发展和普及，公交IC卡收费系统、GPS监控系统、车辆监控系统中积累了大量原始的公交数据，特别是公交IC卡收费系统，里面保存在每位乘客的上车刷卡信息，这些海量的刷卡信息内部蕴含着真实、全面的公交客流信息[1,2]，如何利用数据挖掘技术从这些海量的公交IC卡数据中快速获取真实全面的公交客流信息，也是研究的热点问题。

最近几年，国内外学者在公交IC卡数据分析中做了大量的研究工作。在国外，Jinhua结合AFC及AVC数据获取上车站点，然而国外的城市公交系统与国内的相差很大。在国内，戴宵等[3]提出了对公交卡乘客的刷卡时间进行聚类分析判断乘客上车站点的方法，于勇等[4]结合公交运营调度时刻表所提供的车辆及其发车信息，推算各车次到达各站点的时间，提高了上车站点推算精度。周锐[5]提出了基于IC卡数据的公交站点客流推算方法。赵鹏[6]基于成都公交IC卡数据的乘客上下车站点推算方法研究。徐文远[7]等基于公交IC卡数据的公交客流统计方法。以上的研究存在数据不完整、准确率偏低等问题，所以研究的正确性很难得到保证。

本文针对公交IC卡中海量的刷卡数据，提出了基于hadoop平台的改进k−means算法，在底层HDFS文件系统的支持下，通过k−means算法对公交IC卡刷卡数据进行分析。利用MapReduce算法进行并行计算，通过MapReduce算法极

基金项目：本文系南通市科技资助项目“BP神经网络技术在智能公交IC卡中的应用研究”(项目编号：MS12017026−4).

大地聚类算法的效率，为公交公司制定合理的调度方案提供了重要的依据。

2　数据预处理(Data preprocessing)

本文需要进行计算的数据是南通市公共交通IC卡刷卡数据。公交IC卡刷卡数据字段包括运营公司、IC卡编号、刷卡时间、刷卡金额、卡类型、线路编号、IC卡设备编号和公交车辆编号等字段。在本文的研究过程中，选取IC卡数据的IC卡编号、IC卡类型、刷卡时间、线路编号四个字段属性。数据库表的格式如表1所示。

表1 公交刷卡数据库表

Tab.1 Bus card database table

IC卡编号	IC卡类型	刷卡时间	线路编号
0012369	A	2018-12-13 9:30:51	18

由于公交车在行驶过程中依次停靠公交各个站点，在停靠的过程中乘客依次上车刷卡，又由于公交IC卡刷卡消费数据所记录乘客刷卡时间具有一定的次序性，早上车的乘客刷卡时间早于后上车的乘客，其上车的站点顺序只有两种状况。

①第一，乘车站点相同。在这种情况下，该站点所有的乘客刷卡时间相差不大，相邻两位乘客间的刷卡间隔非常短，大概在几秒之间。该站点第一个上车乘客和该站点最有一个上车乘客刷卡时间差也不是很大，可以归属为同一类。

②第二，刷卡时间早的乘客上车时所在的站点位于刷卡时间晚的之前。在这种情况下，由于公交车从一个站点行驶到另外一个站点，所以相邻两个刷卡间隔比较长。

通过分析乘客刷卡记录，我们可以得出结论，相同站点乘车乘客，刷卡时间间隔较短，乘客在不同站点乘车，其刷卡时间间隔较长，这样可以通过对乘客刷卡记录进行聚类，使得相同站点的刷卡记录归于一类，不同站点的刷卡记录不在一类。

3　聚类算法(Clustering algorithm)

3.1　聚类算法和k-means聚类算法

聚类算法[8]是一种非监督机器学习算法，其实质就是对数据对象划分成子集的过程。聚类分析的算法有多种，可以分为划分法、层次法、基于密度的方法、基于网格的方法、基于模型的方法。k-means算法是属于划分方法中的一种，采用距离作为相似性的评价指标，该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

k-means算法把对象组织成多个互斥的组或簇，采用距离作为相似性的评价指标。假设数据集D包含n个欧式空间中的对象。聚类的目的是把D的对象分配到k个簇$C_1,\cdots,C_k$中，使得对于$1\leqslant i$，$j\leqslant k$，$C_i\in D$且$C_i\cap C_j=\varnothing$。聚类的划分的目的使得簇内高相似性和簇间低相似性为目标。

设数据集集合$D=\{x_1,x_2,\cdots,x_n\}$，$x_i=\{x_{i1},x_{i2},\cdots,x_{ir}\}$，$x_j=\{x_{j1},x_{j2},\cdots,x_{jr}\}$，则样本$x_i$和$x_j$之间的欧式距离为：

$$d(x_i,x_j)=\sqrt[2]{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ir}-x_{jr})^2} \quad (1)$$

误差函数平方和如下：

$$J_c=\sum_{i=1}^{k}\sum_{j=1}^{r^i}(x^j-n^i)^2 \quad (2)$$

其中，k为聚类数目，r^i是第i类样本的个数，n^i是i类样本的平均值。

k-mean均值的算法复杂度为$O(nkt)$，其中n是对象总数，k是用户指定的簇数，t为迭代次数。通常情况下$k<<n$并且$t<<n$。

k-means算法的优点是算法简单，易于实现，而且收敛速度快，计算工作很快就能完成。但是k-means算法也存在着一些缺点。

①分类结果依赖于分类中心的初始化。如果初始化点选择不是太好，容易陷入局部最优解。

②对于离群点数据敏感，离群点对其最终结果影响较大。

3.2　改进的k-means算法

改进的k-means算法在选取初始中心点的时候不采用随机选择的方式，而是通过计算数据集的密度来考虑初始中心点。与传统k-means算法随机选取聚类中心点的方法比，减少了选取数据中心点的随机性和盲目性。步骤如下：

两个样本之间的欧式距离$d(x_i,x_j)$按照(1)式来计算。

中所有样本间的平均距离AvgDist(D)按照式(3)计算数据集D。

$$\mathrm{AvgDist}(D)=\left.2\times\sum_{i=1}^{n}\sum_{j=i+1}^{n}d(x_i,x_j)\right/n\times(n-1) \quad (3)$$

计算数据集中样本i的密度。

$$\rho(i)=\sum_{j=1}^{n}f[d_{ij}-\mathrm{AvgDist}(D)] \quad (4)$$

其中，$f(x)=\begin{cases}1, x<0\\0, x\geq 0\end{cases}$。

$\rho(i)$是所有满足与样本i的距离小于平均距离AvgDist(D)的样本个数。这次样本定义一个簇，簇内平均距离为

$$a(i)=\left.2\times\sum_{i=1}^{\rho(i)}\sum_{j=i+1}^{\rho(i)}d(x_i,x_j)\right/\rho(i)\times(\rho(i)-1) \quad (5)$$

簇间距离$s(i)$定义为样本元素i与样本j之间的距离，样本j的局部密度稍高。若样本点i的局部密度为最大，则$s(i)$等于$\max\{d(i,j)\}$，若存在$\rho(j)>\rho(i)$，则$s(i)$为$\min\{d(i,j)\}$。

$$S(i)=\begin{cases}\min\{d(i,j)\}, \exists j,\ \rho(j)>\rho(i)\\ \max\{d(i,j)\}, \nexists j,\ \rho(j)>\rho(i)\end{cases} \quad (6)$$

定义权值的大小是数，的样本密度$\rho(i)$乘以簇间距离$s(i)$再除以簇内样本平均距离$a(i)$，即

$$W=\left.\rho(i)\times s(i)\right/a(i) \quad (7)$$

传统k-means算法聚类中心是随机选择，这种选择方法往往造成聚类的结果是局部最优，而非全局最优。本文提出基于改进的k-means算法，可以大大地降低这种选择的盲目选择聚类中心对聚类结果造成的非全局最优的问题，大大地

提高了聚类的准确性和稳定性。算法如下：

定义要聚类的数据集D，首先根据式(4)计算样本密度，第一个聚类中心是样本密度最大的一个元素，同时将所有满足式(3)中样本与初始聚类中心的距离小于AvgDist(D)条件的样本元素加入当前簇，并从集合D中删除这些样本，第二个聚类中心按照式(4)—式(7)计算余下元素权值w，找出最大值，并选取对应样本元素作为第二个聚类中心，重复进行，直到集合D为空集。

3.3 Hadoop平台下的改进的k-means算法实现过程

在hadoop程序开发中最重要的就是MapReduce程序的实现，MapReduce程序的开发分为map程序开发和reduce程序开发两个过程。MapReduce的程序设计与实现如图1所示。

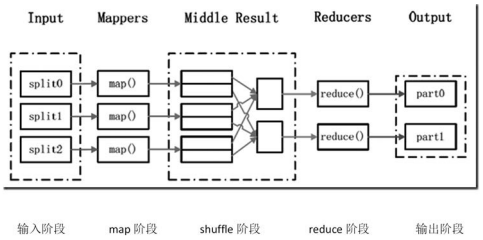

图1 MapReduce运行流程图

Fig.1 Flow chart of MapReduce

首先，将公交IC卡刷卡数据存储在Hadoop分布式文件系统中，然后通过MapReduce并行处理模型计算出K-means算法的输入参数，输入参数是初始聚类中心和k值，然后将计算任务再分配给Map任务节点，完成数据的并行聚类计算。具体步骤如下。

①对存储在HDFS中的IC卡刷卡数据进行初始化操作，产生<Key,Value>键值对。

②Map任务节点计算数据块中样本密度，并根据式(1)—式(7)，计算出最大权值w，并得到一些聚集，计算出每个聚类均值，并把该均值作为该簇的键值Key，Reduce算法根据键值key将具有相同Key值的簇集进行数据合并。

③重新计算出每个簇集的均值，并把计算的结果设置为Value的值，同时对key进行编号，key的号即为簇号。

④通过Map函数计算特征向量与k个初始聚类中心的欧氏距离，根据距离最小原则，找出其距离最小对应簇的簇号，从而得到更新的键值对〈Key，Value1〉。

⑤Reduce函数将每个分区中具有相同Key值的信息进行最后的合并。

⑥重复步骤④和步骤⑤，直到最终聚类结果的误差平方和达到稳定状态，并输出最终k个簇的相应信息。

4 实验结果(Experiment results)

4.1 实验环境

在本实验中，使用两台服务器搭建hadoop集群，每台机器CPU为Intel Xeon处理器，内存128GB。操作系统采用Centos7，搭建ambari大数据管理平台，包括一个master节点和一个slaver节点，来运行k-means和改进的k-means算法。

4.2 实验数据

实验数据来自南通公交2018年8月份南通公交某线路的刷卡数据，刷卡数据包括IC卡编号、IC卡类型、刷卡时间、线路编号等四个字段。

4.3 实验结果

本实验使用精度作为评价聚类性能的评价标准。通过对公交IC卡使用传统的k-means方法和改进的k-means方法进行分析，并计算其精确度，为了更好评价聚类性能，本实验共进行聚类五次。具体的分析如表2所示。

表2 数据集精度比较

Tab.2 Accuracy comparison of data sets

次数	传统的k-means算法	改进的k-means算法
1	0.75	0.81
2	0.68	0.82
3	0.85	0.81
4	0.65	0.82
5	0.59	0.80

5 结论(Conclusion)

本文以海量公交IC刷卡数据为基础，提出了一种在hadoop平台下改进的k-means算法，针对传统的k-means聚类算法中存在的问题，提出了采用一种采用密度参数的改进方法，在选取聚类中心的时候，充分考虑样本数据密度，同时定义了权值，权值的大小有样本密度乘以簇间距离然后除以簇内样本距离而得，通过最大权值来确定聚类初始中心和k值，提高了聚类的准确性和精确性。

参考文献(References)

[1] 孙慈嘉,李嘉伟,凌兴宏.基于云计算的公交OD矩阵构建方法[J].江苏大学学报(自然科学版),2016,37(4):456-461.

[2] 陈锋,刘剑锋.基于IC卡数据的公交客流特征分析——以北京市为例[J].城市交通,2016,14(1):51-58;64.

[3] 戴霄,陈学武,李文勇.公交IC卡信息处理的数据挖掘技术研究[J].交通与计算机,2006,24(01):40-42.

[4] 于勇,邓天民,肖裕民.一种新的公交乘客上车站点确定方法[J].重庆交通大学学报,2009,28(1):121-125.

[5] 周锐.基于IC卡数据的公交站点客流推算方法[D].北京交通大学,2012:27-38.

[6] 赵鹏基于成都公交IC卡数据的乘客上下车站点推算方法研究[D].西南交通大学,2012:16-31.

[7] 徐文远,邓春瑶,刘宝义.基于公交IC卡数据的公交客流统计方法[J].中国公路学报,2013,26(5):158-163.

[8] Jiawei Han,MichelineKamber,JianPei.数据挖掘概念与技术[M].北京:机械工业出版社,2012:4-5.

作者简介：

杨健兵(1976-)，男，硕士，讲师.研究领域：数据挖掘，云计算技术.