文章编号:1671-4598(2023)12-0284-06 DOI:10.16526/j.cnki.11-4762/tp.2023.12.042 中图分类号:TP301 文献标识码:A

Hadoop 平台下基于优化 X-means 算法的大数据聚类研究

张鹏飞[1], 江 岸[1], 熊 念[2]
(1. 广东农工商职业技术学院 计算机学院, 广州 510507;
2. 暨南大学 信息科学技术学院, 广州 510632)

摘要: 针对现有聚类方法对数据处理规模的局限性, 解决数据聚类效果差的问题, 在 Hadoop 平台的支持下提出基于优化 X-means 算法的大数据聚类方法; 利用 Hadoop 平台架构与函数采集大数据样本, 通过缺失补偿、噪声滤波、归一化等步骤, 实现初始样本数据的预处理; 选择大数据聚类中心, 分别提取聚类中心数据与其他所有数据样本的特征, 计算数据样本与聚类中心之间的特征相似度; 以相似度度量结果为聚类判定条件, 利用优化 X-means 算法确定数据所属类型, 最终实现大数据的聚类处理工作; 通过聚类效果测试实验得出结论: 在有、无两种实验条件下, 与传统聚类方法相比, 优化设计方法的查全率和查准率分别提升了 4.75%和 4.5%, 同时优化聚类方法得出数据具有更高利用率。

关键词: Hadoop 平台; 优化 X-means 算法; 大数据聚类

Research on Big Data Clustering Based on Optimized X-means Algorithm Under Hadoop Platform

ZHANG Pengfei[1], JIANG An[1], XIONG Nian[2]
(1. School of Computer, Guangdong Agriculture Industry Business Polytechnic, Guangzhou 510507, China;
2. School of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: In response to the limitations of existing clustering methods on data processing scale and poor performance of solving data clustering , a big data clustering method based on optimized X-means algorithm is proposed with the support of Hadoop platform. The Hadoop platform architecture and functions are used to collect the big data samples, and implement the preprocessing of the initial sample data is through the steps such as missing compensation, noise filtering, and normalization. The big data clustering center is selected to extract the features of the clustering center data and all other data samples respectively, and calculate the feature similarity between the data samples and the clustering center. Using similarity measurement results as the clustering criteria, the optimized X-means algorithm is used to determine the type of data, ultimately achieving the processing of big data clustering. Through the testing experiments of clustering effectiveness, it is concluded that compared to traditional clustering methods with or without two experimental conditions, the recall and precision of the optimized design method are improved by 4.75% and 4.5% respectively,. at the same time, the optimized clustering method has higher data utilization rate.

Keywords: Hadoop platform; optimize X-means algorithm; big data cluster

0 引言

目前数据库技术、数据库管理系统得到了快速的发展, 并得到了广泛的应用。与此同时, 随着数据采集工具的不断完善, 获取数据信息的方式也变得更加宽广, 所积累的数据知识也随之迅速增长。人们都想更好地使用这些数据, 并从这些数据中获取有用的信息和价值。但是, 在获取海量数据的过程中, 提取出有价值的信息变得更加困难。为发现数据源中未知的信息, 提高信息利用率, 提出大数据聚类方法[1]。数据聚类就是把数据按照其固有的属性划分成若干个聚合类, 每个聚合类中的元素尽可能具有相同的属性, 而不同的聚合类之间的属性差异也尽可能大。聚类分析的目标就是要对某一组中的数据进行分类, 从而使得某一组的成员之间既有相似之处, 又有不同之处。聚类分析是一种非指导性的、非监督的、非指导性的学习方法。通常采用聚类分析的方式, 将数据对象划分成多个均在内部关联的类别, 而在不同类别中的对象又有很大的差别。基于以上特点, 在很多实际问题中, 应用聚类方法对大数据进行处理, 聚类后可以把聚类中的各个数据对象看作是

收稿日期:2023-06-13; **修回日期**:2023-07-05。
基金项目:广东省普通高校重点领域专项(新一代信息技术)课题(2023ZDZX1068,2021ZDZX1138)。
作者简介:张鹏飞(1976-),女,硕士,副教授。
引用格式:张鹏飞,江 岸,熊 念. Hadoop 平台下基于优化 X-means 算法的大数据聚类研究[J]. 计算机测量与控制,2023,31(12):284-289,309.

一个整体来处理。

现阶段在诸多领域中应用较为频繁的大数据聚类方法研究成果包括：文献［1］提出的基于分组模型的引力搜索智能大数据聚类方法、文献［2］提出的基于核密度估计的 X-means 聚类方法以及文献［3］提出的云模式事件混沌关联特征提取的大数据聚类方法，其中文献［1］提出方法设计一种特定的解编码策略，即分组编码，可以把数据聚类之间的关联关系映射到相应的解决方案中，对于特定编码，新的引力搜索机制在位置和速度更新策略上设计适合分组编码的更新规则，使分组引力搜索可进行迭代寻优。文献［2］提出方法利用核密度估计的分布结果，对数据集进行密度偏差采样，之后对采样的样本集进行 K-means 聚类。而文献［3］提出方法综合利用了事件混沌关联特征提取算法、大数据关联挖掘算法、决策树学习算法、分析主成分算法等，得出平稳的物联网大数据聚类算法。然而上述传统聚类方法在实际运行过程中存在明显的聚类效果不佳的问题，主要体现在数据分类误差大、聚类数据丢失量大等问题，为此在 Hadoop 平台下，引入优化 X-means 算法。

Hadoop 为用户提供了一种分布式的管理技术，HDFS 是一种基于 MapReduce 的分布式文件系统，是 Hadoop 平台的核心技术。此外，它还为应用程序中的数据提供了很高的吞吐率，这对于拥有大量数据的应用程序来说是非常有用的。HDFS 对 POSIX 协议进行了较大程度的简化，使其能够在文件系统中对数据进行流式访问。使用 MapReduce 的编程模式，无需理解其背后的详细信息，就可以实现并行的应用程序。X-means 算法是动态分群法，是 K-means 算法的延伸，解决了 K-means 算法在运行过程中存在的迭代计算花费大、需要指定聚类中心和聚类数据量，且容易出现局部收敛的情况。以 Hadoop 平台为开发环境，利用优化 X-means 算法对大数据聚类方法进行优化设计，以期能够提升大数据的聚类效果，进而提高大数据的应用效率。

1　基于优化 X-means 算法的大数据聚类方法设计

在聚类分析中，数据的多源异构和数据的高速流动是聚类分析所面临的主要问题，其中所涉及到的数据具有结构化、非结构化、半结构化等多种特征，而且这些特征中存在着大量的不完整、冗余、甚至是错误数据，使得聚类分析的选择和处理变得非常困难。高速数据流对聚类挖掘算法提出了更高的需求，它能在多种约束下对海量信息进行高效的处理，减少对系统存储空间的占用，并提升对流数据的处理效率。大数据聚类方法大体可以分为 3 个步骤：首先准备待聚类处理的大数据，所有的输入数据都非常的复杂、且分散，因此大部分的数据挖掘工作都集中在数据的准备和发掘阶段。只有通过挖掘数据中的潜在价值，聚类分析才有实际意义。实际数据往往具有非一致性、不完备、噪音大、维数高等特点，在应用前必须对其进行预处理。数据预处理技术包括数据清洗、数据整合、数据转换、数据规范等。第二个步骤为特征选择和提取，即对原始数据集进行筛选，并从中抽取具有代表性的特征[2]。比如，在数据集中，某些特征对聚类结果的影响不大，某些特征间存在高度相关，可以通过对这些特征的选取和处理，来降低数据的维数，从而提升模型的质量。在此基础上，对这些有用的特性进行进一步的加工，以便更好地进行计算，比如标准化的数据。最终，根据实际业务需求和解决问题的特点，选取适当的相似度测度，根据相似度测量结果实现对大数据的聚类。在此次优化设计的大数据聚类方法以 Hadoop 平台中的数据作为聚类处理对象，在 K-means 算法的支持下，得出 X-means 算法的优化结果，并在优化 X-means 算法的支持下，完成聚类中心选取、K 值选择等步骤，从而提升优化大数据的聚类效果。

1.1　Hadoop 平台下采集聚类大数据

在分布式计算和分布式存储方面，Hadoop 都使用了主/从结构，它包括 4 个部分：Common 模块、HDFS 模块、MapReduce 计算框架和 Yarn 编程框架。HDFS 模块的主要作用是对 Hadoop 云计算平台中的数据信息进行存储，访问，管理和利用；MapReduce 是一个面向大数据的计算框架，其核心是映射和 Reduce。MapReduce 大数据计算架构可扩展性强，可兼容性强，在加入多个计算节点时，可同时承载所有节点的计算能力，可满足大规模数据的大规模处理需求。MapReduce 框架通过使用 Reduce 函数、映射函数对数据进行处理，有效地解决了数据细节冲突。Yarn 编程框架用来平衡各节点的负载[3]。利用 Hadoop 平台下的 HDFS 模块和 MapReduce 计算框架，通过文件读取的方式，得出待聚类的大数据样本。HDFS 和 MapReduce 计算框架如图 1 所示。

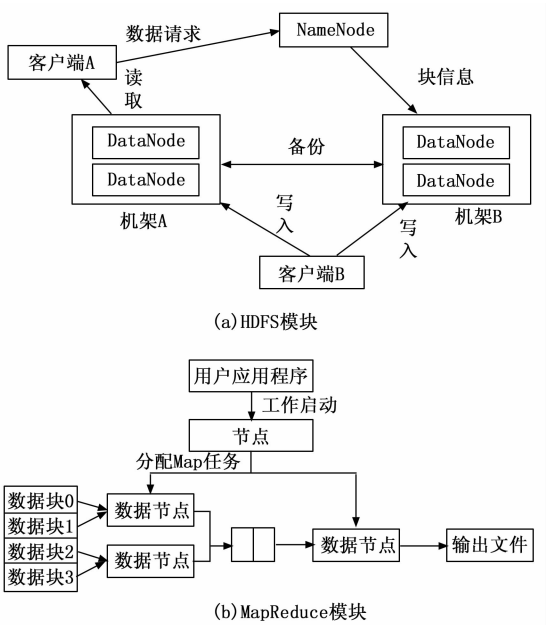

图 1　HDFS 和 MapReduce 计算框架图

在大数据采集过程中，当使用者第一次请求 HDFS 读取某一档案时，会使用打开方式，档案系统会得到档案名称、档案路径等基本资讯，并透过 RPC 机制与档案节点建立连结。这时，NameNode 会查询整个数据块的列表，并选择距离用户最近的一个块，将相应的 DataNode 返回给用户[4]。这个数据节点包含了一个系统中一个文件的相应的信息，以及这个文件中所有的块单位。再用 read 方法读取与用户请求的文件相对应的几个数据块，读完一个块单位后，对其进行校验，如果验证有错误，则会重新读取该块单位的副本，直至块单元校验成功为止。在读完文件内容之后，要用 close 函数来关闭分布式的数据输入流，释放被进程占用的计算和通信资源。Hadoop 平台中 HDFS 模块采集到的大数据结果为：

$$X = f_{\text{HDFS}} \times t_{\text{HDFS}} \tag{1}$$

其中：f_{HDFS}和 t_{HDFS}为大数据的采集频率和时间。在此基础上，采用 Map Reduce 方法，将原始收集到的大数据划分为多个大小相等的数据块，并在此基础上，通过设置参数，对每一个数据块进行备份。

1.2 大数据预处理与分布式存储

为保证大数据的聚类效果，降低干扰数据对聚类结果产生的负面影响，需要对初始采集的数据进行预处理。预处理步骤具体包括：缺失数据补偿、噪声数据过滤和数据归一化，其中缺失数据的补偿过程可以量化表示为：

$$x_{\text{compensate}}(i) = \frac{x(i-1)+x(i+1)}{2} \tag{2}$$

式中，$x(i)$ 为缺失数据，$x(i-1)$ 和 $x(i+1)$ 分别对应的是缺失数据的前后两个相邻数据[5]。为保证大数据的去噪质量，采用协同滤波的方式，即均值滤波与高斯滤波相结合的方式，处理初始采集数据，其中均值滤波是一种线性滤波算法，主要就是在数据集合中赋给目标数据一个模板，在利用模板中的全体数据的平均值来代替原本的数据值。初始数据的均值滤波处理结果为：

$$x_{\text{Mean filtering}} = \frac{\sum_{i=1}^{n_{\text{Big data}}} x(i)}{n_{\text{Big data}}} \tag{3}$$

其中：$n_{\text{Big data}}$为初始采集的待聚类数据量，另外数据高斯滤波的处理过程如下：

$$x_{\text{Gaussian filter}} = e^{-(x(i)^2)/2\sigma^2} \tag{4}$$

式（4）中，变量 σ 表示的是初始采集数据的标准差。在此基础上，对数据进行标准化处理，数据标准化处理结果为：

$$x_g(i) = \frac{x(i) - h_{\min}(X)}{h_{\max}(X) - h_{\min}(X)} \tag{5}$$

式中，$h_{\max}()$ 和 $h_{\min}()$ 分别为最大值和最小值求解函数，按照上述方式可以得出初始采集所有大数据的处理结果，并将预处理结果重新赋值给初始采集数据[6]。最终在 Hadoop 平台下采用分布式存储的方式将预处理完成的数据存储到 HDFS 模块中，完成数据的预处理与存储工作[7]。

1.3 大数据聚类中心

根据密度、距离、邻域等因素，选择初始中心点，使其既可以具有高密度，又可以在某种程度上反映出样本数据的分布情况[8]。首先，在一个数据集中寻找每一个数据对象中的每一个属性的极小值，并将其作为一个新的数据对象，标记为 u。利用公式（6）计算每个数据对象到 u 的距离。

$$d(i) = \sqrt{[x(i)-u]^2} \tag{6}$$

找出距离最大的点作为初始聚类中心，再通过迭代的方式，计算出每个数据对象到各聚类中心之间的距离，从而找到与聚类中心最近的、最远距离的数据对象，并将其作为新的聚类中心，直到聚类中心的数目与聚类的簇数目相等，这样反复进行，最终可以得到多个聚类中心，记为 C_i。

1.4 提取大数据特征

大数据的提取特征向量包括：数据量、数据变化量、数据分布密度等，通过初始大数据的挖掘，可以直接得出数据量特征向量的提取结果，另外数据变化量和数据分布密度特征的提取结果为：

$$\begin{cases} \Delta b = \dfrac{n(t_2)-n(t_1)}{t_2-t_1} \\ \rho = \dfrac{n_{\text{data}}}{W} \end{cases} \tag{7}$$

式（7）中，变量 $n(t_1)$ 和 $n(t_2)$ 对应的是 t_1 和 t_2 时刻采集的大数据样本，n_{data}为采集数据量，W 为数据的存储空间[9]。最终将提取的大数据特征进行融合，得出任意数据的综合特征提取结果，标记为 $\tau(i)$。上述数据特征主要针对的是普通数据，而对于文本、图像、音频等数据形式，还需要提取相应特征，并通过加权融合得出其他形式数据的综合特征提取结果。

1.5 计算大数据间特征相似度

大数据间相似度是大数据聚类的划分条件，定义大数据的相似度矩阵为：

$$S(x_{i,j}) = \begin{cases} -s(x(i),x(j)) \cdot n_{\text{Big data}}, i \neq j \\ (n_{\text{Big data}}-1) \cdot \bar{x} + \vartheta \end{cases} \tag{8}$$

其中：$s[x(i),x(j)]$ 为数据 $x(i)$ 和 $x(j)$ 之间的相似度度量值，$\bar{x}$ 为初始采集数据均值，ϑ 为聚类参考度指标，式（7）中变量 $s(x(i),x(j))$ 的计算公式如下：

$$s[x(i),x(j)] = \frac{\tau(i) \cdot \tau(j)}{\|\tau(i)\| \cdot \|\tau(j)\|} \tag{9}$$

将式（9）的计算结果代入到式（8）中，得出大数据集合中任意两个数据之间的相似度计算结果，并以相似度矩阵的形式输出。

1.6 利用优化 X-means 算法确定数据类型

以 Hadoop 平台作为 X-means 算法的运行支持，在 K-means 算法的基础上确定优化 X-means 算法的运行原理。优化 X-means 算法的运行与 K-means 算法基本一致，K-means 算法根据给定的聚类数 k，采用迭代更新算法对样本

进行分类[10]。在聚类过程中，采用了不同类别的样本之间具有较高的相似性，而不同类别之间具有较低的相似性，从而达到了更好的聚类效果。该方法的具体步骤为：在第一个迭代点上，通过随机选取 k 个初始簇中心；最后，对于剩下的目标，按照目标离每个聚类中心的远近程度，将目标聚类到最近的聚类中；最后，将每一簇中目标的平均作为新的聚类中心[11]。在后续迭代中，若连续两次出现的簇中心不一致，说明仍需要对所有的簇中心进行修正，并重新进行簇中心修正，重新进行下一次迭代；反之，则说明全部目标都已归类完毕，该算法完成。K-means 算法的目标函数为：

$$h_{\min}(e)=\sum_{i=1}^{K}\left|x(i)-C_i\right|^2 \tag{10}$$

式（10）中，变量 e 为大数据集合中所有数据对象的平方误差总和，K 为设置的聚类簇数[12]。针对 K-means 算法在运行过程中存在的局部收敛问题，得出优化 X-means 算法，该算法流程如图 2 所示。

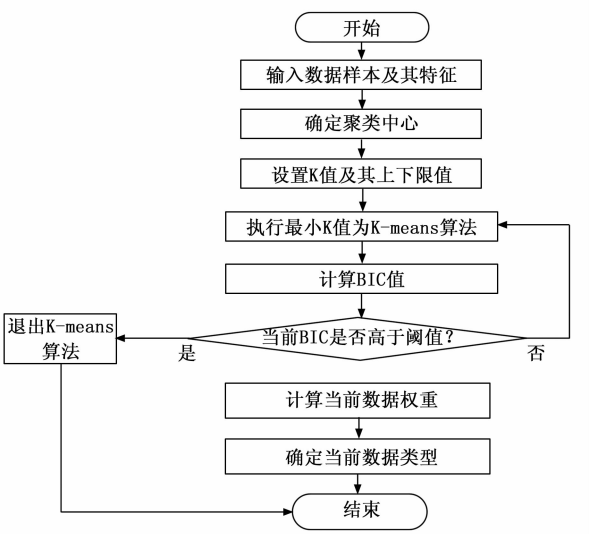

图 2　优化 X-means 算法执行流程图

在 X-means 算法中，只需给出聚类簇数 K 的范围即可完成聚类过程，设置 $K_{\max}$ 和 $K_{\min}$ 为聚类簇数 K 的上下限值，算法将在指定的范围内找到一个最优的聚类数 K 实现聚类划分[13]。根据贝叶斯信息量标准 BIC，X-means 聚类算法工作流程进行了评估，并对簇个数参数 K 进行了反复调整。其中 BIC 指标的计算公式如下：

$$BIC(\psi_{\text{X-means}})=\lambda-\frac{\delta}{2}\cdot\lg X_C \tag{11}$$

其中：$\psi_{\text{X-means}}$为聚类目标，λ 是根据聚类目标数据的似然对数渠道的最大似然点，δ 为 $\psi_{\text{X-means}}$的参数值，X_C 为聚类到 C 中的大数据。针对每一类聚类数据，分别构造分类器。每一个聚类分类器仅关注当前聚类的样本，既可以将大规模优化问题转换为小型优化问题，又可以减少聚类过程的复杂性，减少聚类过程的复杂性[14]。而待处理大数据样本，使用所有簇分类器的预测结果，并根据待聚类样本与各聚类中心之间的距离，动态地设定各聚类的预测权值。在一个数据空间聚类中，如果一个数据空间聚类中的样本和待聚类样本有更高的相似性，那么这个聚类就会有更高的分类可信度和更高的权重[15]。每个簇的权值通过式（12）进行归一化计算：

$$\omega_i=\frac{1}{d(x_i,C_i)}\cdot\left(\sum_{j=1}^{K_{\text{X-means}}}\frac{1}{d(x_i,C_i)}\right)^{-1} \tag{12}$$

其中：$d(x_i,C_i)$ 为大数据 x_i 到第 i 个聚类中心的距离，$K_{\text{X-means}}$为簇类数量。当 BIC 指标高于设置阈值时，执行 K-means 算法操作，得出大数据的分类结果，反复执行上述操作，直到簇数不再发生变化，或 $K_{\text{X-means}}$值高于 $K_{\max}$，直接输出数据类型的划分结果。

1.7　实现大数据聚类

以大数据与聚类中心之间的相似度度量结果为聚类程序的启动条件，在 Hadoop 平台下通过优化 X-means 算法的运行，得出大数据的聚类结果，如式（13）所示：

$$X_{\text{cluster}}=\frac{\sum_{i=1}^{n_{\text{cluster}}}\omega_i\cdot x(i)}{BIC(\psi_{\text{X-means}})},s(i,C_i)\geqslant s_0 \tag{13}$$

式中，变量 s_0 为相似度阈值[16]。将相关数据代入到式（13）中，即可得出以 C_i 为聚类中心的大数据聚类结果，如图 3 所示。

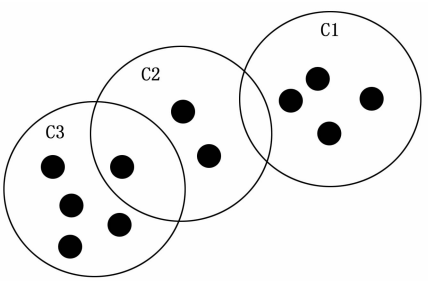

图 3　大数据聚类结果示意图

同理可以得出设置的所有聚类中心的大数据聚类结果，并以可视化的形式输出。

2　聚类效果测试实验分析

为了验证优化设计 Hadoop 平台下基于优化 X-means 算法的大数据聚类方法的聚类效果，设计测试实验。此次实验的主要采用白盒测试和对比测试相结合的方式，其中白盒测试原理就是在已知运行结果的情况下，判断优化设计方法输出结果与预设值之间的偏差。而对比方法则是设置实验对比项，在相同实验环境下，完成相同的运行任务，并通过横向对比，体现出优化设计方法在聚类效果方面的优势。综合上述两种测试方法的基本原理，此次聚类效果测试实验的基本思路为：在不同的数据集环境中，收集不同类型的数据样本，根据数据样本来源，得出数据聚类结果的预设值。采用数据随机置乱的方式，对初始设置的数据样本进行处理，以此作为实验的聚类对象。在此基础上，通过实验聚类方法与多种对比聚类方法的运行，得出相应

的聚类结果，最终通过多聚类质量指标的求解与对比，得出此次聚类效果测试实验的量化测试结果，并反映出优化设计方法是否达到预期效果。

2.1 搭建 Hadoop 平台

Hadoop 平台是基于优化 X-means 算法的大数据聚类方法的开发与运行环境，因此需要对 Hadoop 平台进行配置，Hadoop 集群中共有 5 台机器，其中一台被设置成了主测计算机。Hadoop 平台节点的具体部署情况如图 4 所示。

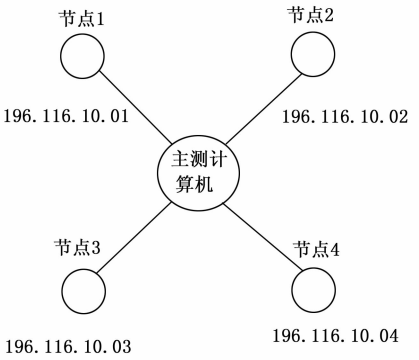

图 4 Hadoop 平台节点部署示意图

每台机器的内存为 4 G，每台机器都安装了 Cygwin 虚拟机仿拟 Linux 环境，安装 eclipse3.3.2，运行 Hadoop0.20.2，java 版本 1.6.0_43。在实际配置过程中，首先在每个虚拟机上复制一个 hadoop 的安装包，并配置下面 6 个主要的文件：1）hadoop-env.sh 文件的作用是配置 hadoop 的环境变量；2）core-site.xml 文件的作用是建立 hadoop 暂存目录、HDFS 路径等；3）hdfs-site.xml 文件的主要目的是为了建立一个 namespace、database、datablock 等等[17]；4）map-site.xml 文件主要用于将 mapReduce 设置为在 yarn 模式下工作；5）yarn-site.xml 文件主要用于对开始失败的自动恢复，资源管理器端口号，以及执行 MapReduce 所需的 shuffle 进程进行配置等；6）slaves 文件对数据节点名进行了配置。在完成了上述的配置文件之后，还需要对 NameNode 节点进行格式化。使用 hdfsnamenode-format 命令来进行操作，最后，使用 start-all.sh 命令来启动 Hadoop 集群。在启动成功之后，可以在浏览器中输入 Hadoop 平台地址信息来查看整个集群的状态[18]。将 Hadoop 设置为完全分布模式，即 Hadoop 集群中每个节点启动相应的进程，完成各自的工作。在 clipse 里建立一个 MapReduce 项目，然后把这个项目输入到 HDFS 里[19]。由于在安装和使用 Hadoop 集群时，必须取得所有虚拟机器的许可，所以可以设置 SSH 无密码登录，以便于启动或停止 Hadoop 群集。每个计算机都会使用 ssh-keygen-trsa 命令来产生一个公钥，并使用 cpid_rsa.pubauthorized_keys 命令来复制这个公钥到其它平台节点设备。

2.2 设定优化 X-means 算法运行参数

根据优化 X-means 算法的运行原理，在实验中需设置聚类簇数 K 的上限值为 150，下限值为 30，相似度阈值为 0.85，最大迭代次数为 100[20]。在实际的实验运行过程中，考虑聚类数据样本的准确情况，对优化 X-means 算法中的其他运行参数进行设定与调整。

2.3 准备待聚类大数据样本

选择多个不同行业的管理系统数据库中存储的数据，作为此次待聚类大数据样本的来源，数据类型具体包括计算机研发数据、军事数据、医药数据、教育数据、交通数据、市场调研数据、经济数据等，部分数据样本的准备情况，如表 1 所示。

表 1 待聚类大数据样本说明表

数据类型	数据总量/GB	属性类型	数据维度
计算机研发数据	37.50	6	7
军事数据	17.63	13	8
医药数据	25.85	8	6
教育数据	21.47	5	4
交通数据	30.58	3	3
市场调研数据	9.82	7	1
经济数据	17.15	4	7

初始准备大数据样本包含但不限于数据、文本、图像、影像、音频等形式，从表 1 中可以看出，此次实验逐步的大数据样本数量共 160 GB。采用随机置乱的方式，调整数据之间的分布关系。为验证优化设计方法是否能够适应多种不同类型的数据状态，在表 1 数据样本的基础上，添加一个数据干扰项，并生成有干扰条件下的数据样本。实验中使用的所有数据样本均存储在 MySQL 数据库中，样本数据可由注册计算机直接访问和调用。

2.4 描述聚类效果测试实验过程

为保证实验结果的可信度，将随机置乱后的数据样本划分成 8 个组别，保证每个组别中包含的数据类型不少于 3 种，实验组别的设置情况如表 2 所示。

表 2 实验组别设置表

实验组别	数据聚类数量	数据聚类类型
1	3	计算机研发数据、医药数据、教育数据
2	5	军事数据、医药数据、交通数据、市场调研数据、经济数据
3	5	计算机研发数据、军事数据、教育数据、市场调研数据、经济数据
4	4	医药数据、教育数据、交通数据、市场调研数据
5	6	计算机研发数据、军事数据、教育数据、交通数据、市场调研数据、经济数据
6	3	计算机研发数据、教育数据、交通数据
7	3	军事数据、医药数据、市场调研数据
8	4	军事数据、医药数据、教育数据、交通数据、市场调研数据

在组别设置过程中，需明确标注各组别中不同类型数据的数据量，以此作为判定优化设计方法聚类效果的比对标准。在配置好的 Hadoop 平台下，利用编程工具完成基于

优化 X-means 算法的大数据聚类方法的开发，并将优化 X-means 算法的运行参数以数据的形式导入到聚类方法的运行程序中。经过特征提取、相似度计算等步骤，得出大数据聚类输出结果。1 号组别的大数据聚类输出结果如图 5 所示。

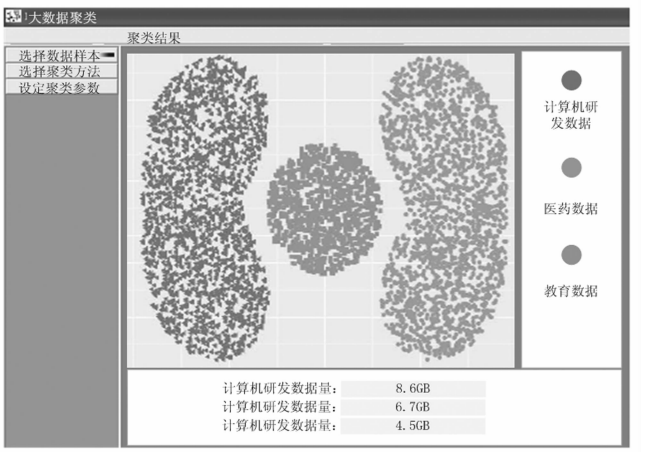

图 5　大数据聚类输出结果

重复上述操作可以得出实验中设置所有组别的聚类处理与输出结果。此次效果测试实验中，设置的对比项包括基于分组模型的引力搜索智能大数据聚类方法和基于核密度估计的 X-means 聚类方法，在相同的实验环境下，对上述对比方法进行开发与运行，重复上述操作得出相应的聚类结果。按照上述方式可以得出有、无干扰条件下的聚类结果。

2.5　设置聚类效果量化测试指标

此次实验分别从大数据聚类质量效果和应用效果两个方面进行测试，其中聚类质量效果的测试指标设置为查全率和查准率，其中查全率反映的是聚类方法是否能够将数据样本中属于任意类型的所有数据划分到数据集合中，而查准率则是判断聚类结果中是否存在分类错误的情况，聚类质量效果量化测试指标的数值结果如下：

$$\begin{cases}\eta_{\text{whole}}=\dfrac{Num_{\text{cluster}}}{Num_{\text{set}}}\times 100\%\\ \eta_{\text{accurate}}=\left(1-\dfrac{Num_{\text{error}}}{Num_{\text{cluster}}}\right)\times 100\%\end{cases}\tag{14}$$

式（14）中，变量 Num_{cluster}、Num_{set} 和 Num_{error} 分别为聚类结果数据量，设置聚类目标数据量以及聚类错误数据量，在上述变量计算过程中需要综合考虑聚类的多种类型，采用多类数据融合计算的方式，得出变量 Num_{cluster}、Num_{set} 和 Num_{error} 的具体取值。聚类的最终目的是实现大数据应用效率的提升，因此设置数据利用率作为聚类应用效果的量化测试指标，该指标的测试结果为：

$$\eta_{\text{use}}=\frac{N_{\text{application}}}{N_{\text{all}}}\times\varphi\times 100\%\tag{15}$$

式中，$N_{\text{application}}$ 和 N_{all} 分别为应用数据量和数据样本总量，φ 为数据应用次数。最终计算得出聚类数据查全率和查准率越高、数据利用率取值越大，说明对应方法的聚类效果越优。

2.6　聚类效果测试实验结果与分析

2.6.1　无干扰条件下的聚类质量效果

在无干扰条件下，通过相关数据的统计，得出反映聚类方法质量效果的测试结果，如表 3 所示。

表 3　无干扰条件下聚类质量效果测试数据表

实验组别	Num_{set}/GB	基于分组模型的引力搜索智能大数据聚类方法		基于核密度估计的 X-means 聚类方法		基于优化 X-means 算法的大数据聚类方法	
		Num_{cluster}/GB	Num_{error}/GB	Num_{cluster}/GB	Num_{error}/GB	Num_{cluster}/GB	Num_{error}/GB
1	20	18.8	1.1	19.4	0.4	19.8	0.2
2	20	18.6	1.2	19.5	0.4	19.9	0.1
3	20	19.3	0.8	19.7	0.3	20	0.2
4	20	19.0	0.9	19.6	0.9	19.9	0.1
5	20	18.4	0.9	19.3	0.7	20	0.2
6	20	19.1	0.7	19.8	0.6	20	0.1
7	20	18.5	1.3	19.4	0.2	19.8	0.1
8	20	19.5	1.1	19.5	0.8	19.9	0.2

将表 3 中数据代入到式（14）中，计算得出两种对比方法的平均查全率分别为 94.5%和 97.6%，平均查准率分别为 94.7%和 97.2%，另外优化设计方法查全率和查准率的平均值分别为 99.6%和 99.2%。

2.6.2　有干扰条件下的聚类质量效果

在有干扰条件下，重复上述操作，得出反映三种方法聚类质量的测试结果，如表 4 所示。

表 4　有干扰条件下聚类质量效果测试数据表

实验组别	Num_{set}/GB	基于分组模型的引力搜索智能大数据聚类方法		基于核密度估计的 X-means 聚类方法		基于优化 X-means 算法的大数据聚类方法	
		Num_{cluster}/GB	Num_{error}/GB	Num_{cluster}/GB	Num_{error}/GB	Num_{cluster}/GB	Num_{error}/GB
1	20	18.0	1.7	18.8	0.5	19.8	0.2
2	20	18.3	1.1	19.3	0.8	19.9	0.1
3	20	18.2	1.9	19.0	0.9	19.9	0.2
4	20	18.6	1.3	19.1	1.7	19.9	0.2
5	20	18.3	1.5	19.4	1.6	20	0.2
6	20	18.1	1.2	18.6	0.6	19.9	0.1
7	20	18.4	1.4	18.9	0.8	19.8	0.2
8	20	18.5	1.4	19.5	1.4	19.9	0.2

通过式（14）的计算，得出三种聚类方法的查全率平均值分别为 91.5%、95.4%和 99.4%，而平均查准率的测试结果分别为 92.1%、94.6%和 99.1%。

2.6.3　聚类应用效果

综合有、无两种实验条件，通过式（15）的计算，得出三种聚类方法应用效果的测试结果，如图 6 所示。

（下转第 309 页）