(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104380282 B
(45)授权公告日 2017.05.17

(21)申请号 201380033471.6
(22)申请日 2013.07.11
(65)同一申请的已公布的文献号
申请公布号 CN 104380282 A
(43)申请公布日 2015.02.25
(30)优先权数据
10-2012-0097498 2012.09.04 KR
(85)PCT国际申请进入国家阶段日
2014.12.24
(86)PCT国际申请的申请数据
PCT/KR2013/006190 2013.07.11
(87)PCT国际申请的公布数据
WO2014/038781 KO 2014.03.13
(73)专利权人 SK 普兰尼特有限公司
地址 韩国京畿道
(72)发明人 金民成 尹度永 李採炫 李俊燮
(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 吕俊刚 刘久亮
(51)Int.Cl.
G06F 17/00(2006.01)
(56)对比文件
US 6115708 A,2000.09.05,
US 6115708 A,2000.09.05,
US 2005222972 A1,2005.10.06,
US 2005049826 A1,2005.03.03,
吉根林 等.一种基于集成学习的分布式聚类算法.《东南大学学报》.2007,第37卷(第4期),
审查员 宋晓毓

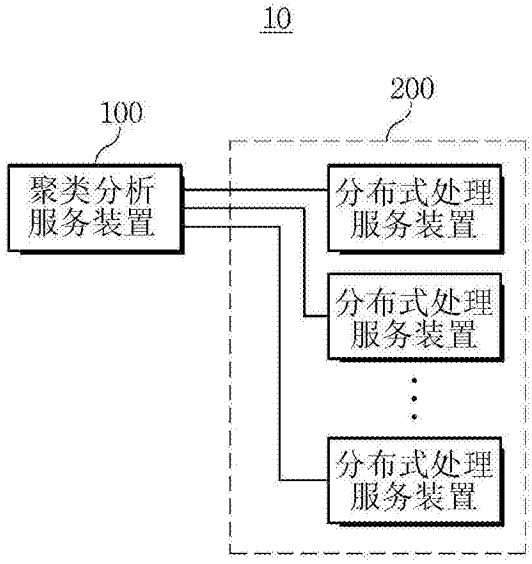
权利要求书2页 说明书12页 附图4页

(54)发明名称

聚类化支持系统和方法以及支持该方法的装置

(57)摘要

本发明涉及聚类化功能支持。公开聚类化支持系统、操作聚类化支持系统的方法和支持该方法的装置，聚类分析支持系统包括：聚类化服务装置，该聚类服务装置被构造成请求分布式处理服务装置基于预定范围内的k值和预设重复数执行k-均值聚类化直至满足预定义的收敛条件为止，如果由分布式处理服务装置计算出k值的中心值，则选择中心值中的最优中心值，通过关于将基于所选择的最优中心值许可的聚类索引的数据应用的指数计算控制最优k值的计算和应用；以及分布式处理服务装置，分布式处理服务装置被构造成根据聚类服务装置的请求基于从聚类服务装置中提供的k值和预设重复数执行k-均值聚类化，计算k值的中心值，向聚类化服务装置提供中心值。



CN 104380282 B

1.一种支持聚类分析的系统，该系统包括：

聚类分析服务装置，该聚类分析服务装置被构造成通过同时执行与针对用于聚类分析而提供的预定范围内的各个k值的预设迭代频率相对应的k-均值的聚类分析来选择对应于所述预定范围内的各个所述k值的中心值中的最优中心值，并且通过关于将基于所选择的最优中心值指派的聚类索引应用于数据的索引计算，来确定所述预定范围内的所述k值中的最优k值；以及

至少一个分布式处理服务装置，所述至少一个分布式处理服务装置包括至少一个数据节点，其中，所述至少一个数据节点被构造成在所述聚类分析服务装置的请求下向所述聚类分析服务装置提供通过同时执行与针对所述预定范围内的各个所述k值的预设迭代频率相对应的k-均值的所述聚类分析而得到的中心值，并且向所述聚类分析服务装置提供关于所述数据的所述聚类索引。

2.一种支持聚类分析的聚类分析服务装置，该聚类分析服务装置包括：

装置存储单元，该装置存储单元被构造成存储数据；

装置输入单元，该装置输入单元被构造成生成输入信号，该输入信号与针对所存储数据的聚类分析而提供的预定范围内的k值、收敛条件以及迭代频率有关；以及

装置控制单元，该装置控制单元被构造成通过同时执行与针对各个所述k值的预设迭代频率相对应的k-均值的聚类分析来选择对应于各个所述k值的最优中心值，并且通过关于将基于所选择的最优中心值指派的聚类索引应用于数据的索引计算，来确定所述预定范围内的所述k值中的最优k值。

3.根据权利要求2所述的聚类分析服务装置，其中，所述装置存储单元存储先前计算的k值。

4.根据权利要求3所述的聚类分析服务装置，其中，所述装置控制单元包括：

数据分配单元，该数据分配单元被构造成分配数据使得与针对各个所述k值的预设迭代频率相对应的k-均值的聚类分析被同时执行，

分析结果选择单元，该分析结果选择单元被构造成选择最优中心值；

分析索引应用单元，该分析索引应用单元被构造成执行关于被指派了聚类索引的数据的k值效率的索引计算，所述聚类索引是通过将具有所述最优中心值的选择的结果应用到所述数据而得到的；以及

最优值更新单元，该最优值更新单元被构造成基于具有所述索引计算的最优结果的k值更新先前存储的k值。

5.根据权利要求2所述的聚类分析服务装置，其中，所述装置控制单元被构造成多次同时自动执行针对多个k值的中心值的结果计算并且在每次计算具有不同的初始值。

6.一种用于支持聚类分析的方法，该方法包括以下步骤：

由聚类分析服务装置基于针对执行聚类分析而输入的预定范围内的各个k值的预设迭代频率以分布式方式向数据节点发送数据；

由所述数据节点执行与各个所述k值相对应的k-均值的聚类分析；

由所述数据节点基于所述聚类分析的结果选择对应于各个所述k值的中心值，并且向所述聚类分析服务装置提供所选择的中心值；

由所述聚类分析服务装置选择所述中心值中的最优中心值并且与所述数据节点共享

所选择的最优中心值；

由所述数据节点向数据指派通过应用所选择的最优中心值而获得的聚类索引；

由所述聚类分析服务装置对被指派了所述聚类索引的数据执行索引计算；以及

由所述聚类分析服务装置基于所述索引计算的结果确定所述预定范围内的所述k值中的最优k值。

7.根据权利要求6所述的方法，其中，在执行所述索引计算的步骤中，所述聚类分析服务装置基于根据采样条件执行索引计算来计算数据，并且基于所计算出的数据执行所述索引计算。

8.根据权利要求6所述的方法，所述方法进一步包括以下步骤：由所述聚类分析服务装置，对被指派了由所述数据节点提供的所述聚类索引的所述数据执行采样。

9.根据权利要求6所述的方法，其中，进行索引计算的步骤包括：

针对每个k值进行聚类索引的计算；以及

选择具有最高聚类索引的k值。

10.根据权利要求9所述的方法，其中，进行索引计算的步骤进一步包括：将多个索引方法应用到针对每个k值的聚类索引的计算中，以从所述多个索引方法中选择相对较高的k值。

11.根据权利要求6所述的方法，其中，在进行k-均值的聚类分析的步骤中，多次同时自动计算出针对多个k值的k-均值聚类化的结果并且在每次计算具有不同的初始值。

聚类化支持系统和方法以及支持该方法的装置

技术领域

[0001] 本发明涉及聚类分析，更具体地，涉及用于支持K-均值聚类化以在分布式处理环境中处理大数据的聚类分析支持系统和方法，以及支持该聚类分析的装置。

背景技术

[0002] 聚类分析，也就是说，聚类化是指对相似数据进行分组。数据是否相似随提前给定的相似性的定义而变化。当各个数据的值被表示为矢量时，主要用几何距离来确定相似性。用来确定相似性的几何距离的一个最具有代表性的示例是欧几里得距离(Euclidean distance)。同时，k-均值聚类化(k-means clustering)是用于将总共n个d维数据分组成k个组。例如，当二维输入数据存在时，k-均值聚类化表示向各个二维输入数据指派范围为从1到k的聚类索引的任务。

[0003] 当使用这种k-均值聚类化时，k直接由用户确定，并且聚类化的结果可以依赖于k而显著变化。因此，在没有关于k值的先验信息或知识的情况下随机地确定该k值，因此要确定k值是非常困难的，并且k值的错误确定可能会导致不希望的结果。由于k-均值聚类化是迭代算法，所以大的n(表示数据的数量)，或者数据的维度的高阶的d可能需要大量的执行时间。即使对于相同的k值，依赖于最初确定的中心(center)值，花费用来收敛的时间，即整个运行时间可以改变或者结果可以改变。这样，传统的k-均值聚类化的效率随k值输入而不同，并且因此其不容易一般化并且需要熟练的操作员控制，并且即使是熟练的操作员，不能连续提供恒定结果的可能性也很高。

发明内容

[0004] 技术问题

[0005] 本发明旨在提供一种能够以稳定的方式提供合适的聚类化效率的聚类分析支持系统和方法，以及支持该聚类分析的装置。

[0006] 具体地，本发明旨在提供一种能够在通过利用适合于分布式环境的数据结构而使k-均值聚类化自动化的同时执行高效k-均值聚类化的聚类分析支持系统和方法。

[0007] 技术方案

[0008] 本发明的一个方面提供一种支持聚类分析的系统，该系统包括聚类分析服务装置以及分布式处理服务装置。该聚类分析服务装置可以被构造成请求分布式处理服务装置基于在预定范围内的k值和预设迭代频率执行k-均值聚类化直至满足预定义的收敛条件为止，并且如果从分布式处理服务装置计算出k值的中心值，则选择所述中心值中的最优中心值，并且通过关于将基于所选择的最优中心值指派的聚类索引应用于数据的索引计算，来控制最优k值的计算和应用。分布式处理服务装置可以被构造成在就聚类分析服务装置的请求下从聚类分析服务装置提供的k值和预设迭代频率执行k-均值聚类化，并且如果k值的中心值被计算出，则向聚类化分析服务装置提供所述中心值。

[0009] 本发明的另一个方面提供一种用于支持聚类分析的聚类分析服务装置，该聚类分

析服务装置包括装置存储单元、装置输入单元以及装置控制单元。该装置存储单元可以被构造成存储数据。该装置输入单元可以被构造成生成输入信号，该输入信号与针对所存储数据的聚类分析而提供的预定范围内的k值、收敛条件以及迭代频率中的至少一方有关。该装置控制单元可以被构造成控制使得基于k值和迭代频率执行k-均值聚类化并计算k值的中心值直至满足收敛条件为止，其中，每当数据被更新时执行所述k-均值聚类化。

[0010] 装置存储单元可以存储根据先前的k-均值聚类计算出的先前k值。

[0011] 装置控制单元可以包括数据分配单元、分析结果选择单元、分析索引应用单元以及最优值更新单元。该数据分配单元可以被构造成分配数据使得对所述数据进行分布式处理。分析结果选择单元可以被构造成如果k值的中心值被计算出，则选择所计算出的中心值中的最优中心值。分析索引应用单元可以被构造成执行关于被指派了聚类索引的数据的k值效率的索引计算，所述聚类索引是通过将具有所述最优中心值的选择的结果应用到所述数据而得到的。该最优值更新单元可以被构造成基于具有索引计算的最优结果的k值更新先前存储的k值。

[0012] 装置控制单元可以被构造成多次同时自动执行针对多个k值的中心值的结果计算并且在每次计算具有不同的初始值。

[0013] 本发明的另一个方面提供一种支持聚类分析的方法，该方法包括以下步骤：由聚类分析服务装置基于所输入的预定范围内的k值以及迭代频率执行中心值矢量的初始化；以分布式的方式向分布式处理服务装置的数据节点发送数据和初始化后的中心值矢量；由所述数据节点基于以分布式的方式发送的所述数据和所述初始化后的中心值矢量执行k-均值聚类化；由所述数据节点向所述聚类分析服务装置提供分析的结果；由所述聚类分析服务装置选择所述分析的所述结果中的最优结果并与所述数据节点共享所选择的结果；由所述数据节点向数据指派通过应用所选择的结果而获得的聚类索引；由所述聚类分析服务装置对被指派了聚类索引的数据执行索引计算；以及使用具有索引计算的最优结果的值更新先前存储的k值。

[0014] 所述方法可以进一步包括以下步骤：由所述聚类分析服务装置向所述数据节点发送采样条件；以及由所述数据节点根据所述采样条件计算数据；以及将计算出的数据提供到所述聚类分析服务装置。

[0015] 所述方法可以进一步包括以下步骤：由所述聚类分析服务装置，对被指派了由所述数据节点提供的所述聚类索引的所述数据执行采样。

[0016] 进行索引计算的步骤可以包括：针对各个k值进行聚类索引的计算；以及选择具有最高聚类索引的k值。

[0017] 进行索引计算的步骤可以进一步包括：将多个索引方法应用到针对每个k值的聚类索引的计算中，以从所述多个索引方法中选择相对较高的k值。

[0018] 在进行k-均值聚类化的步骤中，可以通过多次同时自动计算出针对多个k值的k-均值聚类化的结果并且在每次计算具有不同的初始值。

[0019] 本发明的另一个方面提供一种记录执行聚类分析支持方法的程序的计算机可读记录介质。

[0020] 有益效果

[0021] 正如从所述聚类分析支持系统和方法以及支持该聚类分析的装置中明显的，计算

出具有合适聚类化结果的k值，并且与用户是否进行了输入无关，在k值的预定范围中计算合适的k值，从而支持稳定的分析效率。

[0022] 因此，本发明提供一种具有高可靠性的聚类化数据。

附图说明

[0023] 图1是例示根据本发明的示例性实施方式的聚类分析支持系统的构造的图。

[0024] 图2是例示根据本发明的示例性实施方式的聚类分析服务装置的构造的详细图。

[0025] 图3是例示根据本发明的中心值的示例的图。

[0026] 图4是例示图2的装置控制单元的详细图。

[0027] 图5是例示最优中心值的选择的图。

[0028] 图6是例示聚类分析索引的计算的示例的图。

[0029] 图7是例示根据本发明的分布式处理服务装置的数据节点的构造的示例的图。

[0030] 图8是例示根据本发明的示例性实施方式的聚类分析支持方法的图。

具体实施方式

[0031] 下面将详细描述本发明的示例性实施方式。贯穿本说明书使用相同的标号来表示相同的元件。在实施方式的说明中，为避免造成本发明的主题模糊，这里将省略相关已知功能或结构的详细说明。

[0032] 同时，在说明书及其附图中提出的示例性实施方式仅仅是提供用来完善本发明并帮助本领域技术人员完全理解本发明，因此本发明的范围不受这些实施方式和术语的限制。因此，对于本领域技术人员而言很明显，在不背离本发明的范围的条件下，可以进行各种示例性实施方式。

[0033] 图1是例示根据本发明的示例性实施方式的聚类分析支持系统的构造的图。

[0034] 参照图1，聚类分析支持系统10包括至少一个聚类分析服务装置100和分布式处理服务装置200，该分布式处理服务装置200被构造成为所述聚类分析服务装置100提供项目搜索功能和购买功能，并且可以进还包括用于在所述聚类分析服务装置100和所述分布式处理服务装置200之间进行通信连接的通信网络(未示出)。

[0035] 根据本发明的聚类分析支持系统10被设置成使得聚类分析服务装置100向分布式处理服务装置200的各个数据节点的基于映射化简的映射器(MapReduce-based Mapper)发送收集到的数据和预定范围内的k值，并且每个数据节点对所发送的数据执行对应于该预定范围内的k值和迭代号码的聚类分析直至满足预定义的收敛条件为止，从而提供最小测量值。

[0036] 聚类分析支持系统10被设置成如果提供针对预定范围内的相应k值的最小测量值，则在所述最小测量值中计算出最优中心值，并且将该最优中心值应用到原始数据，从而支持各个数据的聚类分析。另外，聚类分析支持系统10被设置成设置索引应用来测试聚类分析的应用效率，并且通过使用具有最有效索引值的k值来执行k值更新。如上所述，根据本发明的聚类分析支持系统10被设置成同时处理针对预定范围内的k值的聚类分析，并且对处理后的结果进行互相比较，从而以非常快速且精确的方式获得合适的聚类分析。具体地，本发明被设置成在执行聚类分析的过程中，相对于预定范围内的k值反复地且自动地执行

聚类分析，从而与用户是否进行了额外的输入无关，通过根据数据更新的自动聚类分析来支持k值计算和应用。因此，根据本发明的聚类分析支持系统10可以基于稳定可靠的聚类分析支持各种数据应用。

[0037] 为此，通信网络(未示出)被设置以在聚类分析服务装置100和分布式处理服务装置200之间形成有线/无线通信信道。也就是说，通信网络(未示出)可以支持聚类分析服务装置100和分布式处理服务装置200之间的数据传输的信号发送/接收、聚类分析结果的传递以及聚类分析索引的应用结果的传输。具体地，通信网络(未示出)将关于数据和预定范围内的k值的信息从聚类分析服务装置100发送到分布式处理服务装置200，并且将聚类分析的结果从分布式处理服务装置200发送到聚类分析服务装置100。另外，通信网络(未示出)可以将聚类分析服务装置100选择的特定k值发送到分布式处理服务装置200以被应用，并且允许分布式处理服务装置200通过使用所选择的k值将预定索引应用到针对每个数据的聚类分析。

[0038] 聚类分析服务装置100通过使用装置通信单元与通信网络(未示出)连接，并因此接入与该通信网络(未示出)连接的分布式处理服务装置200。聚类分析服务装置100可以向分布式处理服务装置200提供数据和预定范围内的k值。预定范围内的k值可以是在用户输入的预定范围内的自然数。在数据被发送之前，聚类分析服务装置100可以与分布式处理服务装置200的数据节点共享在预定范围内的输入的k值。聚类分析服务装置100向分布式处理服务装置200提供收集到的数据。

[0039] 同时，在从分布式处理服务装置200接收到通过向各个数据应用k值而得到的输出时，聚类分析服务装置100可以在接收到的输出中选择具有最小测量值的中心值信息。聚类分析服务装置100可以请求分布式处理服务装置200将所选择的中心值信息应用到原始数据。如果随着分布式处理服务装置200将所选择的中心值信息应用到原始数据聚类索引被给出，则聚类分析服务装置100可以执行聚类分析索引计算，以检查哪个k值产生最有效的聚类分析。聚类分析服务装置100可以基于聚类分析索引计算来控制k值更新。下面将参照图2详细描述聚类分析服务装置100。

[0040] 分布式处理服务装置200包括多个数据节点，同时连接到聚类分析服务装置100，以基于由所述聚类分析服务装置100提供的数据和预定范围内的k值执行聚类分析。分布式处理服务装置200可以向聚类分析服务装置100提供聚类分析的结果。然后，分布式处理服务装置200可以通过将由聚类分析服务装置100提供的所选k值应用到原始数据来执行聚类分析。

[0041] 根据本发明的聚类分析支持系统10被设置成使得预定范围内的k值在聚类分析期间被同时应用到数据，从而容易地发现具有合适测量值的k值，并基于所发现的k值执行有效的聚类分析。

[0042] 图2是例示根据本发明的示例性实施方式的聚类分析服务装置100的构造的详细图。

[0043] 参照图2，根据本发明的聚类分析服务装置100包括装置通信单元110、装置输入单元120、装置存储单元150以及装置控制单元160。用于参考，应用于根据本发明的聚类分析的k-均值聚类化属于当正确答案不存在时应用的无监督学习方法组，因此精确度和正确答案不能被比较。当提供总共n个d维矢量作为输入时，k-均值聚类化计算出k个中心，总共n个

数据被按照该k个中心分成k个聚类。该计算被迭代地执行，并且找到使下面的算式1最小化的中心。

[0044] 【算式1】

[0045] $$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

[0046] 这里，S_i为各个聚类的索引，并且x_j为第j个输入。μ_i是第i个中心(i＝1,2,…,k)根据下面的算式2更新针对每次迭代的中心。每个输入数据被处理为属于最近的中心，并且新的中心被计算作为属于该新中心的输入数据的平均。

[0047] 【算式2】

[0048] $$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$$

[0049] 在算式2中，$m_i^{(t+1)}$表示第(t+1)个步骤中的第i个中心值。$\left|S_i^{(t)}\right|$表示属于在第t次迭代中被称为‘i’的聚类的数据的数量。继续k-均值聚类化，同时迭代地改变中心值，直至迭代中心值的变化低于预定阀值或者测量值的变化低于预定阀值达预定次数为止。

[0050] 装置存储单元150被构造成存储需要用来操作聚类分析服务装置100的各种程序和数据。具体地，装置存储单元150可以存储被应用k-均值聚类化的整个数据以及通过聚类分析计算出的先前的k值(k_prev)153。可以通过利用基于映射化简的映射器(MapReduce-based Mapper)向分布式处理服务装置200部分地提供数据151。存储在装置存储单元150中的数据151可以被以预定的格式提供，以支持聚类分析。为了使k-均值聚类化自动化，用户可以指定预定范围内的k值，并且对整个k范围内的每个k值以预定的次数同时执行k-均值聚类化分析，并且导出最优值k。更详细地，对整个k范围内的每个k导出最优k值，并且通过索引计算导出最优值k。例如，k被指定为k＝1～5，并且同时执行对k＝1，k＝2，…和k＝5的聚类化分析，或者同时得到对特定k的一定次数的聚类化分析的结果，然后得到最优结果。

[0051] 为此，根据本发明的特定聚类的中心值的索引被限定为三个值的组合。第一列是k-集合ID，其标识与中心值相对应的k值。第二列是NCS(候选集合的数量)ID，通过该NCS ID，对特定k值同时执行k-均值聚类化分析。从对同一k值通过NCS的数量同时执行的分析的结果得到的最小的(最优)测量值被生成作为有代表性的结果值。第三列是中心ID，其为表示在总共k个中心中对特定k值和特定NCS ID的中心值的等级的值。

[0052] 每个中心ID可以依赖于当前中心所属的k-集合ID而具有与k-集合ID相对应的从1到k范围内的值。结果，可以以k-集合ID|NCS ID|中心ID的形式提供存储在装置存储单元150中的数据151中包括的三个值。然而，本发明并不限于此，并且数据151可以被以任何其他数据结构提供，其中上述三个值被区分并且基于这三个值执行数据处理。

[0053] 同时，与传统方法相类似地，用户可以指定单个值作为k值。另选地，用户可以通过指定最小k值和最大k值来指定k值的范围。另选地，用户可以指定期望得到其结果的k值的集合。例如，用户可以指定k值，k＝1、3、4和8。为此，聚类分析服务装置100可以进一步包括提供用于指定k值的屏幕的装置显示单元以及提供装置输入功能的装置输入单元120。

[0054] 例如，当可以作为存储在装置存储单元150中的数据151的示例的特定数据具有键值4|2|3时，该键指示用于针对对k＝4的聚类分析生成第二结果(NCS＝2)的中心值中的第三中心(第三聚类)。当k-集合ID的范围被指定为具有为3的最小k，最大k为5且NCS＝2时，在图3中示出用来区分各个中心值的中心值的键值。同时，当在分布式处理环境下需要各个中心的值时，除了针对映射化简框架(MapReduce Framework)的键之外，在用于交换的值中可以包括d维中心值。

[0055] 为了支持根据本发明的聚类分析功能，装置控制单元160支持数据151的传输、聚类分析执行的结果的收集、k值的选择、以及聚类分析的索引计算，并且具体地，执行k值更新。为此，装置控制单元160具有如图4所示的结构。

[0056] 图4是例示图2的装置控制单元160的详细图。

[0057] 参照图4，根据本发明的装置控制单元160包括数据分配单元161、分析结果选择单元163、分析索引应用单元165以及最优值更新单元167。

[0058] 数据分配单元161可以基于用户通过装置输入单元120输入的输入值执行计算中心值矢量的初始化处理，以与分布式处理服务装置200的数据节点共享由用户输入的预定范围内的k值。也就是说，数据分配单元161可以通过利用指示对同一k值以及预定范围内的k值的k均值聚类的迭代的NCS值来计算中心值矢量的数量。例如，当最小k(min_k)为3，最大k(max_k)为5且NCS为2时，中心集合的总数量为6，并且中心矢量的总数量为3*2+4*2+5*2＝24。同时，用户可以指定针对应用聚类分析的收敛条件。例如，作为收敛条件，可以指定最大可执行迭代频率和测量阀值。根据本发明，同时执行对多个集合(与特定k-集合ID和特定NCS ID对应的中心的集合)的聚类分析。因此，对于满足阀值条件的集合，存储该集合的值，随着k-集合ID|NCS ID|收敛执行数据写入，并且不再计算该集合。同时，执行聚类分析，直至全部集合满足在收敛条件下建立的测量阀值。如果完成了矢量初始化处理和收敛条件设置，则数据分配单元161可以向分布式处理服务装置200的每个数据节点发送中心集合。在该处理中，数据分配单元161可以支持以对在分布式处理服务装置200的每个数据节点之间分布的数据执行分布式处理。

[0059] 分布式处理服务装置200的每个数据节点输出通过将所发送的中心值集合应用于从聚类分析服务装置100中读取的数据而获得的聚类分析执行的结果。每个数据节点利用聚类分析服务装置100执行映射化简方法的化简操作，以发送输出的聚类分析的结果。

[0060] 对于每个化简器，从映射器输出的关于每个键的全部输入数据被发送作为值(value)。与映射器相似，化简器也相对于每个集合(针对特定k-集合ID的特定NCS ID)输出两个结果。当键被表示为(k-集合ID|NCSID|c_i_)时，每个值是与当前键相对应的集合的中心值c_i_最近的输入数据的矢量值。在下面的算式3中表示出作为值输入的全部输入数据的矢量和。

[0061] 【算式3】

[0062] sum_c_i＝sum_ci_1,sum_ci_2,...,sum_ci_d

[0063] 这里，和_c_i表示全部矢量的和。

[0064] 同时，如果在全部的n个数据中被确定为靠近c_i且因此被发送到当前化简器的数据片的数量为n_i，则(k-集合ID|NCS ID|c_i)的新值被表示为算式4。

[0065] 【算式4】

[0066] (c_i1',c_i2',...,ci_d')=1/n_i*<snm_ci_1,sum_ci_2,...,sum_ci_d)

[0067] 分布式处理服务装置200的数据节点输出(键,值),同时将算式4中得到的新值作为值。这里,键=<k-集合ID|NCS ID|中心ID>,并且值=<c_1',c_2',…,and c_n'>。

[0068] 同时,如果键为(k_集合ID|NCS ID|err|),则每个值为与特定k-集合ID|NCS ID对应的聚类分析的全部测量值的计算需要的测量的部分和。对进入当前化简器的全部值求和。在这种情况下,最终测量值是通过将和值除以全部数据的数量(即总数n)而得到的值。数据节点可以输出(键,值),同时将获得的测量值作为值。这里,键=<k-集合ID|NCS ID|诸如'err'虚拟值>,并且值=<和_测量_值(sum_measure_value)/n>。

[0069] 同时,分布式处理服务装置200的数据节点检查相对于全部k-集合ID|NCS ID|的全部中心值的收敛条件。满足收敛条件的中心被指示为收敛的。如果全部k-集合ID|NCS ID|都不是收敛的或者除非达到最大迭代,则数据节点将先前中心集合改变为新获得的值,并且可以反复执行上述处理。

[0070] 如果聚类分析执行的结果被导出,则数据节点可以将该结果发送到分析结果选择单元163。

[0071] 分析结果选择单元163被构造成在根据化简操作从分布式处理服务装置200中接收到通过将中心值集合应用到数据而得到的聚类分析的结果时从分析结果中选择特定中心值集合。也就是说,数据节点输出相对于每个k-集合ID|NCS ID的测量值,并且将测量值发送到分析结果选择单元163。分析结果选择单元163仅存储与相对于针对各个k-集合ID从1到NCS ID范围内的全部频率的测量值当中的最小(最优)测量值相对应的中心ID值。也就是说,分析结果选择单元163仅过滤相对于单个k-集合ID的一个值。例如,当由数据节点对利用k=3～5和NCS=4的全部4000个四维数据执行聚类分析时,分析结果选择单元163可以如图5中所示收集相对于每个k-集合ID的最优中心值。在当k-集合ID为3且NCS为4时的情况下,相对于k=3以并行方式执行四次聚类化分析,并且分析结果选择单元163可以选择与NCS值4相对应的分析执行的结果。类似地,当k-集合ID为4时,分析结果选择单元163可以选择与NCS值3相对应的分析执行的结果。

[0072] 如图5中所示,已选随后的值是测量值。下面的行被以'数字|坐标'的形式提供,并且当k=3时,第一中心的坐标为(17.33、12.00、10.99和1.64),并且属于该聚类的数据的数量为全部4000个数据中的3000个。当k为4时,第三个NC执行结果被选择,并且当基于这些中心值实现聚类分析时,属于每个聚类的数据的数量为1000,1000,954和1046。

[0073] 分析结果选择单元163基于从数据节点收集的结果选择特定分析结果,使得通过利用所选择的数据进行相对于原始数据的聚类分析。为此,分析结果选择单元163可以通过将所选择的中心值信息发送到全部数据节点来执行所选择的中心值信息的共享。

[0074] 分布式处理服务装置200的数据节点可以利用聚类分析服务装置100执行映射步骤,并且在这种情况下,通过利用所选择的中心值信息相对于各个k-集合ID(例如k=3、4和5)执行聚类分析。各个数据节点计算k-集合ID中的各个中心值与当前数据251之间的距离,并指定最近中心值的索引作为当前数据251的聚类索引。另外,数据节点通过利用所选择的中心值作为值同时利用当前输入数据作为键来执行输出。也就是说,数据节点通过将中心值输入到各个k-集合ID中来分析各个数据所属的聚类。因此,单个数据251被分类成全总共k-集合ID个片段,并且在引入全部n个输入数据时,输出多达nX|k-集合ID|的数据。

[0075] 分析索引应用单元165可以收集由具有聚类后的索引的数据节点指派的各个数据251。例如，在当输入四维数据为10、10、20和15时的情况下，可以输出包括10、10、20和153|4|1，10、10、20和154|3|2，以及10、10、20和155|4|1在内的结果。这里，当k-集合ID为3时数据251被分类成属于三个聚类中的第二(1)聚类，当k-集合ID为4时属于第三(2)聚类，并且当k-集合ID为5时属于第五(4)聚类。

[0076] 在接收到各个数据251的聚类分析的结果时，分析索引应用单元165可以利用聚类索引来确定哪个k通过利用分配了聚类索引的数据产生了最优的聚类分析效果。在当在采用聚类之前使用了全部数据时的情况下，需要用来计算聚类索引的时间会很长。因此，分析索引应用单元165可以被设置为如果需要的话执行采样。可以在三个选项中选择采样。也就是说，作为第一个NONE(非)选项，分析索引应用单元165依赖于数据的数量可以不采用采样使得全部数据被使用。第二采样选项操作使得分析索引应用单元165从全部n个数据中随机提取X％的数据而与聚类分布无关。第三选项操作使得分析索引应用单元165在将针对各个聚类的数据的数量考虑在内从针对各个聚类的全部n个k数据中随机提取X％的数据。

[0077] 例如，当假设属于聚类1的项的数量为500，属于聚类2的项的数量为299，并且属于聚类3的项的数量为300时，分析索引应用单元165可以以20％的提取率从聚类1中提取100个项，从聚类2中提取40个项，并且从聚类3中提取60个项。这里，分析索引应用单元165可以依赖于各个聚类的项的数量在将‘最小数量’和‘提取率’考虑在内来提取数据。例如，分析索引应用单元165可以针对20％的提取率和100的最小数量来设置选项。这里，如果以20％的提取率提取的数据的数量不与最小数量相匹配，则分析索引应用单元165可以随机提取数据，使得各个聚类的最小数量大于或等于作为默认值的预定数量。分析索引应用单元165可以向各个数据节点提供用于收集上述各项的选项信息，并从各个数据节点请求与所述选项信息相对应的信息。

[0078] 同时，为了评估聚类分析的有用性，分析索引应用单元165可以采用各种索引。在这种情况下，分析索引应用单元165可以使用具有根据各个聚类之间的距离的密度以及根据混合于其中的特定聚类中的各个数据之间的距离的密度中的至少一方。例如，分析索引应用单元165可以采用邓恩索引和戴维斯-波尔丁索引。邓恩索引中的较高级别指示优异的结果，并且戴维斯-尔丁索引中的较低级别指示优异的结果。聚类分析被基于各个聚类之间的距离(聚类间距离)以及聚类中的各个数据之间的距离(聚类内距离)评估。在这种情况下，测量，也就是用来测量距离的标准可以利用诸如欧几里得距离、曼哈顿距离、切比雪夫距离等的距离索引进行计算。计算聚类间距离的方法的示例包括单一联动(Single Linkage)、完整联动(Complete Linkage)、平均联动(Average Linkage)、矩心联动(Centroid Linkage)、平均到矩心联动(Average to Centroid Linkage)、豪斯多夫度量工作(Hausdorff metrics Job)等。计算聚类内距离的方法的示例包括完整直径(Complete Diameter)、平均直径(Average Diameter)、矩心直径(Centroid Diameter)等。同时上述的邓恩索引利用算式5计算。

[0079] 【式5】

[0080] $$D(U)=\min_{1\le i\le c}\left\{\min_{\substack{1\le j\le c\\ j\neq i}}\left\{\frac{\delta(X_i,X_j)}{\max_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\}$$

[0081] 这里，c是全部聚类的数量，并且U表示全部数据。

[0082] 戴维斯-波尔丁索引方法利用算式6计算。

[0083] 【式6】

[0084] $$DB(U)=\frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left\{\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i,X_j)}\right\}$$

[0085] 这里，c是全部聚类的数量，并且U表示全部数据。最优值更新单元167计算相对于由分析索引应用单元165计算出的索引计算各个k-集合ID的聚类索引，并且选择生成最优值的k。在当采用了多个聚类分析索引来评估时的情况下，通过所述多个聚类分析索引的投票得到最优k。例如，如果所使用的聚类分析索引的总数为三，则在k＝3时聚类分析索引1和聚类分析索引2为最优，在k＝4时聚类分析索引3为最优，并且在k＝5时任何聚类分析索引都不是最优，可以选择k＝3。

[0086] 图6是例示被分类成四个聚类的特定16维数据的上述聚类分析的执行、结果选择、通过共享所选择的结果的数据聚类分析、以及对被指派了聚类索引的各个数据的聚类分析索引的计算结果的图。最优值更新单元167检查在k＝4时全部聚类分析索引为最优，并确定k＝4为最优值。

[0087] 同时，最优值更新单元167被设置为将通过当前聚类分析处理获得的k值与先前获得的k-prev值以及该k-prev值的聚类分析索引值进行比较，如果新得到的k值较优异，则利用该新获得的k值执行更新，并且在执行实际的聚类分析中利用所述新得到的k值。

[0088] 图7是例示根据本发明的可以被包括在分布式处理服务装置200中的数据节点的构造的示例的图。

[0089] 参照图7，根据本发明的分布式处理服务装置200可以包括节点通信单元210、节点存储单元250、以及节点控制单元260。

[0090] 具有上述构造的数据节点通过节点通信单元210形成与聚类分析服务装置100的装置通信单元110的通信信道，并且接收预定范围内的k值和数据151，并且存储接收到的k值和数据151。数据节点可以执行聚类分析。为此，节点存储单元250可以包括用于聚类分析的算法，以及作为用于从聚类分析服务装置100接收数据的程序的映射化简(MapReduce)253。映射化简253被构造成允许多个数据节点在从聚类分析服务装置100接收数据、执行聚类分析、然后为聚类分析服务装置100提供聚类分析的结果的处理中通过分配来接收并处理大数据。映射化简253可以通过利用如上所述的映射器和化简器支持根据本发明的聚类分析执行。

[0091] 同时，节点存储单元250可以暂时地或半永久地存储由聚类分析服务装置100提供的数据。也就是说，节点存储单元250可以存储数据251。在这种情况下，所存储的数据251可以是与存储在聚类分析服务装置100中的数据151明显相同的数据，并且可以是被指派了根

据聚类分析执行的结果的聚类索引的数据。

[0092] 节点控制单元260被构造成进行支持使得在聚类分析服务装置100请求执行聚类分析时对所发送的预定范围内的k值和数据根据预定义的k-均值聚类化方法执行聚类分析，并且向聚类分析服务装置100提供聚类分析执行的结果。节点控制单元260可以对聚类分析服务装置100的例示中已经描述的聚类分析的执行以及聚类索引的指派的控制。

[0093] 例如，节点控制单元260可以允许在各个数据节点上操作的各个映射器或者在同一数据节点上运行的多个数据节点一次读取分布的一行的各个输入。另外，节点控制单元260可以对全部k-集合ID和全部NCS ID执行下面的处理。也就是说，节点控制单元260获取与从第一中心开始的全部k-集合ID中心中的当前输入最近的中心的索引(c*)，并且使用所获取的索引(c*)作为键。值是当前输入数据的矢量值。这样获得的结果(键，值)被输出。这里，键＝(k集合ID|NCS ID|c*)，并且值＝(x_i1,x_i2,..,x_id)。节点控制单元260输出具有最近的中心的距离值(测量)。这里，键＝(k集合ID|NCS ID|err)，并且值＝(1|距离)＝(1|sqrt((c*_2x2)^2),(c*_1x1)^2),…,(c*_nxn)^2))(在欧几里得距离的情况下)。通过上述的处理，对单个输入数据(单行)，节点控制单元260可以支持使得相对于单个数据生成(键，值)＝|k集合ID|X|NCS|X2个输出。

[0094] 根据本发明的分布式处理服务装置200以迭代的方式根据聚类分析服务装置100控制执行聚类分析和应用，使得即使当改变数据时在没有用户的额外干预的情况下持续获得具有最优k值的聚类分析结果。这里，分布式处理服务装置200可以通过采用映射化简框架的组合器来提高执行速度。由于该组合器仅在本地数据节点中执行由化简器执行的功能，所以该化简器需要被改变以不利用映射器的输出而是利用该组合器的输出执行相同的功能。例如，可以假设环境，其中为了根据聚类分析执行的新中心值，映射器输出(键，值)＝(中心ID、数据矢量)，并且化简器接收(键，值)＝(中心ID、数据矢量列表)作为输入，并获取数据矢量列表的矢量和以获得平均值。当如果组合器在该环境中使用时，映射器输出(键，值)＝(中心ID、数据矢量)，组合器接收(键，值)＝(中心ID、数据矢量列表)作为输入，并输出(键，值)＝(中心ID、数据矢量列表的数量|数据矢量的矢量和)，并且化简器接收组合器的输出作为输入以获取平均值。也就是说，单个化简器被设计成接收关于单个键的全部数据，并通过利用该全部数据来计算总和。然而，如果多个组合器被采用，每个组合器计算单个键的部分总和，则化简器在从组合器接收输入的同时不知道属于单个中心ID键的数据的数量。因此，化简器将当计算部分和时经组合器处理的数据的数量与该部分和一起从组合器接收，以计算总和。该方法可以通过相对于单个中心ID键改变(键，值)来实现，使得由各个组合器处理的输入数据的数量被附加地发送到化简器，并且化简器被改变使得总和的平均值被化简器得到。

[0095] 图8是例示支持聚类化的方法的图。

[0096] 参照图8，根据支持聚类分析的方法，聚类分析服务装置100可以从用户收集k值的范围，并且对中心值矢量执行初始化处理(801)。此时，聚类分析服务装置100可以收集聚类分析的收敛条件。随着用户通过装置输入单元120输入或者命令应用预定义值，k值的范围和收敛条件可以被限定。然后，聚类分析服务装置100可以向分布式处理服务装置200的数据节点提供全部中心集合和数据(802)。在这种情况下，聚类分析服务装置100和分布式处理服务装置200的数据节点可以被设置成使得数据根据映射化简(MapReduce)方法通过利

用相应映射器以分布式方式处理。接收到了中心集合和数据的数据节点可以根据接收到的中心集合和数据执行k-均值聚类化分析(803)。

[0097] 如果k-均值聚类的结果被输出，则分布式处理服务装置200的数据节点可以通过利用各个化简器将结果发送到聚类分析服务装置100(804)。具体地，数据节点各自响应于单个中心值可以输出单个输出。在从分布式处理服务装置200的数据节点接收到分析的结果时，聚类分析服务装置100可以在所述分析的结果中选择最优结果(805)。聚类分析服务装置100将选择的结果提供到数据节点，使得所选择的结果被共享(806)。此时，聚类分析服务装置100可以请求使得聚类索引被根据所选择的结果而指派到各个数据。

[0098] 分布式处理服务装置200的数据节点控制使得基于由聚类分析服务装置100提供的选择的结果对各个数据指派聚类索引(807)。数据节点可以向聚类分析服务装置100提供被指派了聚类索引的数据(808)。然后，聚类分析服务装置100可以进行聚类索引应用并进行k值选择(809至812)。

[0099] 也就是说，如果要经过聚类索引的数据的量超过预定值，则聚类分析服务装置100可以在操作809中支持根据预定义的采样方法中的至少一种支持采样应用。然后，聚类分析服务装置100在操作810中进行各个k值的聚类索引的计算，并且在操作811中选择具有最高聚化类索引的k值。聚类分析服务装置100可以支持使得在操作812中比较两个聚类索引并且选择具有较高聚类索引的k值。也就是说，聚类分析服务装置可以应用多个索引方法来计算各个k值的聚类索引，并从索引方法中选择较高的k值。尽管上述进行的说明是关于聚类分析服务装置100直接执行采样进行的，但是本发明并不限于此。也就是说，聚类分析服务装置100可以向分布式处理服务装置200的每个数据节点提供采样条件，接收符合采样条件的数据，并且基于接收到的数据进行操作810至812。然后，聚类分析服务装置100可以将先前存储的值(k_prev)与当前k值比较以执行k值的更新，使得更合适的k值被存储(813)。

[0100] 本发明解决了除非用户在k-均值聚类化期间人工指定合适的K值而在获取优异的聚类化结果中的困难，并且支持使得聚类分析就系统而言并基于分析结果自动执行，聚类分析索引被计算，从而获得最合适的k值。

[0101] 另外，即使具有大尺寸的数据需要基于分布式处理框架进行处理，本发明也可以无限制地应用。另外，如果用户在不预先指定特定k值的情况下指定预定范围，则本发明自动发现最合适的k值，并且该发现处理在没有用户的干预的情况下周期性地执行，因此该值根据数据的变化通过其自身进行调整。

[0102] 同时，尽管已经进行的上述说明涉及聚类分析服务装置100和分布式处理服务装置200彼此分开以支持聚类分析服务，并且相互合作执行映射化简(MapReduce)功能，但是本发明并不限于此。也就是说，分布式处理服务装置200可以被聚类分析服务装置100的装置存储单元150的构造所代替。在这种情况下，装置存储单元150可以具有提供分布式处理服务装置200的数据分布式处理功能的在逻辑上划分的部分或者根据设计者的意图而物理地划分的部分，并且这种划分的部分可以作为上述的数据节点。这里，如果装置存储单元150被设计作为分布式处理服务装置200，则聚类分析服务装置100的装置控制单元160可以允许装置存储单元150的每个部分以可用于多任务处理的形式来设置，使得执行并发接入和数据处理。

[0103] 同时，提供聚类分析的功能可以以通过各种计算设备可执行的程序的形式来实

现，并且可以记录在计算机可读记录介质中。计算机可读记录介质可以在程序指令、数据文件和数据结构中实现，或者在上述的一个或多个的组合中实现。同时，记录在记录介质中的程序可以针对本发明设计和构造，或者通过程序员容易地构造。

[0104] 计算机可读记录介质的示例包括被构造为存储并执行程序指令的硬件设备，例如，诸如硬盘、软盘的磁性介质，以及磁带，诸如CD-ROM和DVD的光学介质，诸如软式光盘的磁光介质、只读存储器(ROM)、随机存取存储器(RAM)、以及闪速存储器。另外，程序指令可以包括由编译器制作的机器代码，和通过解译器可由计算机执行的高级语言。上述的硬件设备可以作为一个或多个软件模块运行，以执行本发明的操作。

[0105] 尽管已经通过特定的术语参照说明书及其附图描述了根据本发明的一些实施方式的聚类分析功能，术语的使用仅仅是用于完善本发明并帮助本领域技术人员彻底理解本发明，因此本发明的范围不限于这些实施方式和术语。因此，对于本领域技术人员而言很明显，在不偏离本发明的范围的条件下，可以进行各种示例性实施方式。

[0106] 工业实用性

[0107] 本发明提供能够在没有用户的控制的情况下自动计算出最优k值，并支持根据该最优k值的聚类分析，使得提高数据聚类分析的可靠性和稳定性的聚类分析支持系统以及支持该聚类分析的方法和装置。

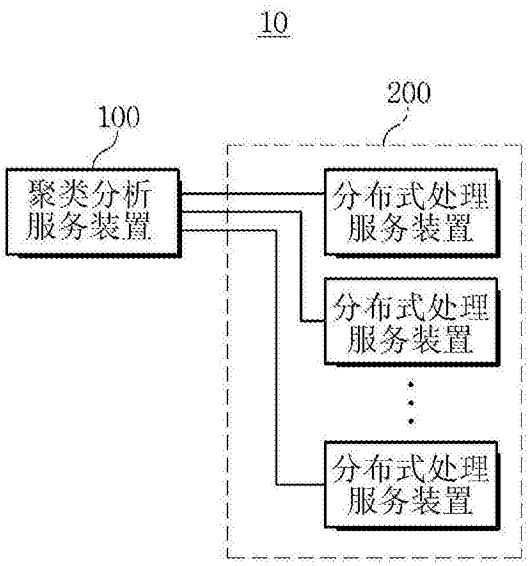

10
100
200
聚类分析
服务装置
分布式处理
服务装置
分布式处理
服务装置
分布式处理
服务装置

图1

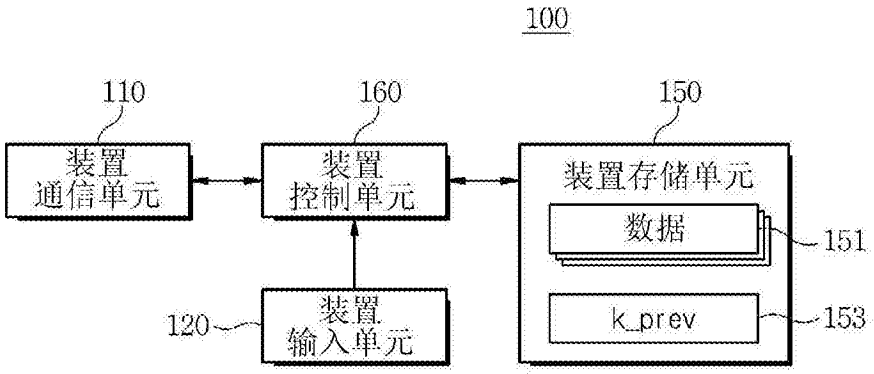

100
110
160
150
装置
通信单元
装置
控制单元
装置存储单元
数据
151
120
装置
输入单元
k_prev
153

图2

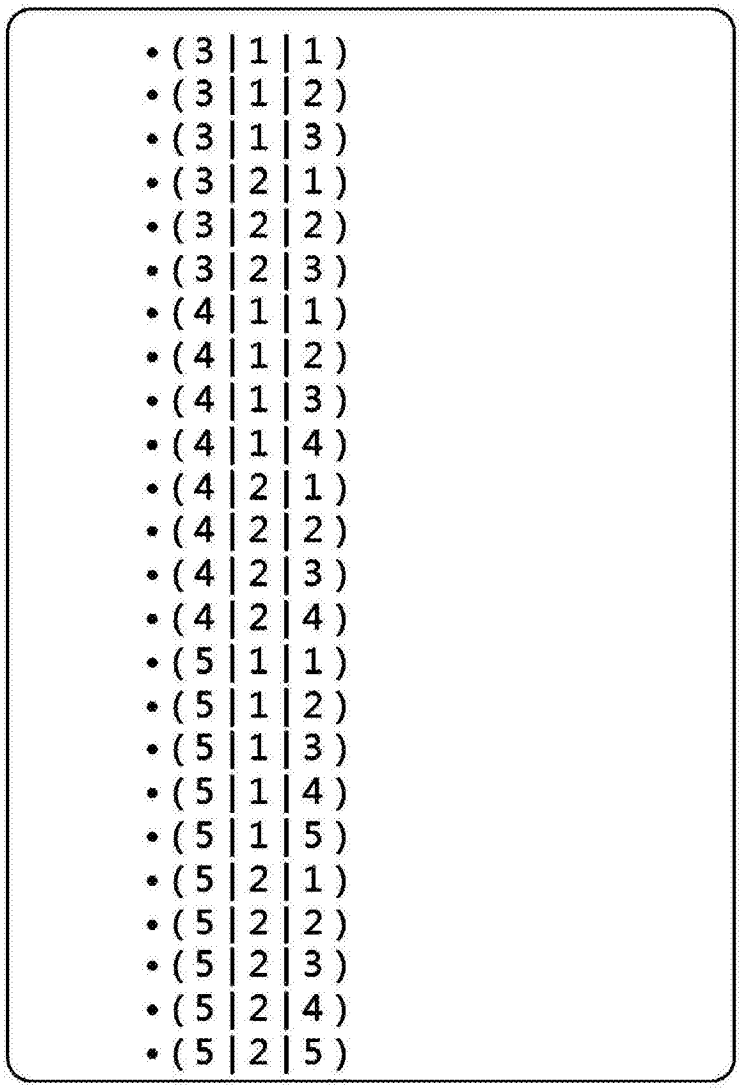

•(3|1|1)
•(3|1|2)
•(3|1|3)
•(3|2|1)
•(3|2|2)
•(3|2|3)
•(4|1|1)
•(4|1|2)
•(4|1|3)
•(4|1|4)
•(4|2|1)
•(4|2|2)
•(4|2|3)
•(4|2|4)
•(5|1|1)
•(5|1|2)
•(5|1|3)
•(5|1|4)
•(5|1|5)
•(5|2|1)
•(5|2|2)
•(5|2|3)
•(5|2|4)
•(5|2|5)

图3

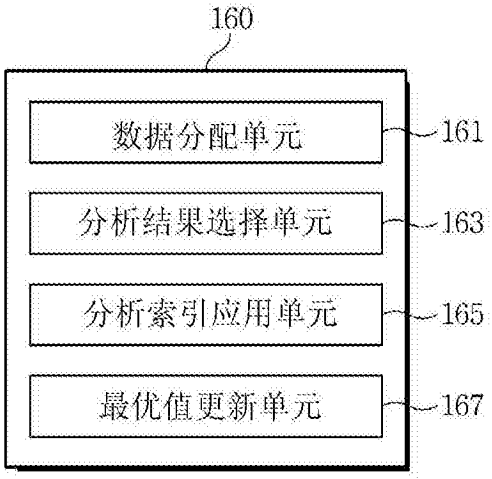

160
数据分配单元
161
分析结果选择单元
163
分析索引应用单元
165
最优值更新单元
167

图4

二维输入数据的执行结果

```
3|4| 所选择 |310.5446153021632
        3000|17.335215,12.002399,10.99198,1.6476852
        252|9.112253,-0.67824936,3.5064828,14.812881
        748|8.920427,0.2394945,2.8139524,15.034133
4|3| 所选择 |11.847766357457289
        1000|19.02186,1.9659603,3.978168,0.96862584
        1000|8.968765,0.008223154,2.98847,14.978377
        954|17.020912,15.0289,13.954832,0.9074227
        1046|16.009405,18.837154,14.995067,2.9720392
5|4| 所选择 |11.04168245593631
        1000|8.968767,0.008223152,2.9884713,14.978378
        1000|19.02186,1.9659607,3.9781685,0.9686255
        384|17.470125,14.620083,13.823077,1.0902998
        624|16.670918,15.296638,14.106458,0.88726085
        992|16.000612,19.034283,15.007315,3.0263195
```

图5

聚类分析索引的计算的示例

- 在生成被分类成四个聚类的总共50000/100000个16维数据之后本发明的应用即，在k=4期望最优聚类分析结果
- 在K=3～6（3，4，5，6），NCS=3的执行

#数据	运行时间	K	轮廓	邓恩索引	戴维斯-泊尔丁索引
50万	SH-40分21秒 DI/DBI -1分24秒 (基于10秒)	3	0.6482	1.3194	0.2947
		4	0.8351	7.3865	0.0676
		5	0.6543	0.4065	0.9634
		6	0.4494	0.4050	0.8057
100万	SH-40小时23分 DI/DBI- 1分58秒 (基于10秒)	3	0.6878	2.0006	0.1979
		4	0.8693	10.1349	0.0493
		5	0.6608	1.0985	0.3697
		6	0.6597	1.0664	0.3017

图6

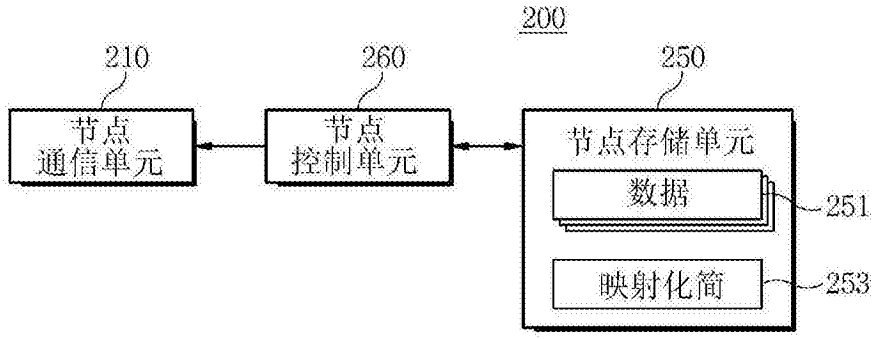

200
210
260
250
节点
通信单元
节点
控制单元
节点存储单元
数据
251
映射化简
253

图7

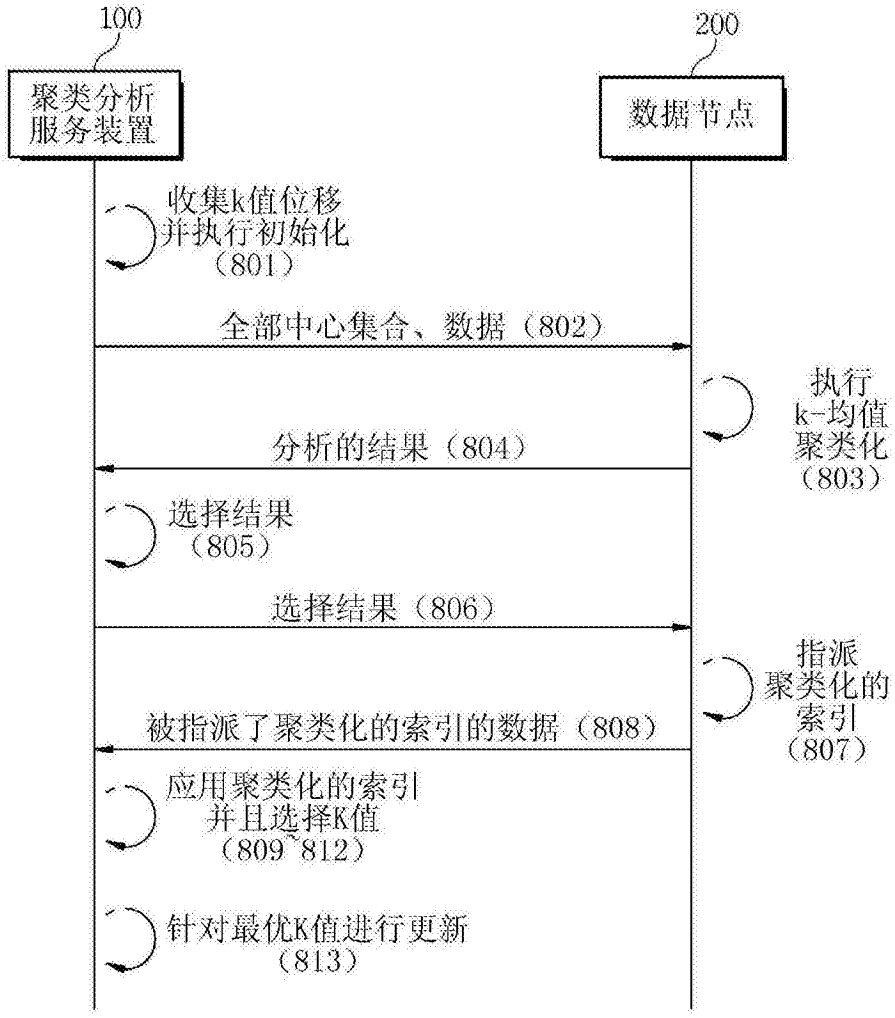

100
200
聚类分析
服务装置
数据节点
收集k值位移
并执行初始化
（801）
全部中心集合、数据（802）
执行
k-均值
聚类化
（803）
分析的结果（804）
选择结果
（805）
选择结果（806）
指派
聚类化的
索引
（807）
被指派了聚类化的索引的数据（808）
应用聚类化的索引
并且选择K值
（809~812）
针对最优K值进行更新
（813）

图8