

基于模糊分区聚类的关联挖掘改进算法

李慧琴, 孙 英, 王俊洁

(楚雄师范学院 信息科学与技术学院, 云南 楚雄 675000)

摘 要: 在存储空间网格分布的不均衡性影响, 导致大数据挖掘的聚类配准性不高, 为了提高数据挖掘效能, 提出一种基于模糊分区聚类的大数据关联挖掘改进算法. 分析数据在云存储空间中的网格结构模型, 提取大数据信息流的关联语义规则性特征量, 对提取的特征量进行自适应加权学习训练, 增强数据的属性特征分布强度. 在存储空间采用模糊分区方法对提取的数据关联特征进行优化聚类, 根据聚类结果进行语义划分, 构建判别统计量和检验准则进行数据挖掘的聚类属性判断, 提高数据挖掘的准确性, 实现大数据优化挖掘. 仿真结果表明, 采用该方法进行大数据挖掘的分区性能较好, 数据的归类准确性较高, 提高了数据的查准率.

关键词: 模糊分区聚类; 大数据; 关联挖掘; 特征提取

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2018)03—0130—05

An Improved Algorithm for Association Mining Based on Fuzzy Partition Clustering

LI Hui-qin, SUN Ying, WANG Jun-jie

(School of Information Science and Technology, Chuxiong Normal University, Chuxiong 675000, China)

Abstract: Affected by the unbalanced distribution of grid storage space, resulting in large data mining cluster registration is not high, in order to improve the efficiency of data mining, proposes an improved algorithm for mining large data association based on fuzzy partition clustering analysis. The grid structure model in the cloud storage space in the data association, semantic rules of feature extraction big data flow of information, the feature extraction of adaptive weighted attribute data processing, enhance the intensity distribution, the extracted data association features optimized clustering using fuzzy partition method in the storage space, the semantic partition based on clustering results, construct discriminant statistics and test criteria for data mining clustering attribute judgment to improve the accuracy of data mining, data mining optimization. The simulation results show that using the method of performance data mining area Better, the accuracy of data classification is higher, and the accuracy of data is improved.

Key words: fuzzy partition clustering; large data; association mining; feature extraction

1 引言

在大数据信息处理环境下,海量大数据存储在云计算网格中,构成云存储数据库.为了实现对大数据的准确调度和访问,需要进行大数据关联挖掘.结合特征聚类和数据检索方法可以实现对云存储空间中的海量数据检索和提取.研究关联数据优化挖掘方法,对提高数据库的优化访问和调度性能具有重

要意义.对云存储空间中大数据的挖掘是建立在对数据的关联规则映射和特征聚类基础上的,通过提取数据的关联特征,结合网格化分区调度方法进行数据信息融合和分组,并根据数据的分组聚类结果进行关联挖掘^[1].随着大数据规模的不断增大,大数据关联挖掘方法越来越受到人们的重视.

从大数据的挖掘集信息处理、数据滤波降噪以及特征分类提取为一体的综合性处理过程,实现数

据挖掘的第一步是进行大数据信息流的时间序列分析和特征重组.采用自适应学习方法进行数据关联性搜索和聚类,常见的聚类方法如 K-means 算法、模糊 C 均值算法和层次分割聚类方法等^[2-3].通过构造大数据特征分布的模糊聚类中心,采用邻近节点搜索和特征采样进行数据分类挖掘,其中,模糊 C 均值聚类方法在进行大数据挖掘的聚类处理中容易陷入局部最优解,在进行动态数据的特征提取时收敛性不好^[4];而采用层次分割聚类方法进行大数据挖掘会受到分割阈值的影响,导致对初始的聚类中心具有较大的敏感性.结合上述方法存在的问题,不少学者进行了大数据关联挖掘改进设计.文献^[5]中提出一种基于频繁项自适应学习大数据挖掘方法,采用分段预白化匹配滤波算进行类间闭频繁项干扰抑制处理,提高数据挖掘的分类能力,但该方法随着数据关联特性扰动的增大,挖掘的准确性会降低;文献^[6]中提出基于近邻传播与密度相融合云数据挖掘方法,反复调整聚类中心来进行大数据的聚类优化,在近邻节点提取数据的密度信息进行特征归集,结合 PID 神经网络训练进行数据分类挖掘,该方法由于计算开销较大,所以挖掘过程的实时性不好.针对上述问题,本文提出一种基于模糊分区聚类的大数据挖掘改进算法.

2 大数据的存储结构分布及特征提取预处理

2.1 大数据云存储结构模型

为了实现对云存储数据库中的大数据关联挖掘和特征聚类处理,需要首先分析云存储中的大数据分布式网络结构模型,在云存储空间中,设分布异构环境下大数据在云存储系统中的节点分布间隔为 d ,每个存储基本块为 $m_{i,j} (i \leq n, j \leq k)$,用有向图表示 $G = (V, E, W, C)$ 大数据分布式存储的网络结构模型,其中,在不同的初始聚类中心,大数据存储的属性分布用子图 $G_1 = (M_1^i, M_1^j, Y_1), G_2 = (M_2^i, M_2^j, Y_2)$ 表示,根据云存储空间中数据存储的负载均衡性和链路吞吐恒定性,则满足 $G_1 \subseteq G_2 \Leftrightarrow Y_1 \subseteq Y_2$,令 $A = \{a_1, a_2, \dots, a_n\}$ 为存储介质的干扰幅值,大数据存储节点的状态特征用 $P(i, j)$ 表示, $p(i, j) = \lim_{t \rightarrow \infty} p\{a_t = i, b_t = j\}$,其中数据采样间隔单元 $i \in [0, n], j \in [0, n]$.在同一时刻, Sink 近邻节点中存储的大数据比特序列具有关联性,则存在如下状态特征分布表达式:

$$p(i, j) = p(i)p(j) = \frac{\frac{\rho^i}{i!}}{\sum_{i=0}^n \frac{\rho^i}{i!}} \frac{\frac{\rho^j}{j!}}{\sum_{j=0}^n \frac{\rho^j}{j!}},$$

$$i, j = 0, 1, \dots, n \quad (1)$$

将数据存储元 $\alpha(i, j)$ 输入数据到存储链路层中,得到关键路径节点采集的大数据时间序列为 x_n ,期望响应为 d_n ,根据数据流图上关键组件进行特征分区^[7],根据数据传输链路层的通道容量确定数据分布的竞争模型如下:

$$\alpha(i, j) = \begin{cases} 0, i = 0 \text{ or } j = 0 \\ 1, n - j < i, i \geq j \\ 1, n - i < j, j \geq i \\ 1 - n_j C_{i/n} C_{i, n-j} \geq i, i \geq j \\ 1 - n_i C_{j/n} C_{j, n-i} \geq j, j \geq i \end{cases} \quad (2)$$

采用 Parallel Sets 变元轴排序方法进行大数据存储结构分区调度,结合模糊分区聚类方法发现数据集的隐含模式,从而提高数据挖掘性能.

2.2 关联语义规则性特征提取

在进行数据存储结构模型和信息流模型构建的基础上,提取大数据信息流的关联语义规则性特征量,根据大数据分布的相似属性排列得到关联规则的概率函数:

$$T(n) = \sum_i R(n, i) + T(n+1) \quad (3)$$

式中, $T(n+1)$ 表示第 $n+1$ 层数据存储节点的关联规则相似度, $R(n, i)$ 表示第 n 层链路的 i 个云存储节点的数据自相关特征量.当 $n=0$ 时, $T(0) = T(1)$; 当 $n=N$ 时, $T(n) = \sum_i R(n, i)$,通过层次渲染方法求得数据在云存储通道内的互信息量为

$$I(i) = \sum_{j \in \text{Intf}(i) \text{ 且 } j \neq i} T(j) \quad (4)$$

式中, $\text{Intf}(i) = B - \frac{L\mu(i)}{128 \ 512}$,表示大数据的语义关联性分布概率密度; B 是的时态分布属性; L 为大数据分组长度; i, j 节点表示具有相似度的大数据分布节点; $\mu(i)$ 表示在存储节点 (i, j) 的语义相似度; p_i 表示连接结点的连通性概率,其对应的权重为 w_i .由此得到数据的关联语义特征提取的节点分布轴排列示意图如图 1 所示.

根据图 1 所示的分轴示意图进行关联特征提取,并对提取的特征量进行自适应加权学习训练,增强数据的属性特征分布强度^[8],得到自适应加权学习函数为

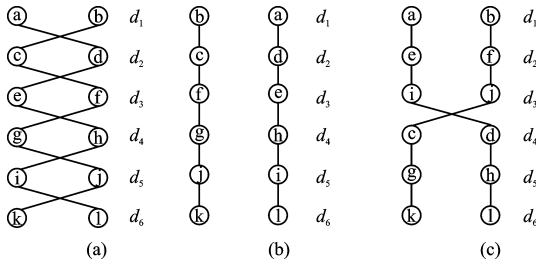


图1 数据的关联语义特征提取的节点分布轴排列示意图

$$\begin{cases} \max U = u_1 + u_2 + \dots + u_n \\ u_i = p_i \\ \sum_{i=1}^n p_i = 1, 0 < p_i < 1 \\ \frac{p_1/(1-p_1)}{\omega_1} = \frac{p_i/(1-p_i)}{\omega_i} = \dots \\ = \frac{p_n/(1-p_n)}{\omega_n} = \frac{1}{K} \end{cases} \quad (5)$$

式中, K 表示加权调节系数. 考虑到 K 值的易变性, 采用同步控制方法进行模糊分区调度, 设置一个相同的权重系数 ω_0 进行大数据的线谱增强处理, 得到数据存取通道内的准确传递概率为 $p_i = r_i p_0$, 初始加权系数 $K_0 = (m-1)\omega_0$, 大数据集合基数满足 $p_0 = \frac{1}{m}$, $m = r_1 + r_2 + \dots + r_i + \dots + r_n$, 优化的大数据关联语义规则特征提取为

$$p_i = \frac{r_i}{m} = \frac{r_i \omega_0}{m \omega_i} = \frac{\omega_i}{\omega_1 + \dots + \omega_i + \dots + \omega_n} \quad (6)$$

通过上述处理, 对提取的关联语义规则特征进行失补偿, 降低了节点竞争导致的存储开销.

3 数据关联挖掘改进算法实现

3.1 模糊分区聚类算法

在分析数据在云存储空间中的网络结构模型, 提取大数据信息流的关联语义规则性特征量并进行自适应加权学习训练处理的基础上, 进行数据关联挖掘改进设计, 本文提出一种基于模糊分区聚类的大数据关联挖掘改进算法, 采用图2所示的模糊分区网格进行关联数据的区域分割处理. 图2中, 海量大数据的关联语义规则特征量的层次网络图可用有向图结构描述为 $G(L_1, L_2, \dots, L_{n-1}, L_n, E)$, 其中 L_i 表示每一个模糊分区网格的数据存储节点, $E \subseteq (\cup (L_i \times L_{i+1}))$, $i = 1, \dots, n$ 代表数据聚类中的关联组 L_{i+1} 与其邻近数据聚类中心 L_i 连线的集合, 在数据聚类中心的分布集合中的任一元素 uv 表示数据聚

类的关联规则属性^[9], 聚类后形成的层次图 $u(u \in L_i)$ 与节点 $v(v \in L_{i+1})$ 的连线为一组有向向量集.

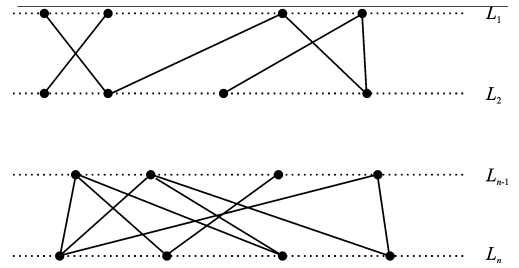


图2 模糊分区聚类网格示意图

设 $G(F) = \{F^1, F^2, \dots, F^m\}$ 为第一层节点按照权值聚类有与 $k(k > 0)$ 关联的变元对象集合, 其中, m 为关联数据挖掘的集合基数, 即多变元对象的数量, $F^i = F^i(c_1^i, c_2^i, \dots, c_{k-1}^i, q^i)$ ($q^i \in R, q^i \geq 0$) 代表表示模糊分区聚类中第 k 个变元对象, $c_1^i, c_2^i, \dots, c_{k-1}^i$ 表示二变元的相关系数关于 $(k-1)$ 变元属性, 得到关联数据挖掘的规则特征分布为

$$r(V_i, V_j) = \frac{\alpha}{|n_i - n_j| + 1} + \beta \sum_l \omega_l \quad (7)$$

式中, $\omega_l = \begin{cases} 1, & |x_l^i - x_l^j| < \epsilon \\ 0, & \text{otherwise} \end{cases}$ ($1 \leq l \leq t$), n_i, t 表示

二维数据的相关度和层次聚类的收敛系数. 用 q^i 表示由 $c_1^i, c_2^i, \dots, c_{k-1}^i$ 确定模糊分区的关联语义集, 称 $G(F)$ 为隶属度函数. 通过上述处理, 在存储空间采用模糊分区方法对提取的数据关联特征进行了优化聚类.

3.2 关联挖掘判决统计分析

根据聚类结果进行语义划分, 构建判别统计量和检验准则进行数据挖掘的聚类属性判断, 得到大数据关联挖掘的判决统计量为

$$Q_s = \frac{\langle (x_n - \bar{x})(x_{n-d} - \bar{x})(x_{n-D} - \bar{x}) \rangle}{\langle (x_n - \bar{x})^3 \rangle} \quad (8)$$

式中, x_n 表示大数据采样序列, d 表示两个变元的相关度, $D = 2d$, $\bar{x}$ 表示均值, $\langle x(n) \rangle$ 表示数据关联规则挖掘样本 $x(n)$ 取均值:

$$\langle x(n) \rangle = 1/N \sum_{n=1}^N x(n) \quad (9)$$

计算任意二维关联数据挖掘的相关度^[10], 得到相关度矩阵:

$$R = \begin{bmatrix} r(V_1, V_1) & \dots & r(V_1, V_{k-1}) \\ \dots & \dots & \dots \\ r(V_{k-1}, V_1) & \dots & r(V_{k-1}, V_{k-1}) \end{bmatrix} \quad (10)$$

由此得到数据关联挖掘的判决准则为

$$p(Q_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left[-\frac{(Q_s - \langle Q_s \rangle)^2}{2\sigma_s^2}\right] \quad (11)$$

$$\int_{-\infty}^{\infty} p(Q_s) dQ_s = 1 \quad (12)$$

如果 Q_0 与 $\langle Q_s \rangle$ 的绝对误差超过阈值 Q_c , 使得

$$p(|Q_0 - \langle Q_s \rangle| > Q_c) \leq 0.05 \quad (13)$$

设定数据关联挖掘的置信度为 95%, 根据均匀正态分布特性, 有:

$$0.025 = \int_{-\infty}^{z_2} p(Q_s) dQ_s = 1 - \int_{-\infty}^{z_1} p(Q_s) dQ_s \quad (14)$$

式中, $z_2 = -z_1$, 当 $S \geq 2.00$, 表示数据挖掘在模糊分区聚类中心以 95% 概率聚敛, 接受原假设, 数据关联挖掘的结果满足收敛条件。

4 仿真实验与性能分析

为了测试本文设计的关联挖掘算法在实现大数据优化聚类 and 挖掘中的应用性能, 进行仿真实验。实验的软件环境为 Matlab 7。硬件环境: Intel(R) Core(TM)2 Duo CPU 的 PC 机和 Professional SP3 系统服务器。测试数据集来自于云存储数据库 CLOUDD2016, 数据库中根据数据属性进行模糊网格分区, 分区的大小为 2 400 MB 的分区, 模糊分区的嵌入维数 m 和迭代补偿 L 分别取值为 25 和 12, 给定初始值 $c_{j, \text{tot}} = 0$, $\alpha_{j, \text{tot}} = 0$, 聚类中心的分布初始 $r = 2$, 对大数据关联挖掘的离散采样率为 $f_s = 10 * f_0 \text{ Hz} = 10 \text{ kHz}$, 训练样本规模为 2 480, 相关度 GGAP 的大小设定为 0.8, 根据上述仿真参量设定, 进行数据关联挖掘仿真, 首先进行测试数据采样, 得到测试样本集的时间-频域分布如图 3 所示。

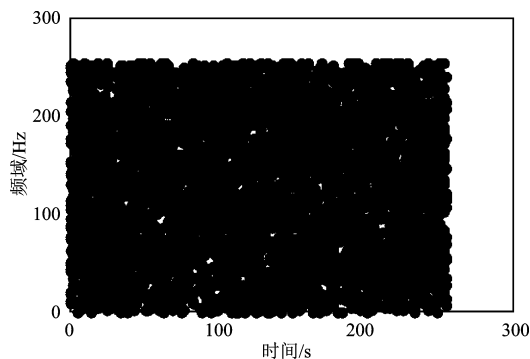


图3 测试样本集的时间-频域分布

以图 3 给出的数据为测试对象, 进行模糊分区聚类, 得到聚类结果如图 4 所示。

分析图 4 得知, 采用本文方法对数据进行模糊分区聚类, 数据在分区聚类中心聚敛程度较好, 在此

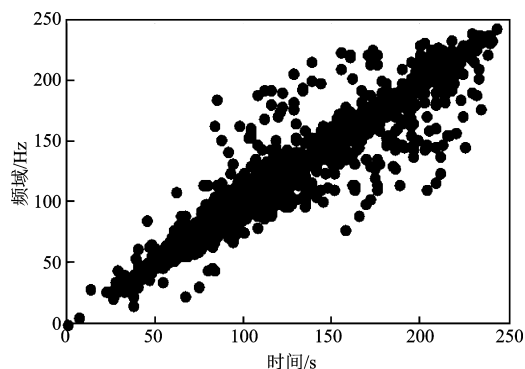


图4 数据聚类结果

基础上, 进行数据关联挖掘, 固定数据的相关度 GGAP, 分别设定数据关联规则维度 NIND 值为 10, 30, 80, 测试每次迭代的吞吐量的关系, 如图 5 所示。固定的值为 30, 测试数据关联挖掘的吞吐量与相关度的关系如图 6 所示。

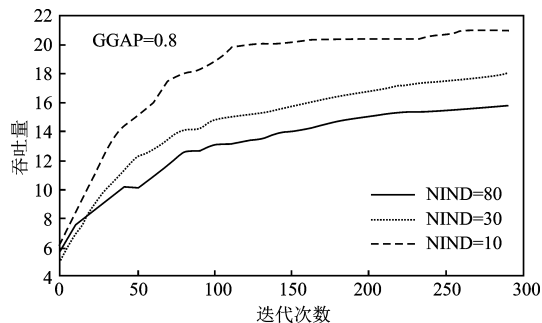


图5 数据关联挖掘的吞吐量与关联规则维度的关系

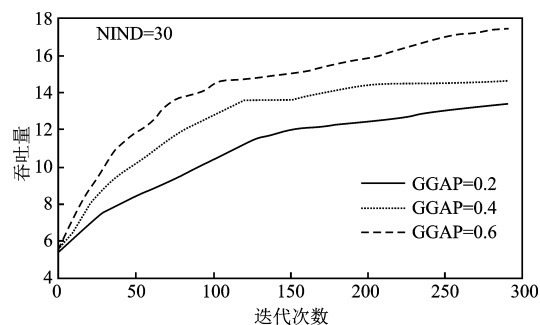


图6 数据关联挖掘的吞吐量与相关度的关系

分析图 5 和图 6 中的结果得知, 采用本文方法进行数据关联挖掘输出的吞吐量较高, 收敛性较好, 具有较好的关联挖掘效果。进一步测试采用本文方法和传统方法进行关联挖掘的误差, 得到对比结果如图 7 所示。分析图 7 得知, 采用本文方法进行模糊分区聚类 and 关联挖掘的误差较低, 性能较好。

5 结束语

通过提取数据的关联特征, 结合网格化分区调

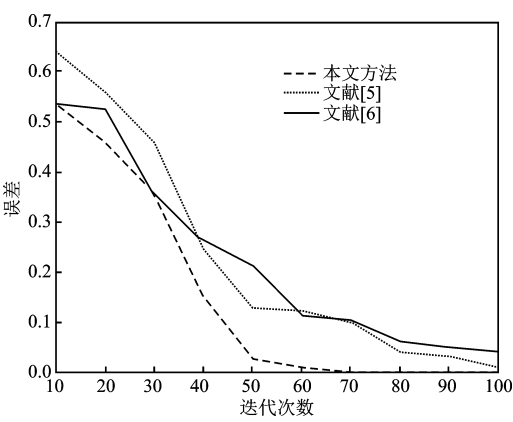


图 7 关联挖掘误差对比

度方法进行数据信息融合和分组,并根据数据的分组聚类结果进行关联挖掘,提高大数据访问和调度能力,本文提出一种基于模糊分区聚类的大数据关联挖掘改进算法.采用 Parallel Sets 变元轴排序方法进行大数据存储结构分区调度,结合模糊分区聚类方法发现数据集集中的隐含模式,在存储空间采用模糊分区方法对提取的数据关联特征进行优化聚类,根据聚类结果进行语义划分和关联挖掘.研究表明,采用该方法进行数据关联挖掘的聚类性能较好,收敛性较好,误差较低,具有很好的应用价值.

参考文献:

[1] 邢淑凝, 刘方爱, 赵晓晖. 基于聚类划分的高效用模式并行挖掘算法[J]. 计算机应用, 2016, 36(8): 2202-2206.

[2] Palomares I, Martinez L, Herrera F. A consensus model to detect and manage non-cooperative behaviors in large scale group decision making[J]. IEEE Trans on Fuzzy System, 2014, 22(3): 516-530.

[3] HUANG Hongwei, GE Xiaotian, CHEN Xuansong. Density clustering method based on complex learning classification system. Journal of Computer Applications, 2017, 37(11): 3207-3211.

[4] JI Yugang, LI Yitong, SHI Chuan. Aspect rating prediction based on heterogeneous network and topic model. Journal of Computer Applications, 2017, 37(11): 3201-3206.

[5] 侯森, 罗兴国, 宋克. 基于信息源聚类的最大熵加权信任分析算法[J]. 电子学报, 2015, 43(5): 993-999.

[6] 毕安琪, 王士同. 基于 Kullback-Leiber 距离的迁移仿射聚类算法[J]. 电子与信息学报, 2016, 38(8): 2076-2084.

[7] 吴鸿华, 穆勇, 屈忠锋, 邓丽霞. 基于面板数据的接近性和相似性关联度模型[J]. 控制与决策, 2016, 31(3): 555-558.

[8] 梁聪刚, 王鸿章. 微分进化算法的优化研究及其在聚类分析中的应用[J]. 现代电子技术, 2016, 39(13): 103-107.

[9] 林楠, 史苇杭. 基于多层空间模糊减法聚类算法的 Web 数据库安全索引[J]. 计算机科学, 2014, 41(10): 216-219.

[10] 余晓东, 雷英杰, 岳韶华, 等. 基于粒子群优化的直觉模糊核聚类算法研究[J]. 通信学报, 2015, 36(5): 74-80.

作者简介:

李慧琴 女, (1975-), 讲师. 研究方向为数据挖掘. E-mail: li_huiqin09@163.com.

孙英 女, (1971-), 硕士, 副教授. 研究方向为计算机应用.

王俊洁 女, (1976-), 硕士, 讲师. 研究方向为虚拟现实、计算机动画.

(上接第 129 页)

[7] ZHAO Chenchen, GONG Zhihui, LONG Yican, et al. The oblique stereo image matching with Shi-Tomasi features[J]. Journal of Geomatics Science and Technology, 2015, 32(1): 66-71.

[8] Chen X H, Jia Y M. Indoor localization for mobile robots using lampshade corners as landmarks: visual system calibration, feature extraction and experiments [J]. International Journal of Control, Automation, and Systems, 2014, 12(6): 1313-1322.

[9] Rublee E, Rabaud V, Konolige K, et al. ORB: An ef-

ficient alternative to SIFT or SURF[C] // Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011: 2564-2571.

作者简介:

王晓彤 女, (1993-), 硕士研究生. 研究方向为移动机器人视觉定位.

黄鲁 (通讯作者) 男, (1961-), 副教授, 硕士生导师. 研究方向为电路与系统. E-mail: luhuang@ustc.edu.cn.