

混合云环境下数据流关联规则挖掘算法

李慧琴, 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: 在混合云环境下数据流具有扰动性, 对关联数据的挖掘的误差较大. 针对散乱点云自适应压缩挖掘算法的抗干扰性不好的问题, 提出一种基于数据信息流相空间重构和关联规则特征提取的混合云环境下数据流挖掘算法. 首先构建混合云环境下大数据信息流时间序列分析模型, 进行数据结构分析, 然后对混合云环境下数据信息流进行高维相空间重构, 在重构的相空间中关联规则特征提取, 以提取的特征作为信息素引导数据定位挖掘, 实现数据挖掘算法改进. 最后通过仿真实验进行性能测试, 结果表明, 采用该算法进行混合云环境下数据挖掘的特征提取聚敛性较好, 数据挖掘的准确概率较高, 抗干扰性能较强, 优于传统方法.

关键词: 混合云环境; 数据挖掘; 关联规则; 特征提取; 相空间重构

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)11-0152-05

Data Stream Association Rule Mining Algorithm in Hybrid Cloud Environment

LI Hui-qin, WANG Jun-jie

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

Abstract: In a hybrid cloud environment, data flow has disturbance, the anti jamming problem of adaptive compression mining algorithm for scattered point cloud is not good, a hybrid cloud environment for extracting data information flow is proposed based on phase space reconstruction and association rule mining algorithm under the characteristics of data stream. First, to construct a hybrid cloud environment information data flow time series analysis model for data structure analysis, and then the hybrid cloud environment data flow information flow for high dimensional phase space reconstruction of association rules in the feature extraction phase space reconstruction, with the extracted features as pheromone positioning guide data mining, data mining algorithm is improved. Finally through the simulation experiment of performance test, the results show that, by using the algorithm in hybrid cloud data mining features extraction has better convergent, higher accurate probability of data mining, anti-interference performance is better than traditional methods.

Key words: hybrid cloud environment; data mining; association rule; feature extraction; phase space reconstruction

1 引言

随着云计算技术的发展, 采用云计算进行大数据信息处理, 能提高数据处理和并行计算的能力, 混合云是建立在公有云和私有云基础上的混合云计算平台, 将公有云和私有云进行混合和匹配, 以获得高效便捷的大数据信息处理效果. 在混合云构建和云计算环境下, 大数据信息流受到网络数据库和本地

数据库中信息流的小扰动和互相关信息特征量的干扰, 导致在混合云环境下对数据流的挖掘准确度不高, 干扰较大, 需要进行混合云环境下数据流的优化挖掘算法设计, 提高对数据库访问和信息检索的能力, 研究相关的混合云环境下的数据挖掘算法同样在云计算任务调度和数据库系统设计中具有重要的意义和价值^[1].

对混合云环境下数据流的挖掘的研究本质还是

数字信息的处理和统计分析,混合云环境下数据流挖掘过程中受到的扰动影响因素较多,具有时变性和随机性变化特点,挖掘的难度较大,准确性不好.传统方法中,通常采用的数据挖掘算法主要有 AR-MA 模型方法^[2]、支持向量机算法、粒子群聚类挖掘算法、FCM 算法等^[3-4],上述方法采用非线性时间序列分析进行混合云环境下的数据结构分析和数据聚类处理,实现数据挖掘目标特征的定位和提取,满足用户的挖掘目的性需求,其中,文献[5]提出一种基于切片原理的海量点云数据简化挖掘算法,通过对海量点云数据的矢量空间重组和特征压缩,降低了数据挖掘的计算开销,提高了实时性,但是该方法在受到较大的网络环境干扰时,数据挖掘的准确度不高;文献[6]提出一种基于小波技术的散乱点云自适应压缩的数据挖掘算法,采用 Apriori 算法对混合云构架下的数据频繁项日志进行关联规则提取和特征检索,采用粒子群优化算法进行数据聚类中心的自适应寻优,采用正态相关状态特征提取,算法同样存在抗干扰性不好,计算开销较大的问题.针对上述问题,本文提出一种基于数据信息流相空间重构和关联规则特征提取的混合云环境下数据流挖掘算法,首先构建混合云环境下大数据信息流时间序列分析模型,对进行数据结构分析,然后对混合云环境下数据信息流进行高维相空间重构,在重构的相空间中进行关联规则特征提取,以提取的特征作为信息索引引导数据定位挖掘,实现数据挖掘算法改进,最后进行了仿真测试,得出有效性结论.

2 数据结构及数据时间序列分析

2.1 混合云环境下关联规则数据流存储结构分析

为了实现混合云环境下数据信息流的关联规则优化挖掘,首先进行数据挖掘模型的总体设计.在大型云存储数据库中,数据通过混合云架构实现信息存储和数据调度,采用基于云计算的 OptorSim 结构进行数据流的信息拓扑和数据结构分析,进行混合云环境下数据流的统计分析.混合云环境下数据流是一组非线性时间序列,采用非线性时间序列分析方法进行空间重构和特征提取^[7],用一个四元组 G 表示混合云环境下数据流的特征分布模型,四元组为 $G = (V, E, W, c)$,采用相空间重构分析方法进行混合云环境下数据流的特征扩展和信息矢量展开,选择最小嵌入维数 m 、最佳时延 τ ,混合云架构数据的信息流模型为

$$x_n = x(t_0 + n\Delta t) = h[z(t_0 + n\Delta t)] + \omega_n \quad (1)$$

式中, $h(\cdot)$ 为混合云环境下数据流的迁移状态函数, ω_n 为数据挖掘中的干扰向量,在数据信息流分布的特征相空间中,混合云环境中的数据信息流向量量化矢量从 $x_n \rightarrow x_{n+1}$ 的演化过程反映了未知混合云环境下数据时间序列的统计特征.采用 Takens 嵌入定理^[8],构建混合云环境下数据的统计测量时间序列 $\{x(t_0 + i\Delta t)\}$, $i = 0, 1, \dots, N-1$,得到混合云架构下的数据流分布的状态空间为

$$X = [s_1, s_2, \dots, s_K]_n = (x_n, x_{n-\tau}, \dots, x_{n-(m-1)\tau}) \quad (2)$$

式中, $K = N - (m-1)\tau$,表示混合云环境下数据流时间序列的正交特征向量, τ 为对混合云环境下数据流采样的时间延迟, m 为在相空间中的嵌入维数, $s_i = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau})^T$ 为一组原始采样的数据测量序列,设定滑动时间窗口,进行混合云环境下数据流的关联规则特征分布模型构建,得到特征分布的状态矩阵为

$$X = [x(t_0), x(t_0 + \Delta t), \dots, x(t_0 + (K-1)\Delta t)] = \begin{bmatrix} x(t_0) & x(t_0 + \Delta t) & \dots & x(t_0 + (K-1)\Delta t) \\ x(t_0 + J\Delta t) & x(t_0 + (J+1)\Delta t) & \dots & x(t_0 + (K-1)\Delta t + J\Delta t) \\ \vdots & \dots & \ddots & \dots \\ x(t_0 + (m-1)J\Delta t) & x(t_0 + (1+(m-1)J)\Delta t) & \dots & x(t_0 + (N-1)\Delta t) \end{bmatrix} \quad (3)$$

式中, $x(t)$ 表示混合云环境下的采样数据时间序列, J 是相空间的时间窗函数. Δt 表示滑动时间窗口宽度.通过上述处理,构建混合云环境下关联规则数据流存储结构模型,在此基础上进行数据特征分析和挖掘算法设计.

2.2 混合云环境下大数据信息流时间序列分析

在上述进行了数据存储结构模型分析的基础上,构建混合云环境下大数据信息流时间序列分析模型,进行数据结构分析,假设混合云环境下关联规则挖掘区域中标量时间序列为 $x(t)$, $t = 0, 1, \dots, n-1$,采用特征信息量化编码进行关联规则特征分解,可以描述为

$$u = [u_1, u_2, \dots, u_N] \in R^{mN} \quad (4)$$

对关联规则特征采用最大似然估计方法进行数据流挖掘区域标定,结合模糊推理规则,求得云计算混合云环境下关联规则数据流挖掘的最大梯度差

$$AVG_x = \frac{1}{m \times n} \sum_{x=1}^n \sum_{y=1}^m |G_x(x, y)| \quad (5)$$

式中, $G_x(x, y)$ 是数据信息流中的语义指向性函数, m, n 分别是混合云环境下关联规则数据流挖掘的自相关特征系数,采用自相关系数进行关联规则挖掘的聚类中心估计,得到自相关系数为

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \quad (6)$$

在混合云构架下,给出数据信息流关联规则分布结点 $G_1 = (M_1^1, M_1^2, Y_1), G_2 = (M_2^1, M_2^2, Y_2)$,采用非线性时间序列分析方法,得到混合云环境下数据流的信道结模型表示为

$$\text{WebJaccard}(X, Y) = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X \cap Y)} \quad (7)$$

$$\text{WebOverlap}(X, Y) = \frac{P(X \cap Y)}{\min(P(X), P(Y))} \quad (8)$$

$$\text{WebDice}(X, Y) = \frac{2P(X \cap Y)}{P(X) + P(Y)} \quad (9)$$

式中, $P(X), P(Y)$ 表示在混合云环境下关联规则分布的概率密度函数, X, Y 为数据传输的负载量, $P(X \cap Y)$ 是关联规则挖掘的互信息概率密度分布, X, Y 为数据召回率。

混合云环境下关联规则数据流的控制参量采用 $I(x, y)$ 表示,采用平均混信息曲线 C 划分混合云环境下关联规则数据挖掘中的聚类中心 R_1 和 R_2 ,在语义信息所索引下,采用模糊 C 均值聚类算法进行数据特征分布的非线性时间状态重组,得到混合云环境下关联规则数据流结构的区域分布函数为

$$E^{cv}(c_1, c_2) = \mu \cdot \text{Length}(C) + \nu \cdot \text{Area}(\text{inside}(C)) + \lambda_1 \int_{\text{inside}(C)} |I - c_1|^2 dx dy + \lambda_2 \int_{\text{outside}(C)} |I - c_2|^2 dx dy \quad (10)$$

式中, c_1 和 c_2 分别两组数据属性分布下的模糊聚类中心, $\text{Length}(C)$ 表示数据的长度, $\text{Area}(\text{inside}(C))$ 表示数据分布区域的归一化时间宽度, μ, ν, λ_1 和 λ_2 表示混合云环境下数据流的自相关系数,均为大于 0 的常数。根据上述分析,采用非线性时间序列分析方法进行数据的特征匹配和融合,表示为

$$C = \text{Min}\{\max(C_i)\} \quad (11)$$

$$\sum_{j=1}^p Z_{i,j} = 1, \forall i \in (1, n), \forall j \in (1, n_i) \quad (12)$$

式中, C_i 表示数据访问时间的置信度, $Z_{i,j}$ 表示关联规则的离散采样样本序列。设混合云环境下数据流具有均匀遍历性,待挖掘的关联规则属性为 $S = \overline{X_1}, \overline{X_2}, \dots, \overline{X_k}, \dots$, 关联规则属性集分布在时间段 $T_1, T_2, \dots, T_K, \dots$ 上的最优解表示数据的挖掘的聚类中心最优解,对混合云环境下数据信息流进行高维相空间重构,在重构的相空间中进行关联规则特征提取。

3 挖掘算法改进实现

3.1 混合云环境下数据信息流相空间重构

在上述进行了混合云环境下关联规则数据流存储结构分析和数据信息流时间序列分析的基础上,进行混合云环境下的数据关联规则挖掘算法改进,提出一种基于数据信息流相空间重构和关联规则特征提取的混合云环境下数据流挖掘算法,假设混合云环境下数据流分布式数据集表示为 X ,由 m 维相空间进行特征的嵌入式重构,在有限数据集中相空间重构的矢量集为

$$X = \{x_1, x_2, \dots, x_n\} \subset R^s \quad (13)$$

式中,第 i 个动态惯性数据特征点在相空间中分布的特征向量 $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$,稳定阶段前的全局最优向量 $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})$,混合云环境下数据流特征空间的聚类权值:

$$\tilde{w}_k^i = \tilde{w}_{k-1}^i \frac{p(z_k/\tilde{x}_k^i) p(\tilde{x}_k^i/x_{k-1}^i)}{q(\tilde{x}_k^i/x_{k-1}^i)} \quad (14)$$

式中, $\tilde{x}_k^i$ 为海量点云数据初始值, $\tilde{w}_{k-1}^i$ 为权重,用非线性时间序列分析方法进行混合云环境下数据流的结构优化,得到第 i 的存储节点在 $k+1$ 时刻的分布式结构特征为

$$x_i(k+1) = x_i(k) + s \left(\frac{x_j(k) - x_i(k)}{\|x_j(k) - x_i(k)\|} \right) \quad (15)$$

对于含有 n 个样本的有限数据集,混合云环境下数据流样本 $x_i, i = 1, 2, \dots, n$ 的聚压缩矢量为

$$x_i = (x_{i1}, x_{i2}, \dots, x_{is})^T \quad (16)$$

对混合云环境下对数据流关联规则进行时间序列采样,把有限数据集 X 分为 c 类,其中 $1 < c < n$,采用模糊推理方法得到数据聚类的中心矢量为

$$V = \{v_{ij} | i = 1, 2, \dots, c, j = 1, 2, \dots, s\} \quad (17)$$

式中, V_i 为聚类中心的信道传播损失,根据数据特征点个体全局寻优性,在相空间重构的分布轨迹中得到第 i 个聚类中心矢量,得到云数据振荡衰减为

$$U = \{\mu_{ik} | i = 1, 2, \dots, c, k = 1, 2, \dots, n\} \quad (18)$$

式中, c 为非线性扰动误差; μ_{ik} 为数数据特征聚类的分岔系数。结合 FCM 聚类算法^[9-10],对混合云环境下数据信息流进行互信息特征提取,得到特征提取的目标函数为

$$J_m(U, V) = \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^m (d_{ik})^2 \quad (19)$$

式中, $(d_{ik})^2$ 为关联规则样本 x_k 与 V_i 的测度距离,

用欧式距离表示为

$$(d_{ik})^2 = \|x_k - V_i\|^2 \quad (20)$$

进一步计算混合云环境下数据流挖掘适应度值为

$$\sum_{i=1}^c \mu_{ik} = 1, k = 1, 2, \dots, n \quad (21)$$

采用差分扰动进行误差修正,在重构的相空间中引导数据定位挖掘。

3.2 特征提取及数据挖据实现

在重构的相空间中进行关联规则特征提取,混合云环境下数据流采集输出的平均互信息函数描述为

$$X_p(u) = \begin{cases} p \sqrt{\frac{1 - j \cot \alpha}{2\pi}} e^{j \frac{u^2}{2} \cot \alpha} \int_{-\infty}^{+\infty} x(t) e^{j \frac{t^2}{2} \cot \alpha - j u t \csc \alpha} dt, \alpha \neq n\pi \\ x(u), \alpha = 2n\pi \\ x(-u), \alpha = (2n \pm 1)\pi \end{cases} \quad (22)$$

式中, p 为混合云环境下数据流分布式存储阶数, α 为数据采样的时间窗口宽度. 用自相关特征描述数据特征定位的 Lyapunov 稳定传递模型表示:

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-i} + \sum_{j=0}^{M_{MA}} b_j \eta_{n-j} \quad (23)$$

式中, a_0 为混合云环境下数据流的关联规则的初始采样幅值, x_{n-i} 为混合云环境下数据流标量时序, b_j 为振荡幅值. 选择最小嵌入维数进行相空间重构,得到 $x(k)$ 在存储空间中的负载均衡模型为

$$A(t) = mx(k) + \sqrt{am} B^H(t) \quad (24)$$

式中, a 为混合云环境下数据流的域间方差系数, $B^H(t)$ 表示互信息相关函数,采用关联规则特征提取等方法进行数据挖掘的信息素定位,关联规则为

$$I(\tau) = - \sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j} \quad (25)$$

式中, τ 为数据特征分布的时间延迟, p_{ij} 表示相空间中关联规则函数分布的概率密度,到混合云环境下数据流的特征映射为 $\Phi: M \rightarrow R^{2d+1}$, 重构的混合云环境下数据流时间序列相空间模型,对于相空间轨迹的任意一点 X_n ,数据挖掘与特征提取之间的映射关系为 $x_n = (x_n, x_{n-\tau}, \dots, x_{n-(m-1)\tau})$,通过上述处理,以提取的特征作为信息素引导数据定位挖掘,实现数据挖掘算法改进。

4 仿真实验与结果分析

为了测试本文算法在实现混合云环境下数据流

广联规则挖掘中的性能,进行仿真实验,实验中,软件工具采用 Matlab2012 进行算法编程设计,实验仿真环境是: IntelCore3—530 1 GB 内存,操作系统为 Windows 7,仿真需要的混合云大数据来自于大型云存储数据库 Deep Web 200 GB,原始数据采样的长度为 1 024,采样的周期为 $T=0.04$ s,混合云环境下关联规则数据流挖掘过程中,受到网络空间特征扰动的干扰强度为 $\text{SNR}=-10$ dB,关联规则标量时间序列的基频为 100 Hz,根据上述仿真环境和参数设定,进行混合云计算环境下的数据流关联规则挖掘,得到原始的数据信息流采样的时域和频域波形如图 1 所示。

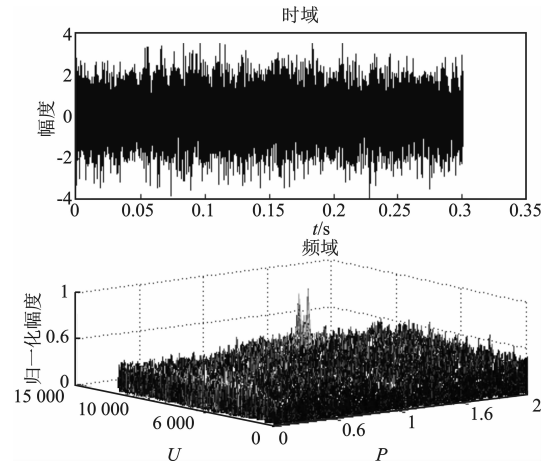


图 1 原始的数据的时域和频域波形

以上述采样的时间序列作为数据测试集,进行数据挖掘仿真分析,对混合云环境下数据信息流进行高维相空间重构,在重构的相空间中进行关联规则特征提取,得到特征提取结果如图 2 所示。

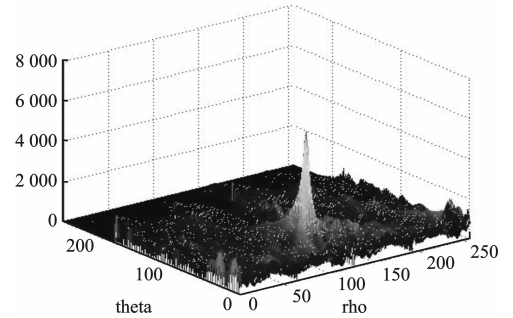


图 2 关联规则特征提取结果

从图可见,采用本文方法进行数据挖掘,通过关联规则特征提取,提高了数据的频域聚焦能力和数据聚类性能,以提取的特征作为信息素引导数据定位挖掘,为了定量对比性能,采用本文方法和传统方法,以数据流关联规则挖掘的包准确传输率和端到

端时延为测试指标,得到性能对比结果如图 3 和图 4 所示.

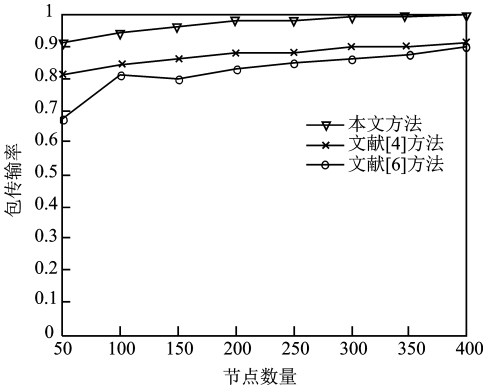


图 3 数据挖掘的包传输率

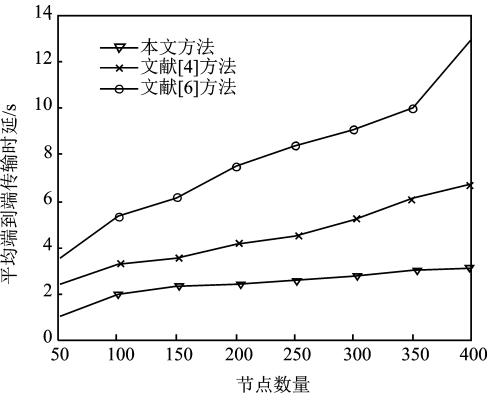


图 4 数据挖掘的端到端传输时延

分析上述仿真结果可知,采用本文方法进行数据挖掘的包传输准确概率较高,说明抗干扰性能较强,时延较低,说明实时性较好,性能指标优于传统方法.

5 结束语

为了提高混合云环境下的数据方位和挖掘能力,提出一种基于数据信息流相空间重构和关联规则特征提取的混合云环境下数据流挖掘算法,构建混合云环境下大数据信息流时间序列分析模型,在重构的相空间中进行关联规则特征提取,以提取的特征作为信息素引导数据定位挖掘,实现数据挖掘算法改进.研究得出,采用本文算法进行混合云环境

下数据挖掘的准确概率较高,包传输准确度较高,时延较低,说明数据挖掘的准确性和实时性较好,具有较高的应用价值.

参考文献:

[1] 吴涛 陈黎飞 郭躬德. 优化子空间的高维聚类算法[J]. 计算机应用, 2014, 34(8): 2279-2284.

[2] 辛宇,杨静,汤楚衡,等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展, 2015, 52(7): 1510-1521.

[3] 陶新民,宋少宇,曹盼东,等. 一种基于流形距离核的谱聚类算法[J]. 信息与控制, 2012,10(3): 307-313.

[4] 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, (5): 2015099.

[5] 官亚勤,赵学胜,王鹏飞,等. 基于切片原理的海量点云并行简化算法[J]. 计算机应用, 2016, 36(7): 1793-1796.

[6] 徐工,程效军. 基于小波技术的散乱点云自适应压缩算法[J]. 同济大学学报(自然科学版), 2013, 41(11): 1738-1743.

[7] Moradi M, Keyvanpour M R. An analytical review of XML association rules mining[J]. Artificial Intelligence Review, 2015, 43(2): 277-300.

[8] Dong G L, Ryu K S, Bashir M, et al. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction [J]. Journal of Medical Systems, 2013, 37(2): 1-10.

[9] 胡志胜,于敬武,述涛. 一种结合了栅格化和特征判断的点云压缩方法[J]. 辽宁工程技术大学(自然科学版), 2015, 34(8): 958-963.

[10] 李牧东,赵辉,翁兴伟,等. 基于最优高斯随机游走和个体筛选策略的差分进化算法[J]. 控制与决策, 2016, 31(8): 1379-1386.

作者简介:

李慧琴 女, (1975-), 讲师. 研究方向为数据挖掘、信息安全、计算机动画方向研究.

王俊洁(通讯作者) 女, (1976-), 硕士, 讲师. 研究方向为虚拟现实、计算机动画. E-mail: 1065781764@qq. com