

基于 K-L 特征压缩的云计算冗余数据降维算法

聂 军

(广东科技学院 计算机系, 广东 东莞 523083)

摘 要: 提出一种基于 K-L (Karhunen-Loeve Transform) 特征压缩的云计算冗余数据降维算法. 在冗余数据的重构相空间中进行高维特征提取, 采用 K-L 特征压缩方法降低云计算冗余数据的维数, 设计改进的 FIR 滤波算法实现冗余数据滤除. 仿真结果表明, 采用该算法能有效实现对云计算冗余数据的特征空间降维和滤除处理, 提高云计算读写速度, 降低计算开销.

关键词: 云计算; 特征压缩; 冗余数据; 降维

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2016)02-0125-05

A Data Reduction Algorithm Based on K-L Feature Compression for Cloud Computing

NIE Jun

(Department of Computer Science, Guangdong College of Science and Technology, Dongguang 523083, China)

Abstract: A K-L (Karhunen-Loeve Transform) based algorithm for reducing the dimension of redundant data in the cloud computing is proposed. High dimensional feature extraction is performed in the reconstructed phase space of redundant data, and K-L feature compression method is adopted to reduce the dimension of redundant data in the cloud computing, and the improved FIR filtering algorithm is designed to realize the redundancy data filtering. Simulation results show that the proposed algorithm can effectively achieve the feature space dimension reduction of the redundant data in the cloud computing, and can improve the speed of cloud computing, and reduce the computational cost.

Key words: cloud computing; feature compression; redundancy data; dimension reduction

1 引言

随着海量数据信息处理的要求和网络技术的发展, 人类进入大数据云计算时代. 云计算不单提供数据计算功能, 还提供和海量数据的存储和管理功能, 这就是所谓的云存储. 云存储是云计算的一个重要分支, 在云计算系统中, “云”的规模可以动态伸缩, 实现多副本容错计算和可靠性数据存储. 由于云计算系统的计算节点同构互换, 导致在数据传输通道产生大量的冗余数据, 通过对冗余数据进行降维处理, 实现冗余数据滤除, 可以降低云计算开销.

因此, 研究云计算系统的冗余数据降维方法, 实现对冗余数据滤波和删除, 在提高云计算性能, 降低计算开销方面具有重要意义^[1].

传统方法中, 对云计算系统隐通道中的冗余数据降维和删除方法主要有基于支持向量机特征分类的冗余数据降维方法、基于神经网络控制的云计算冗余数据降维方法和基于分数阶傅里叶变换的冗余数据空间降维方法等, 该类方法都是通过相空间重构把云计算中的冗余数据嵌入到高维空间中, 再采用子空间降噪和级联滤波等方法进行数据滤除和特征降维, 上述方法存在着计算开销大, 冗余数据的特

收稿日期: 2015-03-30; 修回日期: 2015-05-26

基金项目: 广东省高等学校教学质量与教学改革工程项目(粤教高函[2013]6号)“广东科技学院计算机系软件工程专业综合改革”; 东莞市 2015 年社会科学课题(2015JYZ40); 广东科技学院重点项目课题(GKY-2014KYZD-5)

征维数较高等问题^[2-3]. 对此, 相关文献进行了算法的改进设计与描述, 其中, 文献[4]提出一种基于数据多副本容错状态下冗余数据时频特征分解的删除方法, 通过时频特征分解实现数据降维和删除, 提高了计算能力, 但是该方法需要对冗余数据进行时频特征提取, 在这一过程中, 冗余数据的非线性特征是会进行干扰, 导致冗余数据删除和降维过程中出现误差和失真. 文献[5]提出一种基于熵编码层次分解的冗余数据删除和降维方法, 得到冗余数据的静态特征编码矢量空间, 在这个高维编码空间中, 采用重复数据漂移补偿方法进行数据降维和丢失信息补偿, 提高了冗余数据的滤波和降维能力, 但是该方法需要根据冗余数据的丢失信息流特征进行数据分析, 在先验知识不足的情况下, 特征分析误差较大. 文献[6]采用高阶累积量检测方法进行云计算冗余数据降维, 由于冗余数据的向量特征嵌入在高维相空间中, 导致冗余数据降维和滤除性能不好. 针对上述问题, 本文提出一种基于 K-L 变换 (Karhunen-Loeve Transform) 特征压缩的云计算冗余数据降维算法.

2 云计算中冗余数据生产机制与冗余数据信息模型构建

2.1 云计算中冗余数据生成机制与体系构架

在云计算环境下, 客户端和服务端与以虚拟化为核心的云平台结合, 进行多副本容错计算, 由于云计算的计算中枢系统随着软件技术和硬件结构的变化产生数据传输隐通道, 在隐通道中由于云计算节点同构互换和动态扩展, 产生冗余数据, 所以要进行云计算冗余数据降维处理和滤除, 提高云计算的计算能力, 在冗余数据降维处理系统中, 所开发的软件必须与云计算的应用特征相适应. 云计算中数据存储机制从上到下分别是: I/O, USB 和磁盘层, 通过移动客户端控制云计算节点管理器, 再通过组件调用实现对云计算环境下的海量数据的管理和调度^[7], 云计算系统下数据处理和存储过程描述如图 1 所示.

针对不同的用户和对象, 云资源节点分布在不同的数据隐通道模型中, 用一个四元组 G 表示云计算中冗余数据的存储通道结构模型, $G = (V, E, W, C')$, 设移动学习的云资源网络的第 i 个传输的数据

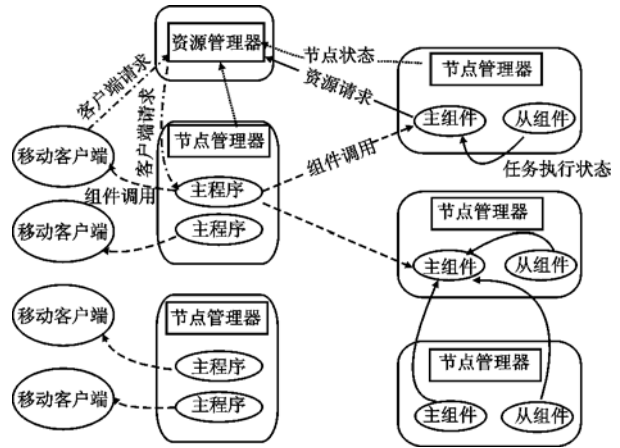


图1 云计算系统下数据处理和存储过程描述

包为 i^h , 它为静态分块数据链上传的数包, 进行文件切分, 得到不同大小的冗余数据块, 实现冗余数据的高维空间投影, 在云计算系统中, 隐通道中的冗余数据切片函数为

$$x_{id}(t+1) = wx_{id}(t) + c_1 r_1 [r_3^{t_0 > T_0} P_{id} - X_{id}(t)] + c_2 r_2 [r_4^{t_g > T_g} P_{gd} - x_{id}(t)] \quad (1)$$

式中, t_0 和 t_g 分别表示动态副本数据的边缘特征分解初始时间向量和在高维空间的迭代步数. 在节点管理器中, 通过组件调用, 得到云计算系统中的冗余数据采样时间序列为 $\{x(t_0 + i\Delta t)\}$, $i = 0, 1, \dots, N-1$, 冗余数据进行文件切换和云存储的排队规模为 $T-1$. 为避免云计算节点的干扰项对冗余数据降维的影响, 需要进行干扰滤波, 通过特征空间降维, 实现对云计算中冗余数据滤除. 通过上述分析云计算中冗余数据生成机制和进行体系构架的基础上, 进行冗余数据信号处理和特征分析.

2.2 云计算中冗余数据的相空间重构和特征提取

在上述进行系统结构分析的基础上, 进行冗余数据的相空间重构分析, 采用信号处理算法实现冗余数据信息处理. 采用校验信息时间-尺度耦合控制方法, 得到冗余数据的相似度特征为 $\{x(t_0 + i\Delta t)\}$, $i = 0, 1, \dots, N-1$, 在云计算系统中, 由于云端接口需要进行分布式中间件调度, 导致冗余数据产生时间漂移, 假设冗余数据信息流的时间序列长度 N , 把有限数据集 X 分为 c 类, 得到云计算下冗余数据的相空间重构表达式为

$$X = [x(t_0), x(t_0 + \Delta t), \dots, x(t_0 + (K-1)\Delta t)] =$$

$$\begin{bmatrix} x(t_0) & x(t_0 + \Delta t) & \cdots & x(t_0 + (K-1)\Delta t) \\ x(t_0 + J\Delta t) & x(t_0 + (J+1)\Delta t) & \cdots & x(t_0 + (K-1)\Delta t + J\Delta t) \\ \vdots & \cdots & \ddots & \cdots \\ x(t_0 + (m-1)J\Delta t) & x(t_0 + (1+(m-1)J)\Delta t) & \cdots & x(t_0 + (N-1)\Delta t) \end{bmatrix} \quad (2)$$

式中, $x(t)$ 表示冗余数据的时间序列模型, J 是相空间重构的空间分割尺度, m 是重构嵌入维数. 在构建的冗余数据高维相空间中^[8], 进行云计算冗余数据特征提取, 假设冗余数据降维过程中的有限矢量集合为

$$X = \{x_1, x_2, \cdots, x_n\} \subset R^s \quad (3)$$

在云计算系统冗余数据有限矢量集合中, 通过对数据流形向量的遍历历经, 得到数据集中的冗余特征是含有 n 个样本的有限元结合, 其中数据的采样样本 $x_i, i = 1, 2, \cdots, n$, 状态空间的高维特征矢量为

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{is})^T \quad (4)$$

在不同的量化步长下把冗余数据进行熵编码, 对编码矢量分为 c 类, 其中 $1 < c < n$, 采用冗余数据的高阶累积量特征压缩处理方法, 得到数据的聚类中心为

$$V = \{v_{ij} | i = 1, 2, \cdots, c, j = 1, 2, \cdots, s\} \quad (5)$$

式中, V_i 为冗余数据的第 i 个干扰向量, 对聚类中心矢量进行 SVD 分解, 得到分解结果为

$$U = \{\mu_{ik} | i = 1, 2, \cdots, c, k = 1, 2, \cdots, n\} \quad (6)$$

定义冗余数据的降维目标函数:

$$J_m(U, V) = \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^m (d_{ik})^2 \quad (7)$$

式中, m 为加权重, $(d_{ik})^2$ 为冗余数据采样标量样本 x_k 与 V_i 的欧式距离, 为

$$(d_{ik})^2 = \|x_k - V_i\|^2 \quad (8)$$

当满足

$$\sum_{i=1}^c \mu_{ik} = 1, k = 1, 2, \cdots, n \quad (9)$$

分析约束条件式(7), 结合中心极限定理, 求得冗余数据特征空间压缩的目标函数的极值为

$$\mu_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \quad (10)$$

$$V_i = \frac{\sum_{k=1}^m (\mu_{ik})^m x_k}{\sum_{k=1}^n (\mu_{ik})^m} \quad (11)$$

在干扰向量抑制的情况下, 给定初始值, 结合模糊度指标 m , 实现对冗余数据的特征提取.

3 基于 K-L 变换的特征压缩和冗余数据降维滤波算法改进实现

3.1 K-L 变换的特征压缩

根据上述进行冗余数据特征提取结果发现, 由于通过高维相空间重构, 导致冗余数据的特征矢量嵌入在高维空间中, 需要进行特征压缩和降维处理, 降低计算开销, 传统方法采用高阶累积量检测方法进行云计算冗余数据降维, 由于冗余数据的向量特征嵌入在高维相空间中, 导致冗余数据降维和滤除性能不好. 为了克服传统方法的弊端, 本文提出一种基于 K-L 变换特征压缩的云计算冗余数据降维算法. K-L 变换 (Karhunen-Loeve Transform) 是建立在统计特征分析基础上的离散时频变化, K-L 变换是一维变换, 它在数据压缩技术中占有重要地位. 本文采用 K-L 变换进行冗余数据的特征压缩过程描述所示.

令 T_i 表示 i^{th} 的冗余数据在云计算系统中传输时间, T_d 表示它传输时间的期望平均, 采用 K-L 特征压缩变换, 得到冗余数据分布的随机场中的时间检测间隔表示为 $j_i = T_i - T_d$, 其中 T_i 为直接收发双方相关信道联通因素, 在高维特征空间中的构建向量分布随机场. 计算冗余数据的 l 维特征向量 $X(l, n_i)$ 在信源中的协方差矩阵 $\hat{S}_w$, 求出特征空间能量集中的 l 个特征值 $\lambda_1, \lambda_2, \cdots, \lambda_l$ 和相应的在高维相空间中的特征向量矩阵 $Y = [y_1, y_2, \cdots, y_l]$. 假设 $\hat{S}_w$ 为

$$\hat{S}_w = \sum_{i=1}^c p_i \frac{1}{n_i} \sum_{k=1}^{n_i} [(X_k^{(i)} - m_i)(X_k^{(i)} - m_i)^T] \quad (12)$$

式中, p_i 表示第 i 类冗余数据的先验目标信息尺度, n_i 表示同于数据的控制参量样本数, m_i 表示第 i 类冗余信息特征状态 n_i 个的压缩特征 $\{X_k^{(i)}, k = 1, 2, \cdots, n_i\}$.

在给定的高阶嵌入维数下, 计算冗余数据的类间聚敛度矩阵:

$$S_b = \sum_{i=1}^c p_i (m_i - m)(m_i - m)^T \quad (13)$$

式中, $m = \sum_{i=1}^c p_i m_i$ 是表示冗余数据的量化分解定量化递归向量, 对上述特征参量并阈值大小方式进行排列: $J(X_1) \geq J(X_2) \geq \cdots \geq J(X_l)$; 取前 d 个较大

的 $J(\mathbf{X}_j)$ 冗余数据状态特征, 降到 d 维面, 实现对冗余数据高维特征空间的 K-L 特征压缩, 最后生成特征压缩后的状态空间矩阵:

$$\mathbf{W} = [y_1, y_2, \dots, y_d] \quad (14)$$

采用通过上述处理, 云计算中的冗余数据特征向量由 l 降到 d 维, 从而减低了计算维度和开销。

3.2 冗余数据滤除实现

在上述进行特征降维的基础上, 进行冗余数据滤波处理, 用 $\mathbf{V} = [v_1, v_2, \dots, v_n]$ 表示冗余数据的特征向量, 采用 K-L 变换特征压缩算法, 产生一组云计算任务调度数据集 V 后, 采用 p_i 表示 $x(t)$ 出现在冗余数据分布区间 i 中产生的冗余数据的概率。 $p_{ij}(\tau)$ 表示信号 $x(t)$ 出现在区域 i 的时间延迟, 为了提高对冗余数据的滤除能力, 采用 FIR 滤波系统, 当 V 为连续变量, 根据冗余数据的特征压缩和降维结果, 进行特征分解, 得到冗余数据滤波函数为

$$H(t) = F^\alpha[x(t)] = \int_{-\infty}^{\infty} K_p(t, u)x(t)dt \quad (15)$$

式中, p 为冗余数据滤除的高阶傅里叶函数域, 为一实数, 冗余数据特征旋转向量角度 $\alpha = p\pi/2$ 。 $F^\alpha[\cdot]$ 表示为变换算子, $K_p(t, u)$ 是 IMF 分量。冗余数据在相空间中训练样本集为 $X = [X_1, X_2, \dots, X_k, \dots, X_N]^T$, 信息流特征编码为 $X_k = [x_{k1}, x_{k2}, \dots, x_{km}, \dots, x_{kM}]$, 冗余数据通过 K-L 特征压缩和滤波处理后, 得到输出向量模型为

$$F_\alpha(m, n) = A_\alpha e^{(j/2)\cot\alpha \cdot m^2\Delta u^2} \cdot e^{(j/2)\cot\alpha \cdot n^2\Delta t^2} \cdot e^{-j\frac{\text{sgn}\sin(\alpha) \cdot 2\pi \cdot m \cdot n}{N}} \quad (16)$$

式中, A_α 为幅值, 综上分析, 实现了算法的改进设计, 改进算法的实现流程伪代码步骤如下所示。

- 1: 初始化 $df_{\text{model}} = \text{explode}(D * h_{\sigma_s} * h_{\sigma_f})$
- 2: 初始化区域 I 的中心为目标位置 (x, y) 。
- 3: for $f = 2 \rightarrow |V|$ do
- 4: $df = \text{explode}(f) * h_{\sigma_s} * h_{\sigma_f}$
- 5: $(x', y') = \text{argmax}_{(x, y)} C(df, df_{\text{model}})$
- 6: $(x, y) = (x', y')$
- 7: $df_{\text{model}} = \rho df_{\text{model}} + (1 - \rho) df(x, y)$
- 8: end for

4 仿真结果及分析

为了测试本文算法在实现云计算中冗余数据降维和滤除中的性能, 进行仿真实验。在云计算节点同

构互换和动态扩展过程中, 产生数据隐通道, 数据隐通道中的冗余数据格式分别为 DOC、TXT 等, 对云计算系统隐通道中的处理数据进行采样, 得到原始数据的时间序列波形如图 2 所示。

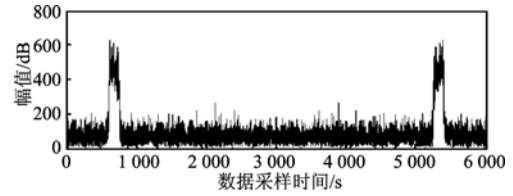


图2 云计算系统隐通道中原始数据的时间序列波形

以上述数据为测试样本, 进行冗余数据降维和滤除仿真实验分析, 从图 2 可见, 原始的数据集中含有大量的冗余数据, 导致目标数据的特征细节不能得到有效体现, 云计算性能不好, 对此, 采用本文方法, 进行相空间重构, 把冗余数据信息嵌入到高维特征空间中, 得到数据高维相空间重构结果如图 3 所示。

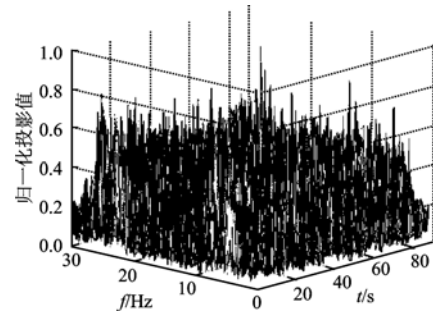


图3 云计算数据高维相空间重构

以图 3 重构相空间为基础, 进行 K-L 特征压缩, 实现冗余数据降维处理, 得到特征压缩结果如图 4 所示。

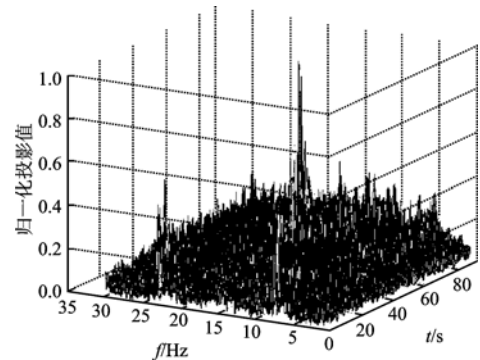


图4 冗余数据特征压缩结果

从图 4 可见, 采用本文方法, 通过 K-L 特征压

缩,冗余数据的状态特征的谱峰值较为尖锐,空间聚焦性能较好,降低了数据特征空间维数,提高了冗余数据的检测和滤除能力,以此为基础,进行冗余数据滤波,得到冗余数据滤除后的云计算隐通道数据输出结果如图5所示.从图5可见,采用本文方法实现了冗余数据滤除,降低了云计算开销.

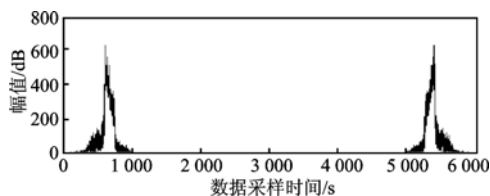


图5 冗余数据滤除后的输出波形

为了对比算法性能,采用不同方法,以相同规模下的数据量进行云计算,分析计算时间开销和云计算的读写速度,对比结果如图6所示.定量分析可知,采用本文方法,比传统方法节约时间开销14.87%,读写速度提高明显,展示了较好的应用价值.

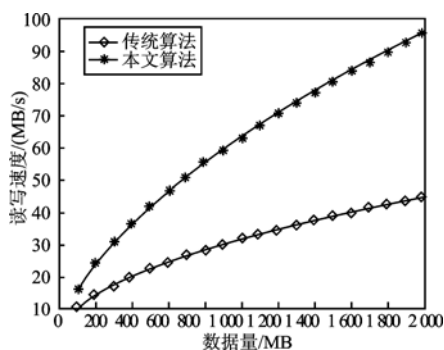


图6 计算性能对比

5 结束语

在云计算系统中,由于云计算节点同构互换和动态扩展,导致在数据传输隐通道产生大量的冗余数据,通过对冗余数据进行降维处理,实现冗余数据滤除,可以降低云计算开销.传统方法采用高阶累积量检测方法进行云计算冗余数据降维,由于冗余数据的向量特征嵌入在高维相空间中,导致冗余数据降维和滤除性能不好.本文提出一种基于 K-L 变换

特征压缩的云计算冗余数据降维算法,分析了云计算中冗余数据生成机制与网络体系构架,进行了冗余数据的相空间重构和特征提取,采用 K-L 特征压缩方法可降低云计算冗余数据的维数,基于 FIR 滤波实现冗余数据滤除.实验结果表明,采用该算法,能有效实现对云计算冗余数据的特征空间降维和滤除处理,提高云计算读写速度,降低计算开销.

参考文献:

- [1] 谢平. 存储系统重复数据删除技术研究综述[J]. 计算机科学, 2014, 41(1): 22-30.
- [2] 蒋海波, 王晓京, 范明钰, 等. 基于水平纠删码的云存储数据布局方法[J]. 四川大学学报: 工程科学版, 2013, 45(2): 103-109.
- [3] 敖莉, 舒继武, 李明强. 重复数据删除技术[J]. 软件学报, 2010, 21(5): 916-929.
- [4] 付印金, 肖依, 刘芳. 重复数据删除关键技术研究进展[J]. 计算机研究与发展, 2012, 49(1): 12-20.
- [5] FREY D. Probabilistic de-duplication for cluster-based storage systems[C]//The 2012 ACM Symposium on Cloud Computing. New York: ACM Press, 2012: 315-321.
- [6] WU Y, LEE W, LIN C. Cloud storage performance enhancement by real-time feedback control and de-duplication[C]// The 2012 Wireless Telecommunications Symposium. London: WTS, 2012: 1-5.
- [7] 周悦, 岳林蓓, 张力心, 等. 基于粒子群变异的克隆选择算法的结构故障检测与分类[J]. 信息与控制, 2015, 44(4): 436-441.
- [8] Rotondo D, Puig V, Nejari F, et al. A fault-hiding approach for the switching quasi-LPV fault tolerant control of a four-wheeled omnidirectional mobile robot[J]. IEEE Transactions on Industrial Electronics, 2014, 99(1): 1-11.

作者简介:

聂军 男, (1976-), 硕士, 讲师, 高级工程师. 研究方向为计算机软件与应用、云计算与应用、智能计算理论及应用.
E-mail: 13739149@qq.com