

大规模高维数据集中局部异常数据挖掘算法

张凯斐¹, 刘继华¹, 张菊芳²

(1 吕梁学院 计算机科学与技术系, 山西 吕梁 033000;

2 山西能源学院 基础部, 山西 太原 030000)

摘 要: 提出一种基于 FFD 的大规模高维数据集中局部异常数据挖掘算法. 将 FFD 首次应用在挖掘中, 通过引用无线传输技术, 将所提方法的宗旨定为对作业级与任务级的实现, 以提高局部异常数据抗干扰能力. 所提方法利用 FFD 的强控制能力实现无线传输技术与挖掘进程的数据互通, 利用 FIFO 挖掘思想依次进行数据本地化与挖掘, 并对挖掘流程与目标函数进行了重点设计. 实验结果证明, 所提方法的可靠性强, 挖掘效率高, 挖掘任务完成量大.

关键词: 大规模高维数据集; 局部异常数据; 挖掘算法; 本地化

中图分类号: TP311.52

文献标识码: A

文章编号: 1000-7180(2018)03-0116-04

Local Outlier Mining Algorithm for Large Scale High Dimensional Data Set

ZHANG Kai-fei¹, LIU Ji-hua¹, ZHANG Ju-fang²

(1 Department of Computer Science and Technology, Lvliang University, Lvliang 033000, China;

2 Basic Courses, Shanxi Institute of Energy, Taiyuan 030000, China)

Abstract: This paper proposes a mining algorithm of large scale and high dimensional data based on FFD local concentration of abnormal data. The FFD was first applied in mining, by referencing the wireless transmission technology, the proposed method is the aim for the realization of the working class and the task level, in order to improve the anti-jamming ability of local abnormal data. The data exchange method based on robust control ability of FFD wireless transmission technology and the data mining process, followed by local mining and using FIFO mining method and the mining process and the objective function of the key. The experimental results show that the proposed method has strong reliability, high mining efficiency and large amount of mining tasks.

Key words: Large scale high dimensional data set; Local outlier data; Mining algorithm; Localization

1 引言

数据挖掘是人工智能、模式识别、机器学习和统计学等学科的交叉, 是适应信息化社会从海量的数据集中, 提取有用知识的需要而产生的新的学科. 随着对局部异常数据挖掘研究的深入, 出现了许多异常检测技术, 在实际应用领域得到了广泛应用^[1]. 虽然局部异常数据挖掘技术近年来取得了一定的发展, 但因为均存在各种各样的问题, 局部异常数据挖掘方法的应用十分有限. 对此, 本文提出一种新的大规模高维数据集中局部异常数据挖掘算法.

2 无线传输技术

分析无线传输技术原理有利于帮助人们提高对挖掘算法的理解能力, 只有得知异常输出数据的出处与特点, 才能够合理响应用户需求, 实现高水准的局部异常数据挖掘^[2]. 图 1 表述的是无线传输技术原理.

如图 1 所示, 图 1 以无线传输技术基本传输结构为例讨论其原理, 传感线圈是其传输媒介, 无线传输技术通过传感线圈的耦合能力, 将主电路中产生的高频逆变传输给副电路^[3]. 高频逆变通过谐振电

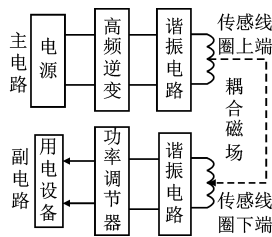


图 1 无线传输技术原理

路后会产生电动势,电动势与传感线圈上端发生感应,在传感线圈上、下端之间产生耦合磁场,顺着耦合磁场方向进行传输,由副电路上的谐振电路进行接收,通过功率调节器将作用在用电设备上。

在大规模高维数据集中,传感线圈上、下两端之间的隔空性以及传输路径上干扰的不可控制性将导致传输距离与传输精度等性能的无限制降低,这些不利影响在进行局部异常数据传输中尤为严重^[3],数据挖掘在其中的主要作用就是通过调整数据传输结构来赋予数据本身一定的抗干扰能力,大力提高传输性能。在提出大规模高维数据集中局部异常数据挖掘新算法时应首要考虑两点要求,如图 2 所示。

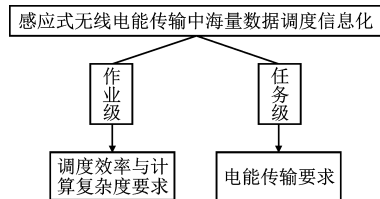


图 2 挖掘要求描述

有关作业级和任务级的意义表述如下：

- (1)作业级,指在保证数据传输质量的基础上,加快挖掘效率,减少计算复杂度;
- (2)任务级,指新方法本身不应对无线传输技术原理做出较大变更^[4],防止传输任务不可用与不可行。

3 大规模高维数据集中局部异常数据挖掘新方法

3.1 基于 FFD 的局部异常数据挖掘算法

文献[5]曾将 RFD(Request for Discussion,精简功能设备)应用在数据挖掘中,虽然 RFD 不具备数据传输能力,但对协调传感线圈干扰具有很好的作用。将 FFD 应用于数据挖掘是一次大胆尝试,FFD 即“Full Functional Device”,译名为“完整功能设备”,这是一种拥有数据收发与处理功能的信息网

络^[6],与 RFD 相比,FFD 的控制能力更强,能够实现与无线传输结构的数据互通,并大力提高传输数据抗干扰能力。

将局部异常数据在 FFD 信息网络中进行挖掘处理的难点是:非本地化的挖掘工作无法获取较高效率^[7]。因此,将数据进行本地化处理在所提基于 FFD 的局部异常数据挖掘中是一项非常重要的任务。FFD 在进行非本地向本地化的转换处理上具有明显优势,这实际上相当于一个格式转换问题与数据转存问题^[8]。为了提高挖掘效率,格式转换与转存用时是需要严格限定的,用时过长会导致挖掘时间延长,无法实现作业级,用时过短会导致传输任务不可用、不可行,无法实现任务级。

3.2 局部数据挖掘流程与目标函数设计

设无线传输结构中含有 3 个传输任务,这 3 个传输任务中含有局部异常数据,需要进行挖掘。所提方法分数据本地化与挖掘两个步骤进行^[9],具体流程如图 3、4 所示。

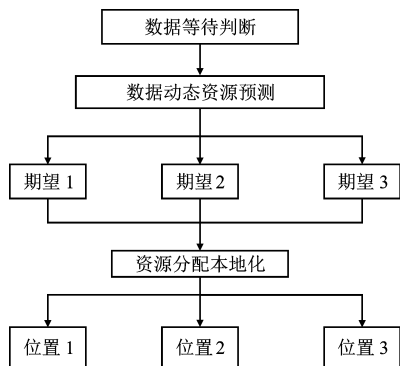


图 3 数据本地化

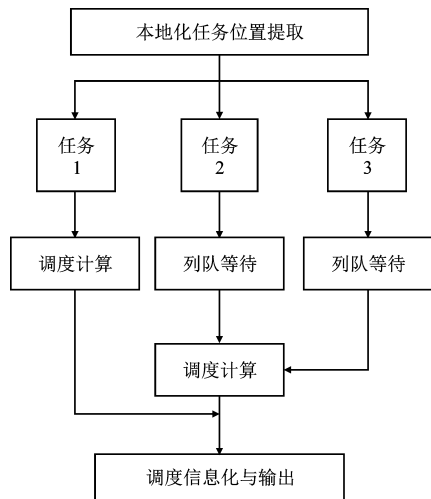


图 4 挖掘算法

由图 3 可知,所提方法先将无线传输任务引入

FFD 信息网络中,按顺序等待进行数据资源预测.数据资源预测的目的是提取传输任务中的有效数据,用来表征资源对任务的可用性,即实现任务级^[10]. FFD 只对可用性任务数据进行本地化,在局部异常数据挖掘中节约了大量时间,可实现作业级.

数据期望的计算公式为

$$E_j = H \times T \times P_j \times P_h \tag{1}$$

式中, j 为任务数据列队次序; H 为挖掘请求平均值; T 为挖掘起点时间; P_j 为第 j 个任务数据在 FFD 节点上的存储量; P_h 表示挖掘请求概率.

设 N 是 3 个任务上的数据总量, k 是任务数据原存储量, S 是格式转换与转存用时范围,有

$$H = \frac{TN}{3} \tag{2}$$

$$P_j = \frac{kj}{N} \tag{3}$$

$$P_h = \frac{3S}{T} \tag{4}$$

将以上三个公式带入式(1),可得数据期望目标函数,即:

$$E_j = \frac{STNkj}{N} \tag{5}$$

因为 S 是一个范围值,所以数据期望 E_j 的计算结果也是一个范围值,将这个范围中的任务数据提取出来转存在 FFD 信息网络中,令数据本地化得以实现.

由图 4 可知,FFD 设计一种 FIFO(First Input First Output,先入先出队列) 挖掘思想,这样设计有利于提高无线传输精度,减少漏传、误传. 在提取到无线传输任务数据后,先将第 1 个任务数据列入挖掘程序中,让其他任务列队等待,当第 1 个任务数据完成挖掘计算,顺次进行其他任务挖掘计算,最后将计算结果汇总输出.

挖掘计算需要使用两个参数,即 P_{rj} 和 R_j . R_j 表示与第 j 个任务数据同时存在于在相同 FFD 存储节点上的其它未挖掘数据数量, P_{rj} 表示第 j 个任务数据与其他数据之间的挖掘竞争概率,有

$$P_{rj} = \frac{R_j}{k_j} \tag{6}$$

当 $P_{rj} \geq 0.75$ 或 $P_{rj} \leq 0.25$ 时,则第 j 个任务数据不值得进行优先挖掘,否则,将第 j 个任务数据引入优先挖掘序列中,按照 P_{rj} 值由高到低的顺序排列. 输出原则同样使用 FIFO,优先对 P_{rj} 值高的数据进行挖掘,实现 FFD 对局部异常数据的挖掘.

4 实验

4.1 实验目的

为了对所提基于 FFD 的大规模高维数据集局部异常数据挖掘算法进行深度研究,验证首次将 FFD 应用在数据挖掘的可行性,设计如下三个实验目的:

- (1)对基于 FFD 的无线传输中局部异常数据挖掘算法的功能进行测试;
- (2)对所提方法挖掘效率进行对比验证;
- (3)对所提方法挖掘任务的完成量进行对比验证.

4.2 功能测试

4.2.1 功能测试参数设置

在基于 FFD 的大规模高维数据集局部异常数据挖掘算法中,功能测试是对所提方法进行传输测试、数据本地化测试以及挖掘测试. 传输测试是指对无线传输结构(如图 1 所示)的功能进行验证,是第一个测试过程,首先将无线传输结构与网络服务器相连,在服务器上登录管理员账号,登录成功后开始依次进行三项功能测试. 为了保证网络参数、计算机性能、数据传输带宽等条件不对本次实验造成影响,实验参数应按照表 1 中的约束严格设置.

表 1 功能测试实验参数

实验环境	参数	描述
硬件环境	中央处理器	酷睿 4 核 3.5 GHz
	内存	3 GB
	网卡	10/100/1 000 Mb/s 自适应
	局域网	千兆
	广域网	100 兆光纤
	操作系统	WIN7
	数据库	SQL Server
软件环境	服务器内存	8GB ECC DDR4
	服务器结构	C/S 结构

4.2.2 功能测试用例

功能测试用例的作用是对方法测试期望与实际测试结果进行预先编制,并设置测试极限条件,以便在功能测试过程中更好地挖掘出所提方法的错误运行状态. 功能测试结果主要为了验证所提方法的可用性,实验对功能测试用例的要求是需要其遍布所有方法运行程序,并且极限条件需要符合实际情况,不可过于苛刻. 功能测试用例列表如表 2、表 3 所示,其中,表 2 所示为传输测试用例,表 3 所示为数据本地化测试以及挖掘测试用例.

表 2 功能测试用例表 1

用例编号	1
描述	传输测试
测试大纲	1. 用户登录 2. 数据采集与传输 3. 与挖掘程序连接状态
期望	无误

表 3 功能测试用例表 2

用例编号	2
描述	1. 数据本地化测试;2. 挖掘测试
测试大纲	1. 开启服务器与数据库的计算程序 2. 计算结果输出;3. 传输到挖掘作用方位
期望	无误

4.2.3 挖掘效率与挖掘任务完成量验证

在所提方法的挖掘效率与挖掘任务完成量实验中,FFD 默认使用 Jmeter 挖掘器^[10],延迟时间自动设置,实验参数与表 1 相同. 向 SQL Server 数据库中写入 10 个不同的数据样本,每个样本中含有的数据点均超过 5 000 点,最高数据点数量为 20 万点.

在无线传输技术的支持下,使用几种不同方法对这些数据样本进行数据挖掘,提取全程挖掘时间以及挖掘结束后的挖掘任务完成比率,如表 4、表 5 所示.

表 4 不同方法挖掘时间对照表

样本编号	挖掘时间/s		
	FFD	RFD	点云聚类
1	42.5	58.2	61.4
2	75.3	89.4	92.5
3	52.2	86.3	135.1
4	91.7	108.2	198.4
5	103.5	138.4	248.5
6	25.3	48.5	88.5
7	134.5	240.8	286.7
8	161.4	201.7	316.4
9	171.2	226.7	343.8
10	201.1	284.2	394.2
平均值	105.87	148.24	216.55

从表 4、表 5 中可以初步看出,FFD 的平均挖掘效率与平均挖掘任务完成比率均要优于 RFD 与点云聚类. 为了方便进行整个局部异常数据挖掘算法各个进程性能对照,将表 4、表 5 绘制成折线图,如图 5、图 6 所示.

表 5 不同方法挖掘任务完成比率对照表

样本编号	挖掘任务完成比率/%		
	FFD	RFD	点云聚类
1	92.54	83.25	83.28
2	95.67	82.96	63.11
3	96.75	82.04	67.79
4	92.18	79.56	67.41
5	93.89	81.18	64.93
6	96.17	84.02	68.04
7	93.25	80.46	64.29
8	95.74	82.96	66.78
9	94.78	82.63	70.48
10	95.12	79.41	62.56
平均值	94.61	81.85	67.87

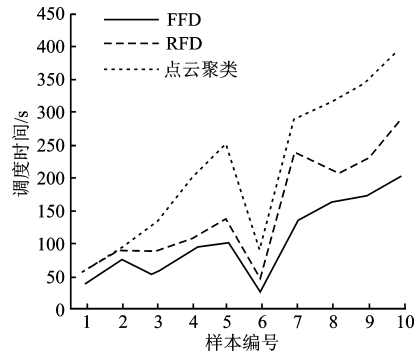


图 5 不同方法挖掘时间折线图

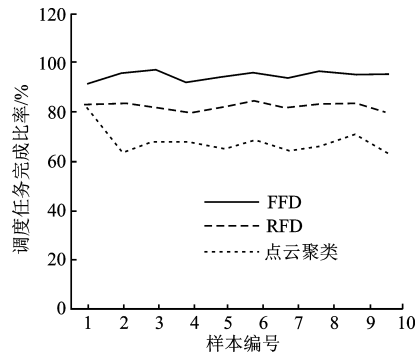


图 6 不同方法挖掘任务完成比率折线图

从图 5、图 6 所示的折线图中能够更加清晰地看出基于 FFD 的大规模高维数据集中局部异常数据挖掘算法与使用 RFD 和点云聚类进行数据挖掘算法的性能优劣情况. 结果表明,所提方法具有挖掘效率高、挖掘任务完成量大的优点,进一步确定出在无线传输技术支持下,将 FFD 应用在局部异常数据挖掘算法高度可行性,这与 FFD 提供的强控制与强抗干扰能力是密不可分的.

(下转第 124 页)

参考文献:

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504.

[2] Chen H, Qi X, Yu L, et al. DCAN: Deep contour-aware networks for object instance segmentation from histology images. [J]. Medical Image Analysis, 2017, 36:135-146.

[3] 张立民, 刘凯. 基于深度玻尔兹曼机的文本特征提取研究[J]. 微电子学与计算机, 2015, 32(2):142-147.

[4] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions [M]. World Scientific Publishing Co. Inc, 1998.

[5] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9:249-256.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// International Conference on Neural Information Processing Systems. Curran Associates Inc, 2012: 1097-1105.

[7] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[8] Lin M, Chen Q, Yan S. Network In Network[C]//

ICLR. IEEE, 2014.

[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[10] Rumelhart D E, Hinton G E, Williams R J. Learning internal representation by back-propagation of errors [J]. Nature, 1986, 323(323):533-536.

[11] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. Computer Science, 2015.

[12] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]// ACM International Conference on Multimedia. ACM, 2014:675-678.

作者简介:

杨远飞 男, (1992-), 硕士. 研究方向为深度学习. E-mail: yyf287852761@163. com.

曾上游 男, (1974-), 博士, 教授. 研究方向为非线性动力学、计算神经科学.

冯燕燕 女, (1993-), 硕士. 研究方向为深度学习.

周悦 男, (1993-), 硕士. 研究方向为深度学习.

潘兵 男, (1993-), 硕士. 研究方向为深度学习.

(上接第 119 页)

5 结束语

本文提出基于 FFD 的大规模高维数据集中局部异常数据挖掘算法, 将 FFD 应用于进行数据挖掘是一次大胆尝试, 是对 RFD 应用弊端的思考与改进. 本文在进行 FFD 挖掘算法设计时, 着重考虑了挖掘效率与挖掘任务完成比率的均衡挖掘性能, 达成了以往所提方法无法比拟的挖掘性能, 并轻松通过了传输、数据本地化以及挖掘三项功能测试, 表明本文方法具有很强的可用性, 达到了期望效果, 而且具有挖掘效率高、挖掘任务完成量大优点.

参考文献:

[1] 李永红, 张继福, 苟亚玲. 相关子空间中的局部离群数据挖掘算法研究[J]. 小型微型计算机系统, 2015, 36(3):460-465.

[2] 农晓锋. 高维数据集中局部离散文本数据挖掘方法研究[J]. 现代电子技术, 2017, 40(19):89-105.

[3] 董本清, 彭健钧. 复杂网络数据流中的异常数据挖掘算法仿真[J]. 计算机仿真, 2016, 33(1):434-437.

[4] 焦润海, 张谦, 陈超. 基于 Spark 改进的最大频繁项集挖掘算法[J]. 计算机工程与设计, 2017, 38(7):

1839-1843.

[5] 梅毅, 熊婷, 罗少彬. 复杂属性环境下 NoSQL 分布式大数据挖掘方法研究[J]. 科学技术与工程, 2017, 17(9):239-243.

[6] 张继福, 李永红, 秦啸, 等. 基于 MapReduce 与相关子空间的局部离群数据挖掘算法[J]. 软件学报, 2015, 26(5):1079-1095.

[7] 陈宏. 基于关联规则挖掘算法的用电负荷能效研究[J]. 电子设计工程, 2017, 25(4):79-82, 86.

[8] 刘继华, 强彦. 云计算下的一种数据挖掘算法的研究[J]. 科技通报, 2016, 32(12):133-137.

[9] 杨晓波. 基于投影模式支持集的数据挖掘算法研究[J]. 计算机应用与软件, 2017, 34(7):273-276.

[10] 刘晓丹. 分布式环境中保护隐私数据挖掘算法研究[J]. 自动化与仪器仪表, 2016, 32(10):155-156.

作者简介:

张凯斐 男, (1981-), 硕士研究生, 讲师. 研究方向为数据挖掘等. E-mail: zhangkaifei035@163. com.

刘继华 女, (1975-), 硕士研究生, 副教授. 研究方向为软件工程、云计算.

张菊芳 女, (1977-), 硕士研究生, 副教授. 研究方向为非线性泛函分析.