文章编号:1002-2082 (2020) 06-1305-06

基于数据挖掘的光纤通信网络异常数据检测研究

马莉莉,刘江平
(内蒙古农业大学 计算机与信息工程学院,内蒙古 呼和浩特 010018)

摘 要: 为了提高光纤通信网络中异常数据的识别能力,提出了基于熵目标函数最优化的异常数据检测算法。首先,对数据样本进行属性分类,依据异常数据特征密度指标完成邻域区间半径的选取;其次,通过对高阶统计量的大数据聚类度循环迭代,完成特征提取参数的优化;最后,由样本属性概率计算熵目标函数的最优值,并利用最优值完成异常数据检测。实验对 1 000 组通信数据进行测试,结果显示,该算法的检测精度均值约为 95.7%,其数据融合率、检测耗时与平均误检率均优于 2 种传统方法。该算法具有精度高、收敛快、误检率低的优势,具有一定的应用价值。

关键词: 光纤通信网络;异常数据识别;特征提取;熵目标函数
中图分类号: TP391.4 **文献标志码:** A **DOI:** 10.5768/JAO202041.0608003

Research on abnormal data detection of optical fiber communication network based on data mining

MA Lili, LIU Jiangping
(College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China)

Abstract: In order to improve the recognition ability of abnormal data in optical fiber communication networks, an abnormal data detection algorithm based on the optimization of entropy objective function was proposed. Firstly, the attributes of the data samples were classified, and the radius of the neighborhood interval was selected based on the abnormal data feature density index. Secondly, the clustering degree of the big data with high-order statistics was iterated and the optimization of the feature extraction parameters was completed. Finally, the optimal value of the entropy objective function was calculated according to the sample attribute probability, and it was used to complete the abnormal data detection. The experiment tested 1 000 sets of communication data. The results show that the average detection accuracy of this algorithm is about 95.7%, and its data fusion rate, detection time-consuming and average false detection rate are better than that of the two traditional methods. It can be seen that this algorithm has the advantages of high accuracy, fast convergence and low false detection rate, which has the certain application value.

Key words: optical fiber communication network; abnormal data recognition; feature extraction; entropy objective function

引言

随着计算机网络应用与光纤传感技术的不断发展,基于光纤传输的互联网络被大量应用于生活、生产等领域,由于光纤传感网络具有通信容量大、传输距离远、成本低等特点[1-2],所以得到了广泛推广及应用。随着光纤传输网络中用户端以指数量级增长,通信故障造成的异常数据也不断增多。在规模巨大的互联网数据中快速分析识别异

收稿日期:2020-04-08; 修回日期:2020-05-06
基金项目:国家自然科学基金(61703056);内蒙古教育厅项目(NJZY18062)
作者简介:马莉莉(1977–),女,硕士,副教授,主要从事计算机网络安全研究工作。E-mail: malilinuc@sina.com
通信作者:刘江平(1980–),男,副教授,主要从事软件工程及程序设计方面的研究工作。E-mail: liujiangpingnuc@sina.com

常数据并加以定位是十分困难的，因为数据传输过程中不可能将所有信息全部获取后才完成分析及信息识别，所以必然有很多信息在传输过程中被省略，这些省略的信息中不可避免的存在有效信息，从而导致信息熵的增加，降低了现有通信故障检测效率[3]。

光纤通信网络[4]中异常数据的识别需要从当前所有数据中提取异常数据的数据特征及模式结构，将异常数据特征与模式结构作为训练样本进行学习，从而对大数据中其他故障终端产生的异常数据进行快速精确识别[5]。利用该算法找出异常数据所对应的各变量间的逻辑关系。光纤传感网络中数据量巨大，采用大数据挖掘技术能够更好地将异常数据特征提取出来。传统的识别方法主要包括：BP 神经网络[6]、时序分析法[7]、遗传算法[8]、粗糙集[9] 等。BP 神经网络可适用于此类非线性问题并且具有一定的自学习能力，但在数据量很大时容易陷入局部极值问题[10]；时序分析法是基于时间顺序对数据进行统计分析的方法，本质是统计规律的总结，它是网络中异常数据分析的常见方法，对具备先验知识的数据分类简单易实现，但预测精度较差[11]；遗传算法可实现多个体同时比较，有利于多参数协调优化，其本质是参数权值的动态调整，这与网络数据交互是十分相似的，但其算子参数选择大多靠经验完成，在海量数据中容易陷入局部极值解[12]；粗糙集的最大优势是能处理不完整、不精确的数据，对不确定特征属性的识别具有一定帮助，但易受噪声影响，稳定性差[13-15]。由此可见，现有算法各有特色，但对于日渐庞大的数据规模和异常数据种类，采用信息熵作为目标函数完成异常数据的挖掘可以限定实际寻优范围。本文利用自主机器学习的数据挖掘技术实现多算法融合，设计了熵目标函数最优化算法。该算法的优势在于信息熵解算本身就是面向海量数据的，且是针对信息值的，不需要先验数据特征。

1 异常数据特征属性分类

1.1 样本属性分类

采用挖掘技术[16]中的聚类算法对异常数据进行特征聚类。设所有待检测数据点集合为 M，其中存在 N 个异常数据样本集合。异常数据样本对应的权值为 $c_j(t)$，j=1, 2, ⋯, K；异常数据聚类权值为 $c'_i(t-1)$，i=1, 2, ⋯, K'。异常数据样本对应的权值由其可能造成的错误严重程度给出，通常由数据用户提供。将 K_t 个异常数据样本 $x_j(1)$ 归类到 K 个聚类中心，则异常数据聚类中心可表示为

$$c'_i(1)=\sum_{j=1}^{K_t}\mu_{ij}c_j(1)+\sum_{i=1}^{K'}\mu_{ij}c'_i(0) \tag{1}$$

式中：μ_{ij} 是异常数据样本相对聚类中心的模糊隶属度（集合成员若被定义为 0 或 1，则成员介于 [0, 1] 之间的集合可称之为“模糊隶属度”，用于表达具有不确定性的数据。），$1\leqslant i\leqslant K'$，$1\leqslant j\leqslant K_t$。

设存在 n 个 d 维的异常数据特征集合，表示为 $X=(x_1,x_2,\cdots,x_n)$，则每个特征 x_i 所对应的密度指标可表示为

$$D_i=\sum_{j=1}^{K_t}\exp\left\{-\frac{\left\|x_i-x_j\right\|^2}{(r_a/2)^2}\right\} \tag{2}$$

式中：r_a 为异常数据特征 x_i 的邻域区间半径，设该区间中密度最大值为 x_1，则密度指标为 D_1。若 x_l 是第 l 次的异常数据聚类中心，其密度指标有 D_l，则(2)式有：

$$D_i=D_l\exp\left\{-\frac{\left\|x_i-x_l\right\|^2}{(r_b/2)^2}\right\} \tag{3}$$

式中，r_b 为异常数据特征密度指标的邻域区间半径。由此可见，异常数据的特征可由 D_{k+1}/D_1 的比值进行分选，比值越小，则其聚类结果越好。

1.2 特征提取优化

基于属性特征密度的判据可以完成特征分类，但是 D_{k+1}/D_1 阈值的选取直接影响了聚类质量，故本文设计了利用高阶统计量作为模型补偿参数的特征提取优化算法。设数据集合为 $M=\{m_1,m_2,\cdots,m_m\}$，个体最优解集合为 $P_i=\{p_{i1},p_{i2},\cdots,p_{id}\}$，全局最优解集合为 $P_g=\{p_{g1},p_{g2},\cdots,p_{gd}\}$，在则异常数据判断更新策略有

$$x_{id}^{(t+1)}=x_{id}^t+C_1*r_1\left(p_{id}^t-x_{id}^t\right)+C_2*r_2\left(p_{gd}^t-x_{id}^t\right) \tag{4}$$

式中：x_{id} 表示第 i 个节点在第 d 维中的异常数据集合中的一个数据点；$\{C_1, C_2\}$为优化加速系数；$\{r_1, r_2\}$为 [0,1] 的随机值，由此构建的模型可使数据具有更好的相关性。首先，将求解分布聚类的最大值，有

$$d_{\max}=\left|\max[d_{ij}(t)]\right| \tag{5}$$

然后，求解平均粒度，有：

$$d_{\mathrm{mid}}=\left|\sum_{j=1}^{m}\sum_{i=1}^{d}d_{ij}(t)\right|\cdot(md)^{-1} \tag{6}$$

式中: $d_{ij}(t)$为第 j 个采样点 i 维上的分布聚类; d 为异常数据维度; m 为总样本 M 中的数据个数。最后, 设高阶统计量的数据聚类度是 k, 则其函数可表示为

$$k=\frac{|d_{\max}-d_{\mathrm{mid}}|}{d_{\max}} \tag{7}$$

对 k 值的循环迭代即可实现对特征提取参数优化选择。

2　异常数据检测与实现

2.1　函数构建

在上述异常数据特征优化提取的基础上, 对光纤传感网络中的所有待测数据进行检测。由光纤故障导致的异常数据类型有很多, 使异常数据的属性结构各不相同, 故单纯采用传统的样本方差或平方差形式会造成识别误差大的问题。本文提出了基于熵目标函数最优化的异常数据检测算法, 根据光纤网络中异常数据随机性强的特点, 引入熵描述数据的不确定度, 设 t 时刻异常数据特征为 $X(t)$, 对第 i 个样本属性而言, $P(x_i(t))$为样本属性 $x_i(t)$的概率, 则熵 H 有:

$$H(x_i(t))=-\frac{n\sum_{i=1}^{n}P(x_i(t))\log_2 P(x_i(t))}{\log_2 n} \tag{8}$$

结合样本方差 S^2 有:

$$S^2(x_i(t))=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t)-\frac{1}{n}\sum_{i=1}^{n}x_i(t)\right)^2 \tag{9}$$

将(8)式和(9)式作为异常数据判别依据后, 熵目标函数有:

$$F(x_i(t))=\sum_{i=1}^{n}\left\{\alpha S^2(x_i(t))+\beta H^2(x_i(t))\right\} \tag{10}$$

式中: α 和 β 为权重系数, $\alpha+\beta=1$, $\alpha>0$, $\beta>0$。该函数即为异常数据检测函数。

2.2　算法实现

为了获得光纤通信网络中异常数据识别的最优函数值, 实现算法步骤如下:

1) 在初始时刻 $t=0$ 时, 将光纤网络中已有的异常数据特征参数载入算法, 自动检索半径设为 R, 信息特征阈值设为 T, 迭代次数为 i;

2) 将光纤网络中异常数据属性特征 $X(t)$构建的目标函数 $F(x_i(t))$作为目标值, 将 τ 设为目标的值, 构建符合检索半径 R 的适应度函数:

$$f(R,\tau)=\frac{1}{1+\mathrm{e}^{-F(x_i(t))}} \tag{11}$$

3) 带入初值后获得标的值的初值 $F(R,0)$, 经过算法迭代令 $F(R,0)$趋近 F 的全局最优解 F_{best}, 从而得到 $f(R,0)$, 判别依据有:

$$F(R,\tau)=F(R,\tau)+(1-rand())(F(R,\tau)-F(R,\tau-1)) \tag{12}$$

4) 在所有待测数据集合 M 中循环运行(9)式和(10)式, 从而获得 F_{best} 和 $f(R,0)$, 当 $f(R,\tau)>f(R,0)$, 更新 F_{best} 和 $f(R,0)$; 否则, 进入下一个数据检测循环;

5) 将 $P(x_i(t))$映射到搜索域中, 导入下式:

$$f(R,\tau)=\frac{(1+f(R,\tau))^{1-\gamma}}{R+\sum_{i}(f(R,\tau))^{\gamma}} \tag{13}$$

将输出数据与 $f(R,0)$进行比较, 若大于 $f(R,0)$, 用 $F(R,i)$替换 F_{best}, $f(R,i)$替换 $f(R,0)$; 若小于等于 $f(R,0)$, 则设 $i=i+1$, 转入判断下一个数据的异常判断概率计算, 直至结束;

6) 循环得到迭代后的 $f(R,i)$, 与 T 进行比较, 当 $f(R,i)<T$ 时, 依据 $P(x_i(t))$重新设定 τ 进行循环, 当 $f(R,i)>T$ 时, 结束循环, 输出 F_{best};

7) 将 F_{best} 带入异常数据检测函数计算标的值, 结束运行。

由此完成算法, 流程图如图 1 所示。

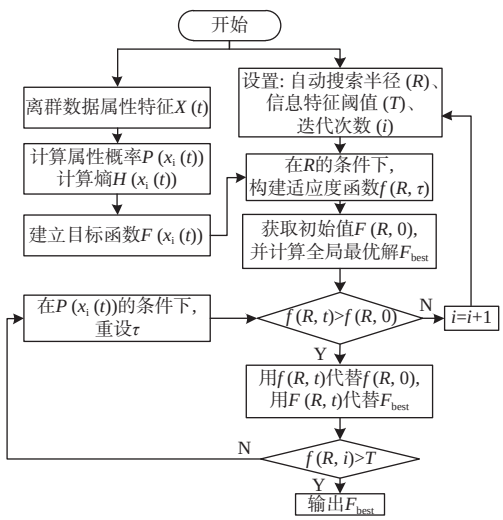

图 1　熵目标函数优化算法流程图

Fig. 1　Flow chart of optimization algorithm for entropy objective function

2.3　算法评价

为了验证算法具有更好的适用性, 主要从数据融合率、检测精度、检测耗时以及误检率 4 个方面进行分析。数据融合率用于考察算法对不同数据的融合能力; 检测精度用于考察对异常数据物理

位置的计算精度；检测耗时用于描述算法的处理速度；误检率用于检出异常点与总点数之比，是最直接反映算法性能的评价参数。

3 实验

为了验证本算法的异常数据检测精度及运算速度，实验在 VS2005 平台下利用 C#语言完成。系统包括主频 3.0 GHz 的 CPU、2 GB 的内存和 Xeon e5 型服务器。将异常数据状态信息、光时域反射仪（optical time domain reflectometer, OTDR）测试信息进行对比，从而进行评价。

3.1 数据状态分类

针对实验室光纤网络服务器系统 2019 年的通信记录信息，分别将状态信息、OTDR 测试信息以及数据占用率等进行了对比，并按照光纤网络系统中不同的状态组合进行了对比，$M\{M_1, M_2, \cdots, M_n\}$就是数据集合，则测试数据如表 1 所示。

对所有光纤通信网中的状态数据进行汇总，然后利用本算法进行分类识别，对异常数据进行标记，并与 OTDR 测试结果对比，分析算法数据识别的能力。为了保证训练效果，取 50 组正常数据（P）与 50 组各类不同异常数据（Q）构建样本，分别采用时序分析法（常用的数据规律统计方法，与其对比可以体现出本算法处理结果与数据统计规律的符合程度。）与 BP 神经网络（常用的参数权重调整方法，与其对比可以体现出本算法最终参数选择的适应度。）进行对比。

表 1 异常数据与光纤网络状态测试表

Table 1 Abnormal data and fiber network state test

	占用率%	网络状态延迟	路由状态	本算法状态判断	OTDR
M_1	75	无	连接	正常	正常
M_2	23	无	连接	正常	异常
M_3	92	有	断开	异常	异常
M_4	82	有	连接	异常	正常
…	…	…	…	…	…
M_n	46	有	连接	异常	正常

3.2 结果对比

验证集为 1 000 组随机通信数据，训练后分别求取 3 种算法的数据融合率、检测精度、检测耗时以及误检率，结果如图 2 所示。数据融合率为原始数据与融合数据之差再与原始数据的比，该指标反映了算法特征分类的能力。当在特征提取过程中选取的聚类精度不同时，相同属性的数据融合效果如图 2(a)所示，3 种方法结果在精度较低时相近，随着识别精度的提高，本算法略优于其它两种算法；如图 2(b)所示，本算法的检测精度基本不随样本个数的增大而减小，其均值约为 95.7%，时序

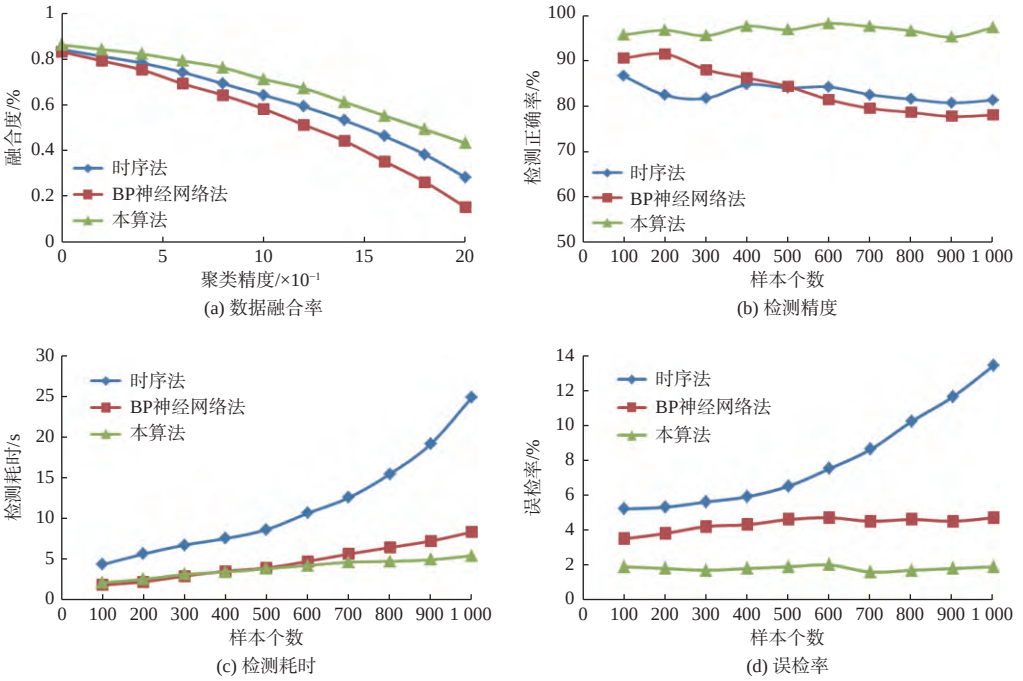

图 2 不同算法数据处理性能对比

Fig. 2 Comparison of data processing performance by different algorithms

法均值约为 84.6%, BP 神经网络法会随着数据量增大而明显下降; 如图 2(c)所示, 本算法与 BP 神经网络法的收敛时间相近, 时序法计算速度受样本总量的影响较大; 如图 2(d)所示, 本算法和 BP 神经网络法的误检率波动较小, 本算法效果最好, 平均误检率仅为 1.67%, BP 神经网络法次之, 为 4.05%, 而时序法受样本个数增加而出现较大偏差。综上所述, 本算法对异常数据的属性分类具有很好的效果, 在检测精度与误检率上相比传统方法均具有一定提升。

4　结论

本文提出了一种基于熵目标函数最优化的异常数据检测算法。利用对数据属性的特征分类完成对异常数据特征的提取, 再通过高阶统计量的大数据聚类迭代完成样本数据熵目标函数的最优计算。实验对 1 000 组光纤通信数据进行分类, 并与传统检测方法进行对比, 结果显示, 本算法在检测精度上具有明显优势, 并且在数据融合率、检测耗时以及误检率方面也略强于传统算法, 具有一定的应用价值。

参考文献:

[1] LYU Yanxia, WANG Cuirong, WANG Cong, et al. Online classification algorithm for uncertain data stream in big data[J]. Journal of Northeastern University: Natural Science, 2016, 37(9): 1245-1249.
吕艳霞, 王翠荣, 王聪, 等. 大数据环境下的不确定数据流在线分类算法[J]. 东北大学学报: 自然科学版, 2016, 37(9): 1245-1249.

[2] GU Yingying, WANG Li, HUA Baocheng, et al. 3D point cloud filtering method for pose measurement application of space non-cooperative targets[J]. Journal of Applied Optics, 2019, 40(2): 210-216.
顾营迎, 王立, 华宝成, 等. 一种面向空间非合作目标位姿测量应用的三维点云滤波算法[J]. 应用光学, 2019, 40(2): 210-216.

[3] JIA Qi. Fault location and monitoring of optical fiber lines[J]. China New Telecommunications, 2017, 19(1): 74.
贾琦. 光纤线路故障的定位和监测[J]. 中国新通信, 2017, 19(1): 74.

[4] ZHANG Taijiang, LI Yongjun, ZHAO Shanghong, et al. Design of space optical backbone network simulation platform based on OPNET and STK[J]. Journal of Applied Optics, 2019, 40(5): 901-909.
张泰江, 李勇军, 赵尚弘, 等. OPNET和STK联合的空间光骨干网络仿真设计[J]. 应用光学, 2019, 40(5): 901-909.

[5] CHEN Yang, ZHAO Shanghong, WANG Xiang, et al. BER analysis of high-altitude OFDM-FSO modulation system under exponentiated Weibull atmospheric turbulence model[J]. Laser & Infrared, 2018, 48(7): 832-837.
陈阳, 赵尚弘, 王翔, 等. 指数威布湍流模型下高空OFDM-FSO系统误码率分析[J]. 激光与红外, 2018, 48(7): 832-837.

[6] CHEN Y, LI L J. Very fast decision tree classification algorithm based on red-black tree for data stream with continuous attributes[J]. Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition, 2017, 37(2): 86- 90.

[7] ZHAO Taifei, WANG Wenke, LIU Long. A preferential shared path protection algorithm for WDM optical network[J]. Laser Technology, 2012, 36(3): 408-412.
赵太飞, 王文科, 刘龙. WDM光网络中一种优先共享通路保护算法[J]. 激光技术, 2012, 36(3): 408-412.

[8] YANG Jingming, HOU Yuhao, SUN Hao, et al. Modified NSGA-Ⅱ-DE with two-dimensional information ordering strategy and magnitude threshold[J]. Control and Decision, 2016, 31(9): 1577-1584.
杨景明, 侯宇浩, 孙浩, 等. 采用数量级阈值与二维信息排序策略的NSGA-Ⅱ-DE算法[J]. 控制与决策, 2016, 31(9): 1577-1584.

[9] GUO H P, LIU H B, WU C G, et al. Logistic discrimination based on G-mean and F-measure for imbalanced problem[J]. Journal of Intelligent & Fuzzy Systems, 2016, 31(3): 1155-1166.

[10] HUANG X, WANG Z, LI Y X, et al. Design of fuzzy state feedback controller for robust stabilization of uncertain fractional-order chaotic systems[J]. Journal of the Franklin Institute, 2014, 351(12): 5480-5493.

[11] LIU He, WANG Tao. Research on breakpoint fault detection method of optical fiber communication LAN[J]. Modern Electronics Technique, 2017, 40(16): 174-17.
刘贺, 王涛. 光纤通信局域网断点故障检测方法研究[J]. 现代电子技术, 2017, 40(16): 174-17.

[12] GUAN Xiaoying, CHEN Guo, LIN Tong. Feature selection method based on differential evolution and genetic algorithm with multi-criteria evaluation and its applica-

tions[J]. Acta Aeronautica et Astronautica Sinica, 2016, 37(11): 3455-3465.
关晓颖, 陈果, 林桐. 特征选择的多准则融合差分遗传算法及其应用[J]. 航空学报, 2016, 37(11): 3455-3465.

[13] MA Zongmei, ZHANG Ruiping. Traffic anomaly identification of optical fiber communication based on big data background[J]. Laser Journal, 2019, 40(7): 75-78.
马宗梅, 张睿萍. 大数据背景的光纤通信流量异常辨识研究[J]. 激光杂志, 2019, 40(7): 75-78.

[14] LIU Xuejun, YUAN Bixian, ZHUO Sichao, et al. Multimode optical fiber short distance data acquisition communication system based on MSP430[J]. Electronic Design Engineering, 2017, 25(6): 96-99.
刘学君, 袁碧贤, 卓思超, 等. 基于MSP430的多模光纤短距离数采通信系统[J]. 电子设计工程, 2017, 25(6): 96-99.

[15] CHEN Junda, WANG Tianshu, ZHANG Xinmeng, et al. Free space optical communication system based on wide-spectrum partially coherent laser[J]. Journal of Applied Optics, 2019, 40(1): 157-161.
陈俊达, 王天枢, 张欣梦, 等. 自由空间宽谱部分相干光通信系统[J]. 应用光学, 2019, 40(1): 157-161.

[16] JU Chunhua, ZOU Jiangbo. An incremental classification algorithm for data stream based on information entropy diversity measure[J]. Telecommunications Science, 2015, 31(2): 92-102.
琚春华, 邹江波. 基于信息熵差异性度量的数据流增量集成分类算法[J]. 电信科学, 2015, 31(2): 92-102.