



基于决策树的企业信息系统故障自动诊断分析方法

金鑫, 闫龙川, 刘军, 张书林

(国家电网公司信息通信分公司, 北京 100761)

摘要: 目前, 大型企业信息系统规模和复杂度快速增长, 但对故障的诊断分析仍主要依赖传统的人工经验, 这不仅耗时、耗力, 还影响对故障的及时处理。针对这一问题, 创新性地提出了基于决策树的企业信息系统故障自动诊断分析方法, 根据信息系统运行监控指标告警信息, 实现对信息系统故障的自动诊断。利用某大型国有企业的实际生产运行数据, 提取典型告警数据特征对该方法进行了验证, 并在 R 语言环境下对决策树模型及其训练方法进行了仿真和对比分析。实验结果证明, 该方法可以较为准确地实现故障自动快速诊断, 有助于提高信息系统故障诊断分析效率。

关键词: 自动诊断分析; 信息系统故障; 决策树; R 语言

中图分类号: TP319

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017058

Enterprise information system automatic fault diagnosis and analysis method based on decision tree

JIN Xin, YAN Longchuan, LIU Jun, ZHANG Shulin

State Grid Information & Telecommunication Branch, Beijing 100761, China

Abstract: With the rapid growth of the scale and complexity of enterprise information systems, traditional fault diagnosis and analysis methods relying on human experiences and manual operations cost more and more labor and time. To solve this problem, an automatic algorithm was proposed. The algorithm exploits information from system operation monitoring indicators and alarm data, based on decision tree, to automatically diagnose and analyze faults of enterprise information systems. The algorithm was verified, and the decision tree model and training method was simulated and analyzed comparatively under R language environment, using alarm data extracted from real operation data of a typical large-scale enterprise system. The experiment results show that this algorithm is able to achieve fast automatic fault diagnosis accurately, and is much helpful on improving efficiencies of information system fault processing.

Key words: automatic diagnosis analysis, fault of information system, decision tree, R language

1 引言

随着企业信息系统规模的日益庞大、集成程度的不断提高、技术架构的日益复杂、与企业经营管理和安全生产

的紧密结合、应用频度的大幅增加, 故障影响的范围广、涉及的用户多、带来的损失大, 且信息系统各组件存在耦合关系, 使故障具有传导特性, 当一个组件或某项指标出现故障时, 通常会引起其他组件或指标联动异常, 导致信息



系统故障点难于定位排查。在企业信息系统故障处置过程中,如何在复杂的各类监控告警和异常现象中,快速准确定位出故障点,实现信息系统故障快速处置和业务及时恢复,是所有大型企业需要迫切解决的问题。传统的信息系统故障分析处置主要依靠人工完成,处理时效和准确率受限于工程师的个人能力、经验、水平,差异较大、分布不均、效率低下,难以适应大部分复杂企业信息系统的实际情况。

决策树算法是一种通过对数据进行处理、利用归纳算法生成可读的规则和决策树,然后使用决策树对新数据进行分析的典型分类方法,因其具有分类精度高、生成模式简单、抗噪声数据等优点,在数据挖掘中受到广泛关注,在网络设备、电网、光纤通信网等故障诊断中具有广泛的应用。参考文献[1]设计了基于决策树的网络故障专家系统模型,参考文献[2]提出了基于决策树的变电站故障诊断知识的获取与表示,参考文献[3]设计了一种基于优化的决策树数据挖掘算法的光纤网络通信故障检测系统,参考文献[4]提出了采用故障诊断树设计ASON故障诊断系统推理机的方案。

本文将决策树算法理论与企业信息系统故障数据相结合,提出基于决策树的企业信息系统故障的自动诊断分析方法,以期发掘出信息系统指标间的关联规则,实现根据信息系统告警综合信息快速定位故障点,提高信息系统故障的处置效率。利用某大型国有企业实际生产运行中产生的信息系统故障及告警数据,提取典型告警特征和故障分类特征,采用决策树生成算法,在R语言环境下搭建仿真实验平台,验证该方法的实现效果和应用价值。

2 运行指标选择及故障类型

实际生产中,企业级大型信息系统需应对高并发访问、海量数据处理、高可靠性运行等一系列问题,还需实现高性能、高可用性、易伸缩、可扩展、安全等各种技术架构目标^[5]。信息系统一般包括应用服务器、数据库服务器、缓存服务器、负载均衡、集群、存储服务器、文件系统服务器、搜索引擎服务器等部分,典型信息系统架构如图1所示。

为便于对信息系统运行指标的研究、选取和对信息系统架构的分析,大型信息系统切分通常有两种方式:分层和分割。分层方式是将信息系统在横向维度切分,通过上层对下层的依赖和调用组成完整的系统,有利于信息系统分层次界面、清晰逻辑结构;分割方式是将信息系统在纵向方面切分,将不同功能和服务分割开,包装成高内聚、

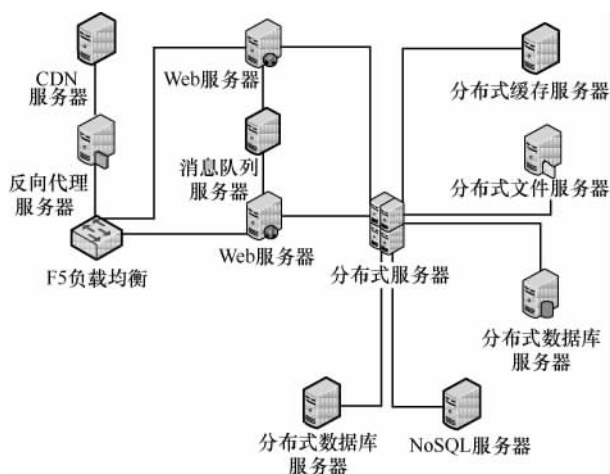


图1 典型信息系统架构

低耦合模块,有利于不同模块的分布式部署。本文立足于通用信息系统的故障自动诊断分析方法,尽量弱化不同信息系统间的差异,且考虑到信息系统故障时产生的逐级传导效应,故采用分层方式对典型信息系统结构进行简化处理。如上,可将信息系统分为业务层、应用服务层、数据层3个层次,典型信息系统分层的架构如图2所示。

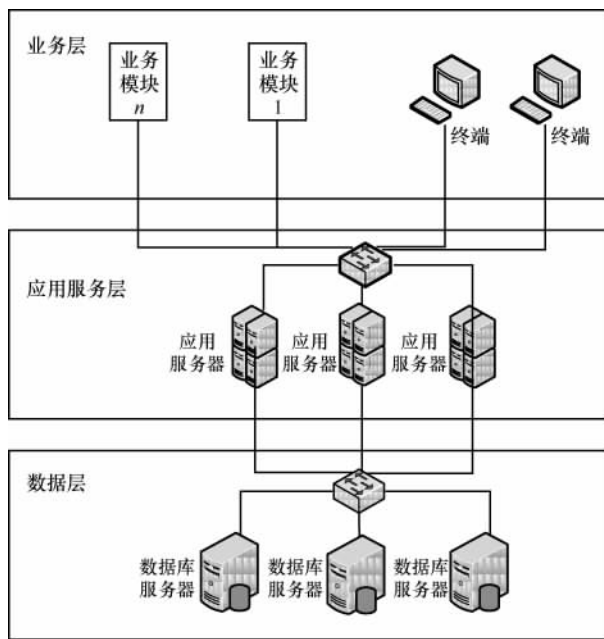


图2 典型信息系统分层的架构

通过对信息系统各层次进行细化,分析各层的主要构成组件和运行指标,并对生产运行中信息系统故障产生的主要告警数据以及定位到的信息系统故障点数据进行归纳,具体见表1。本文将从表1中择优选取运行指标和故障点类型。

表 1 信息系统各分层的运行指标及故障点

序号	层次	构成组件及运行状态指标	故障点
1	业务层	在线用户数和健康运行时长指标、URL 地址探测、系统可访问性、业务使用情况等	应用服务业务量、用户权限、监控工具运行等
2	应用服务层	应用节点连通性、应用节点可用性、端口、进程、线程、事件、内存、日志、CPU、TCP 连接、磁盘、文件使用率、会话连接、代码效率、数据库可连接性、JVM 内存、中间件线程等	应用服务 CPU、应用服务 TCP 连接、应用服务磁盘、应用服务进程、应用服务内存、应用服务线程、应用服务代码质量、应用服务电源、中间件数据库连接池、中间件内存、负载均衡配置、隔离装置运行等
3	数据层	数据库节点连通性、dbtime、I/O 性能、集群等待事件、行锁情况、监听服务、内存、会话数、swap 空间、日志空间、数据库表空间、归档空间、数据库密码、集群 RPC 通信时长、SQL 语句等	数据库 SQL 语句、数据库归档空间、数据库监听服务、数据库密码、数据库内存、数据库存储过程、数据库响应时长等

3 决策树算法

机器学习中,决策树是一个预测模型,是一种依托于分类、训练的预测树,并能根据已知进行预测和归类。决策树算法具有规则清晰、便于理解的特点。经典的决策树算法有 ID3(iterative dichotomiser 3,迭代二叉树 3 代)算法、C4.5 算法和 CART(classification and regression tree,分类回归树)算法。

ID3 算法的核心思想是自顶向下地贪婪搜索遍历可能的决策树空间构造决策树,以信息增益度量属性选择,选择分裂后信息增益最大的属性进行分裂,即递归地选择分类能力最强的特征对数据进行分割。

一个属性的信息增益就是由于使用这个属性分割样例而导致的期望熵降低,一个属性 A 相对样例集合 S 的信息增益 $Gain(S,A)$ =集合 S 的经验熵-属性 A 给定情况下 S 的经验条件熵,计算式可表示为:

$$Gain(S,A)=Entropy(S)-\sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

其中, $V(A)$ 是属性 A 所有可能值的集合, S 是样本集合, S_v 是集合 S 中属性 A 的值为 V 的子集,即 $Gain(S,A)$ 是由于给定属性 A 的值而得到的关于目标函数值的信息。

ID3 算法的目的在于减少树的深度,但是忽略了叶子数目的研究,C4.5 算法是对 ID3 算法的改进,在预测变量的缺值处理、剪枝技术、派生规则等方面做了较大改进。C4.5 算法采用增益比率(GainRatio)进行属性选择,增益比率度量是用增益度量 $Gain(S,A)$ 和分裂信息度量 $SplitInformation(S,A)$ 来共同定义的,可表示为:

$$GainRatio(S,A)=\frac{Gain(S,A)}{SplitInformation(S,A)} \quad (2)$$

其中,分裂信息度量被定义为(分裂信息用来衡量属

性分裂数据的广度和均匀性):

$$SplitInformation(S,A)=-\sum_{l=1}^c \frac{|S_l|}{|S|} \lg \frac{|S_l|}{|S|} \quad (3)$$

其中, S_1 到 S_c 是 c 个值的属性 A 分割集合 S 形成的 c 个样例子集,分裂信息实际上就是集合 S 关于属性 A 的各值的熵。

还可以使用 CART 算法生成决策树,CART 算法采用一种二分递归分割技术,将当前的样本集分成两个子样本集,使得生成的每个非叶子节点都有两个分支,生成的决策树是二叉树。CART 算法将每个属性的所有划分按照能减少的杂质量进行排序,杂质减少被定义为划分前的杂质量减去划分后每个节点的杂质量再乘以划分所占样本比率之和,使用杂质度量方法(基尼指数 $Gini$)选择最优特征,一个节点 B 的 $Gini$ 不纯度定义为:

$$Gini(B)=1-\sum_{k=1}^c p_k^2 \quad (4)$$

其中, k 是类, c 是类别集中因变量数目, p_k 是观测点中属于 k 类的概率。

决策树算法构造决策树来发现数据中隐藏的分类规则,决策树构造的输入是一组带有类别标记的例子,构造的结果是一棵二叉树或多叉树,构造精度高、规模小的决策树是决策树算法的核心内容。决策树构造可以分两步进行:决策树的生成和决策树的剪枝。本文研究使用 C4.5 算法和 CART 算法。

4 实验及应用

本文收集某大型国有企业实际生产运行中关于应用系统或数据库方面的信息系统故障分析报告,从中采集故障告警数据及故障点,对数据进行梳理、清洗,选取典型的信息系统故障运行指标告警特征和故障点特征,分别采用 C4.5 算法和 CART 算法生成决策树,使用 R 语言程序包进



行仿真实验,利用70%的故障数据完成了对决策树的训练,利用30%的故障数据完成了故障自动诊断的测试。

经研究,选取的故障运行指标告警共18项,包括:在线用户数和健康运行时长指标中断告警、页面探测异常告警、系统无法正常使用告警、应用节点日志告警、线程超时或阻塞告警、业务量大告警、应用服务内存告警、应用服务CPU告警、应用无法连接数据库告警、数据库ping不通告警、数据响应慢告警、数据负载高告警、数据库内存告警、数据库监听服务告警、数据库归档空间告警、数据库密码不可用告警、数据库SQL语句效率告警;选取的信息系统故障点数据共12项,包括:数据库归档空间问题、数据库监听服务异常问题、数据库密码失效问题、数据库响应问题、数据库

SQL语句不合理问题、数据库服务器内存问题、中间件无法连接问题、中间件内存问题、应用服务线程问题、应用服务业务量大问题、应用服务器内存问题、应用服务器CPU问题。

采用R语言环境下的rpart分组实现,利用sample对数据源进行抽样分割,形成训练集数据和测试集数据,利用rpart.control设置生成树的交叉验证折数、生成树的深度、最小分支节点数、复杂度等生成树参数,利用rpart进行因变量和自变量的特征选取、算法选取以及生成树生成,利用prune对生成树进行剪枝修正,利用rpart.plot描绘生成树,利用predict进行预测,利用sum、table等公式计算预测正确率。利用C4.5算法和CART算法生成的决策树分别如图3和图4所示。

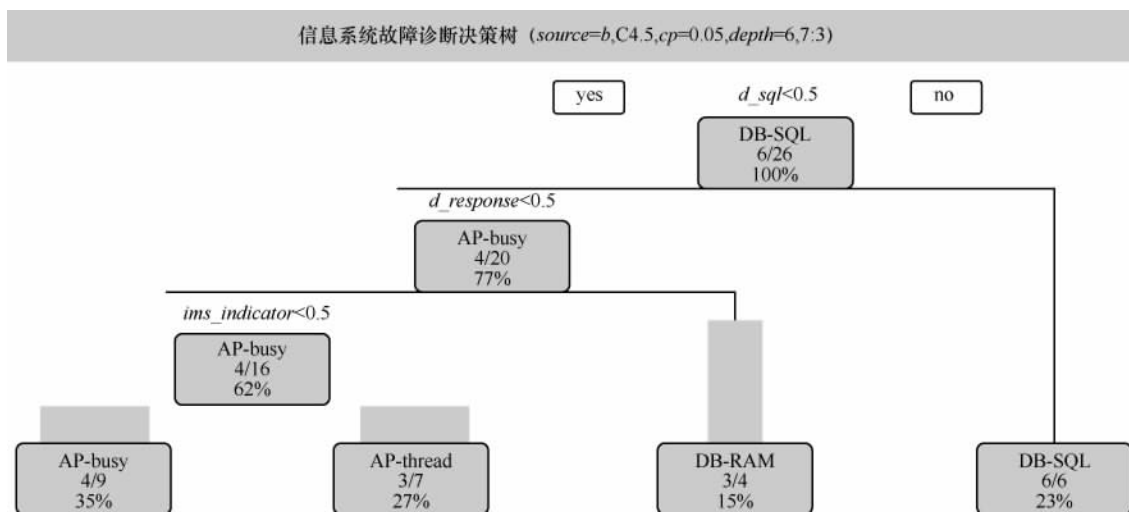


图3 C4.5算法生成的决策树

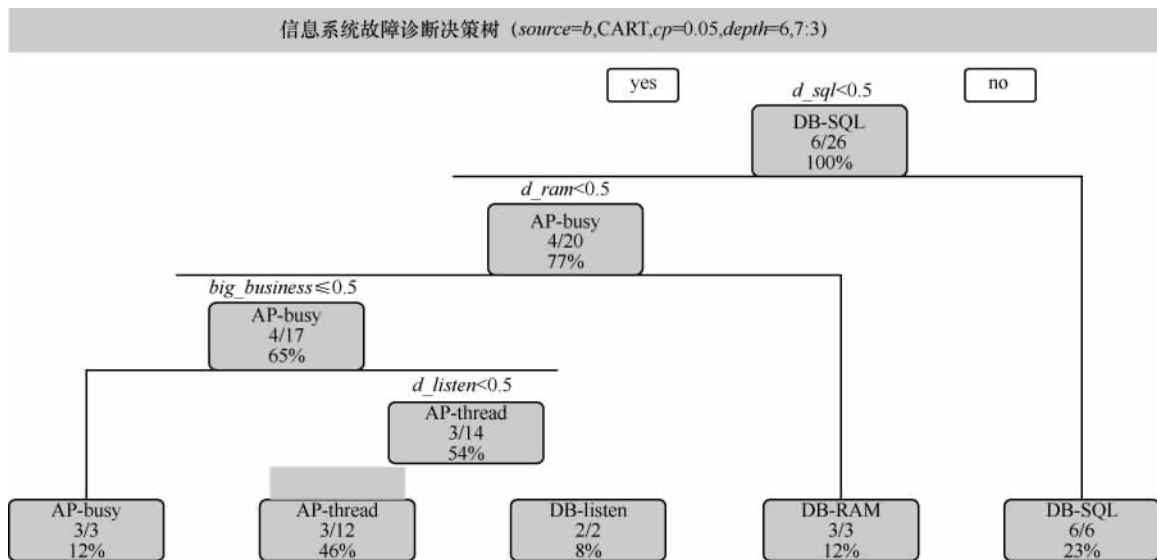


图4 CART算法生成的决策树

由图 3 可知,C4.5 算法生成的决策树规则认为,在所有输入告警中,分类效果最好的告警依次是:数据库 SQL 语句告警、数据库响应慢告警、在线用户数和健康运行时长指标中断告警。在故障自动诊断过程中,当用户将信息系统各类告警信息输入后,决策树将依据生成规则,给出故障点预测结果。

由图 4 可知,CART 算法生成的决策树规则认为,在所有输入告警中,分类效果最好的告警依次是:数据库 SQL 语句告警、数据库服务器内存告警情况、应用服务业务使用量情况、数据库监听服务运行情况。在故障自动诊断过程中,当用户将信息系统告警输入后,决策树将依据生成规则,给出故障点预测结果。

对两种算法的预测正确率进行对比,对比结果见表 2。由表 2 可知,两种算法的预测正确率随复杂度 cp 存在相同的变化规律,即 cp 越低,预测正确率越高。在 $cp=0.001$ 的场景下,C4.5 算法的预测正确率相较 CRAT 算法略高,更宜于应用。

表 2 不同算法的训练数据预测正确率对比

复杂度	C4.5 算法正确率	CART 算法正确率
$cp=0.1$	61.5%	50%
$cp=0.05$	61.5%	65.4%
$cp=0.001$	84.6%	76.9%

此外,数据源质量和数量对预测正确率影响很大,准确可靠、丰富全面的数据源对于训练出符合规则的、正确的、智慧的决策树至关重要,正确的决策树才能做出合理的预测,给出故障点诊断结果。

5 结束语

本文创新性地将决策树算法引入信息系统故障自动诊断,能够实现良好的故障自动诊断分析效果。实验证明,基于该方法可以实现根据信息系统运行指标告警数据进行故障自动诊断分析,还可以在便携式计算设备上实现秒级决策,有效节约人力成本,大幅缩减故障定位时间,对信息调度监控故障处置辅助诊断和决策具有广阔的应用前景。

在后续的工作中,将继续提升信息系统故障告警数据的数量和质量,优化告警特征选择和决策树算法,提高信息系统故障自动诊断的准确性。并引入故障历史处置操作数据,研究故障处置操作自动推荐算法,实现企业信息系统故障诊断分析与处置的一体化处理,推进此方法在企业

信息系统故障处置中的实用化水平和自动化程度。

参考文献:

[1] 曲朝阳,高宇峰,聂欣.基于决策树的网络故障诊断专家系统模型[J].计算机工程,2008,34(22):215-217.
QU Z Y, GAO Y F, NIE X. Network fault diagnosis expert system model based on decision tree [J]. Computer Engineering, 2008, 34(22): 215-217.

[2] 白建社,樊波,黄文华,等.基于决策树的变电站故障诊断知识表示与获取[J].电力系统及其自动化学报,2004,16(2):5-8.
BAI J S, FAN B, HUANG W H, et al. Knowledge representation and acquisition based on decision tree for substation fault diagnosis[J]. Proceedings of the EPSA, 2004, 16(2): 5-8.

[3] 黄勇,胡连城,刘增良.基于决策树算法的工业光纤网络通信故障检测系统[J].现代电子技术,2014(20):134-136.
HUANG Y, HU L C, LIU Z L. Industrial fiber network communication fault detection system based on decision tree algorithm[J]. Modern Electronics Technique, 2014(20): 134-136.

[4] 车雅良,林初善,潘青,等.ASON 网络故障诊断系统推理机的设计与研究[J].广东通信技术,2015,35(3):64-67.
CHE Y L, LIN C S, PAN Q, et al. Design and research of ASON network fault diagnosis system inference engine [J]. Guangdong Communication Technology, 2015, 35(3): 64-67.

[5] 李智慧.大型网站技术架构:核心原理与案例分析[M].北京:电子工业出版社,2013.
LI Z H. Large website technology architecture: core principles and case studies[M]. Beijing: Publishing House of Electronics Industry, 2013.

[作者简介]



金鑫(1985-),女,国家电网公司信息通信分公司工程师,主要从事信息运维工作。

闫龙川(1979-),男,国家电网公司信息通信分公司高级工程师,主要从事信息运维工作。

刘军(1970-),男,国家电网公司信息通信分公司高级工程师,主要从事信息通信运维及管理工作。

张书林(1968-),男,国家电网公司信息通信分公司工程师,主要从事信息通信运维及管理工作。