基于关联规则的态势预测方法

陈 珍，夏靖波，陈 婉，李 博
（空军工程大学信息与导航学院，西安，710077）

摘要 态势预测是网络态势感知的重要环节，可以为网络管理员提供必要的决策支撑。为了实现对网络的大数据管理模式，针对当前预测算法无法充分利用大数据优势的局限，提出了基于关联规则的态势预测方法。该方法综合考虑了大数据的特点和态势预测的需求，给出了方法的基本思想和实现流程。实验结果表明，提出的方法与传统预测方法相比，通过寻找数据间的关联物而不是非线性匹配来达到预测的目的，大大降低了计算的时间复杂度，提高了预测效率。

关键词 态势预测；关联规则；大数据
DOI 10.3969/j.issn.1009-3516.2016.04.016
中图分类号 TP393 **文献标志码** A **文章编号** 1009-3516(2016)04-0085-05

A Situation Forecast Method Based on Association Rules

CHEN Zhen, XIA Jingbo, CHEN Wan, LI Bo
(Information and Navigation College, Air Force Engineering University, Xi′an 710077, China)

Abstract: Situation forecast is a key link in network situation awareness, because this can provide necessary decision support for network managers. In order to realize the big data model of network management, a forecast method based on association rules is proposed aimed at the problem that the current forecast methods fail to take advantages of big data. The method takes the characteristics of big data into full account combined with requirement of situation forecast, and the basic idea and processes of the method are given. The experiment results show that the proposed method reduces time complexity greatly through finding relevance instead of nonlinear registration, and improves accuracy of forecast compared with the traditional methods.

Key words: situation forecast; association rules; big data

近年来，云计算技术迅速发展，移动互联网、物联网应用大规模爆发，那些由社交媒体、视频、音频、邮件、文档信息和网页所产生的海量数据正在以惊人的速度增长，人类进入了大数据时代。大数据时代的到来不只是带来了海量丰富的数据信息，也逐渐开启了人们生活、工作与思维上的变革[1]。

网络态势感知是网络管理的未来发展的必然方向[2]，包括觉察、理解（评估）、预测及可视化4个环节。其中，态势预测网络态势感知的重要环节，是指在态势评估的基础上，对网络在未来一段时间内整体的发展趋势做出预测，从而为网络管理者提供有力的决策支撑。常用的预测方法包括回归分析、贝

收稿日期：2015-09-28
基金项目：陕西省科技计划自然基金(2012JZ8005)
作者简介：陈 珍(1990—)，女，内蒙古巴彦淖尔人，硕士生，主要从事网络态势感知、大数据与云计算研究.E-mail:cz10184527@163.com

引用格式：*陈珍，夏靖波，陈婉，等. 基于关联规则的态势预测方法[J]. 空军工程大学学报：自然科学版，2016，17(3)：85-89. CHEN Zhen, XIA Jingbo, CHEN Wan, et al. A Situation Forecast Method Based on Association Rules[J]. Journal of Air Force Engineering University: Natural Science Edition, 2016, 17(3): 85-89.*

叶斯网络、马尔科夫链和人工神经网络等,这些方法多是通过复杂的运算进行数值匹配,更多注重的是预测的精确性,且针对的是小样本数据,当面对海量数据时,时效性就无法得到保证。精确的预测必然会以时间消耗为代价,在小数据时代,为了避免因抽样带来偏差放大而不得已追求精确性;而大数据时代,云计算等技术的发展,使得处理大规模数据成为可能,快速获得一个事务大概的轮廓和发展脉络远比严格的精确性重要且容易的多[3]。大数据关注的是数据的相关性或关联性,这也是大数据预测的关键。因此在大数据环境下,对网络态势做出预测,只需要从网络历史数据中找到是什么造成网络现在的状态,而没必要求解为什么。应用相关关系,可以使分析预测事务比以前更容易、更迅速、更清楚。

基于以上分析,针对现有预测方法的局限性,结合大数据预测的特征[3],本文提出了基于关联规则的网络态势预测方法。关联规则是数据挖掘领域中一个非常重要的课题,相关文献已有学者将其应用到预测领域,文献[4]将关联规则算法用于电信网络告警,文献[5]研究了关联规则在干旱预测中的应用,文献[6]建立了基于关联规则的城市电力负荷预测模型,文献[7]则运用关联规则进行股票预测。与常用的预测方法相比关联更侧重于挖掘数据之间有价值的关联知识,通过寻找"关联物",而不是复杂的非线性匹配,达到预测的目的。

1 基本概念

关联规则[8]是指从有噪声的、模糊的、随机的海量数据中,挖掘出隐藏的、事先不知道的、但是有潜在关联的信息或知识的过程,所发现的信息和知识通常用关联规则或频繁项集的形式表示[9]。下面介绍关联规则的形式化描述及相关定义。

设 $I=\{i_1,i_2,\ldots,i_m\}$ 是项的集合,其中 $i_k\{k=1,2,\ldots,m\}$ 表示项,如果 $X\subset I$,集合 X 被称为项集,如果一个项集包含 K 个项,则称为 k-项集。事务二元组 $T=(T_{\mathrm{ID}},X)$,T_{ID} 是事务唯一的标识符,称为事务号,数据集 $D=(t_1,t_2,\ldots,t_n)$ 是由 t_1,$t_2,\ldots,t_n$ 事务组成的集合。如果项集 X 是事务 t_j 的子集,则称事务 t_j 包含项集 X。项集的一个重要性质是它的支持度计数,即包含特定项集的事务个数,项集的支持度计数 $\delta(X)$ 可以表示为:

$$\delta(X)=|\{t_i\mid X\leqslant t_i,t_i\subset T\}| \tag{1}$$

关联规则可以描述为 $A\Rightarrow B$ 的蕴含式,其中,$A\subset I$,$B\subset I$,且 $A\cap B=\varnothing$。关联规则的强度可以用它的支持度(Support)和置信度(Confidence)度量,支持度表示 D 中事务包含 $A\cup B$ 的百分比,它是概率 $p(A\cup B)$,而置信度表示包含 A 的事务也包含 B 的百分比,它是条件概率 $p(B/A)$。支持度和置信度这 22 种度量的形式定义如下:

$$s(A\Rightarrow B)=p(A\cup B)=\frac{\delta(A\cup B)}{\delta(D)} \tag{2}$$

$$c(A\Rightarrow B)=p(B/A)=\frac{\delta(A\cup B)}{\delta(A)} \tag{3}$$

关联规则的任务就是在事务 D 中找出大于用户给定的最小支持度和最小置信度的规则。关联规则的挖掘问题一般可以分解为以下 2 个子问题:①发现频繁项集:发现满足最小支持度阈值的所有项集,这些项集称为频繁项集;②产生关联规则:从问题①发现的频繁项集中找出所有满足要求的置信度的规则,这些规则称为强规则。

2 关联规则用于态势预测

虽然态势预测方法在人工智能领域已经得到充分研究,但基于关联规则的态势预测方法研究并不多见。由于关联规则善于发现态势与指标之间隐含的关联关系,通过优先选择强规则用于决策支撑,有可能最大程度的减少损失。

2.1 算法设计

关联规则应用于态势预测的基本思想是:将势指标体系中的每一类指标作为规则前件的属性,网络状态作为规则后件的属性,每一类的指标属性是一个事务 t,所有的网络状态集合是一个事物 t,收集的所有类型的 t 组成事务集合 D,D 由不同类型的表构成,其中行集 T 表示 t 的类型,如脆弱性、容灾性、威胁性、稳定性、网络状态等;列集为特征项的集合,即每一类中具体包含哪些内容,如网络状态这一事物 t 包含的特征项有优秀、良好、正常、预警、瘫痪等。算法的目的就是通过分析海量历史数据找出指标与网络态势之间相互影响关系,即关联规则,这样在发现一些指标异常的同时就可以预测接下来的网络状态的发展趋势。由于影响网络态势的因素众多,等症状全部发生就可能已经到了无法弥补的境况,如果能在发现个别重要症状的同时及时采取相应的措施,就可以最大程度的避免灾难,尽可能的减少损失。

算法的具体步骤如下:

Step1 收集历史监测数据,作为训练样本,并对各指标值进行相应的预处理。

Step2 采用了类似的单词计数的过程对训练集进行扫描,找出满足最小支持度的 k-频繁项集(k

$=1,2,\cdots,m$），并保存。其中最小支持度阈值 min_sup 有用户自己定。

Step3　设定最小置信度阈值 min_conf，产生指标间的强关联规则集。

Step4　将强关联规则集中的规则按照长度最短、置信度最大、支持度最大进行多关键字排序。这样有利于快速定位接下来异常发生点。

Step5　运用大量的历史实例数据对产生的强关联规则集进行验证，去掉具有偶然性的巧合关联，尽可能的精简强关联规则集。

算法的伪代码描述如下：

输入：训练集 D，最小支持度阈值 min_sup 最小置信度阈值 min_conf

输出：关联规则集 rules

initial $k=0$；　// k 表示扫描次数

$L=\varnothing$；$C_1=I$；　//L、C 分别表示频繁项集、候选项集的集合，初始把所有的单个项目作为候选项集

rules=$\varnothing$；

Repeat

$k=k+1$；

$L_k=\varnothing$；

for $I_i \in C_k$ do

$c_i=0$；//每个项目集的初始计数设为 0

for $t_j \in D$ do

　for $I_i \in C_k$ do

if $I_i \in t_j$ then

　$c_i=c_i+1$；

　　if $c_i >=$ min_sup then

　　$L_k=L_k \cup I_i$；

　　$L=L \cup L_k$；

Until $L_{k+1}=\varnothing$；

CF $=L_j/L_i (i<j)$；//求各规则置信度

if$\cup$CF $>=$ min_conf then

　$R=$ getrules()；

rules $=$rules $\cup R$；

Output rules；

为提高算法的效率，Apriori 算法运用了“频繁项集的子集都是频繁项集，非频繁项集的超集都是非频繁项集”这一重要性质有效地对频繁项集进行了修剪，大大减少了计算候选项集支持度及规则置信度的计算量。

2.2　基于关联规则的态势预测流程

基于关联规则的网络态势预测流程包括数据准备、建立预测规则挖掘模型和预测关联规则检验 3 个步骤，流程见图 1。

数据准备的主要任务是对收集的原始数据进行必要的预处理。原始指标之间由于类型不同、量纲不同而存在着不可公度性，如果直接用来作为态势因子，就可能造成不合理的现象发生。因此为了确保计算的合理性，尽可能的反映实际情况，必须对指标作一定的预处理。此外，关联规则通常适用于指标取离散值的情况[10]，对于一些连续型指标必须进行离散化处理才能用于规则挖掘。

建立预测规则挖掘模型就是根据算法设计生成强关联规则集，这是应用关联规则进行态势预测的核心部分。由于影响网络态势的因素有很多，有用户层面的，也有网络及网络设备层面的，因此接下来需要做的就是对所采集到的数据进行一些必要的数理统计，例如某些指标取值在什么范围对应网络的什么状态，从而构造出一些与网络状态密切相关的属性。各指标与态势间的关系是隐含且复杂的，采用此模型可以找出它们之间的联系，有助于态势预测和决策制定。

预测关联规则检验就是用实际的历史实例去验证生成的关联规则，看规则集是否有重复、巧合或是遗漏，根据检验结果调整 min_sup 和 min_conf，完善规则集。如果调整阈值后仍有未匹配的规则，那么直接将此规则加入规则集。

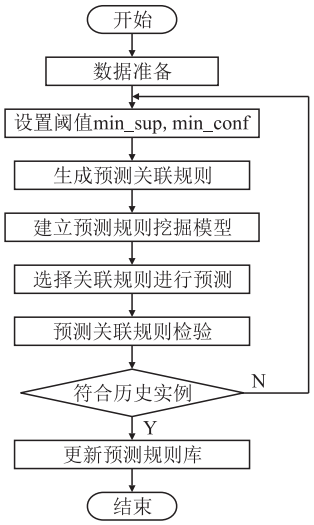

图 1　基于关联规则的态势预测流程图

Fig.1　Flow chart based on association rules of awareness prediction

3　仿真实验

3.1　实验数据与内容

为了证明本文提出算法的有效性，实验对算法的参数设置及其性能进行了进一步的探索研究。实验数据通过模拟网络平台获得，利用 NS2 仿真软件，以美国教育科研网 Abilene 骨干网络拓扑为例，搭建了网络模拟平台，见图 2。为了更逼真地模拟现实网络，模拟过程中两两节点间选择以随机的方式进行相互通信。

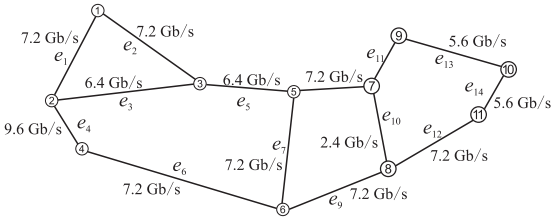

图 2 Abilene 骨干网络拓扑

Fig.2 Abilene backbone network topology

实验内容包括:

1)通过准确率的变化设置合适的 min_sup、min_conf 参数值,以探索该算法中参数设置对预测准确率的影响。

2)首先随机选取不同时间段,采集规模相当的不同数据集,将本文方法与常用的几种态势预测方法进行准确率对比;然后采集不同规模的数据集,通过比较预测时间进一步说明该方法的有效性。

3.2 实验结果与分析

3.2.1 参数选取

关联规则用于态势预测关键是要设置最小支持度 min_sup 和最小置信度 min_conf 2 个参数,而参数对算法准确率的影响是值得探索的问题。图 3 反映了参数对预测准确率的影响。

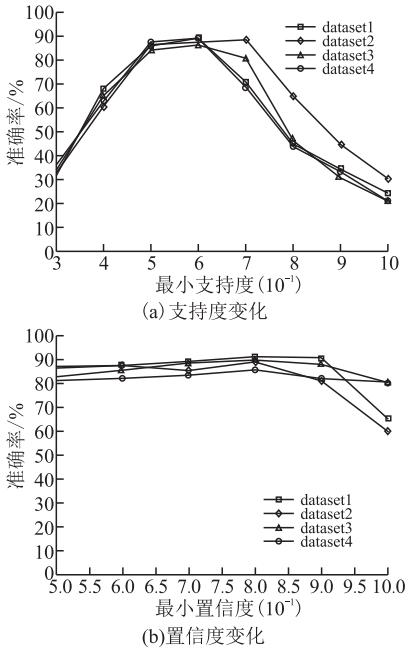

图 3 参数对准确率的影响

Fig.3 The influence of parameters on the accuracy

由图 3(a)可以看出,随着最小支持度逐渐变大,算法的准确率先是迅速上升,增加到 0.5 时逐渐变缓,增加到 0.6 左右时,准确率开始明显下降。图 3(b)中因为较小的置信度对预测并没有太大的研究意义,所以直接从 0.5 开始取值,可以发现随着置信度的变大,预测准确率在刚开始时并没有较大的起伏,当增加至 0.85 左右时,准确率开始下降,故本文取 min_sup=0.6,min_conf=0.85。从理论层面分析虽然支持度和置信度越大,规则的可信度就越高,但随着参数的增大,规则数目也在迅速减少,这就意味着有价值的规则数目也在减少,过少的规则数目会带来漏测问题,反而会降低预测效率。因此,参数的设置对预测准确率有直接影响,高的可信度并不能导致高的预测精度,充足的规则数才是基于关联规则的态势预测算法高效的必要条件[15-16]。

3.2.2 性能比较

为了进一步证明本文提出方法的有效性,实验将常用的几种态势预测方法与该方法在随机采集的 10 个数据集上进行了准确率的比较。其中 ARAM[12]是研究时间序列预测的重要方法,在短期内具有较高的预测准确度;GRNN[13]是神经网络预测模型,属于浅层学习结构;SVR[14]是支持向量机在回归学习中的应用,在小样本条件下具有较好的预测性能。实验结果见图 4。

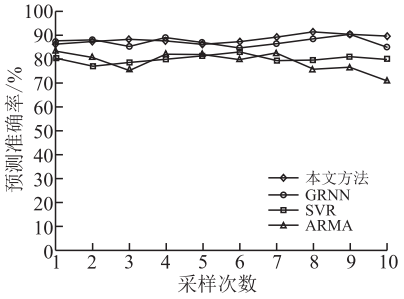

图 4 预测准确率比较

Fig.4 Comparison of prediction accuracy

从图 4 中可以看出,ARMA 算法准确率波动较大,稳定性欠缺;SVR 模型虽然相对稳定,但总体准确率有待提高;本文算法和 GRNN 算法稳定性较好,且都表现出了较高的准确率。从算法的复杂度来说,本文算法原理[17],不涉及复杂的非线性运算,ARMA 和 GRNN 则是通过复杂的数值匹配进行预测,数据规模增加时,精确性必然会以时间为代价。为进一步说明本文算法的有效性,实验又在不同规模的数据集上对几种算法的时间性能进行了对比,结果见图 5。

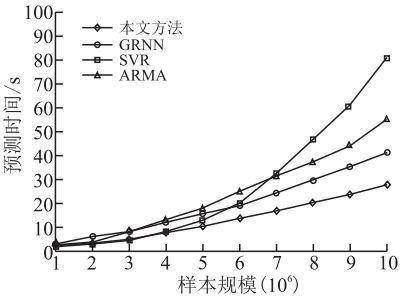

图 5 数据规模对预测时间的影响

Fig.5 The influence of data scales on time

可以看到随着数据规模的增加，SVR 的预测时间成指数型增加；ARMA 和 GRNN 的增长介于指数和线性之间，本文算法则更接近线性增长，在时间效率上均优于其他算法。而且本文算法容易并行化，可以通过 MapReduce 计算模型进一步提升其时间效率[18]可扩展性。

4　结语

本文针对现有预测方法无法充分利用大数据优势的不足，结合大数据应用的特点及现状，提出了基于关联规则的态势预测的方法。给出了方法的基本思想和实现流程，并通过实验探索了相关参数设定与预测准确率之间的影响关系，最后验证了该方法的优越性。与传统方法相比，该方法放弃了复杂的非线性计算，更加重视数据间隐藏的、潜在的关系，比较符合大数据用于预测和决策的技术发展趋势。但是，该方法对数据规模比较敏感，进一步提高其对大规模数据的预测效率是下一步的研究重点。

参考文献(References)：

[1] Viktor Mayer－SchÖnberger, Kenneth Cukier. Big Data[M]. Eamon Dolan/Houghton Mifflin Harcourt,2013.

[2] 柏骏，夏靖波，赵琪，等. 栅格化信息网络态势感知能力建设及其模型[J]. 空军工程大学学报：自然科学版，2014，15(2)：48-50.
BAI Jun，XIA Jingbo，ZHAO Qi，et al. Model and Construction of Situation Awareness Ability for Grid Network[J] Journal of Air Force Engineering University ：Natural Science Edition，2014，15(2)：48-50.(in Chinese)

[3] 吕本富，陈健. 大数据预测研究及相关问题[J]. 科学促进发展，2014，10(1)：60-63.
LÜ Benfu，CHEN Jian. The Predication Based on Big Data and Related Issues[J]. Science & Technology for Development，2014，10(1)：60-63.(in Chinese)

[4] 于漫，胡明，金刚，等. 关联规则算法的电信网络告警作用[J].吉林大学学报，2010，28 (3)：264-269.
YU Man，HU Ming，JIN Gang，et al. Association Rules Algorithm Applied to Telecommunications Network Alarms[J]. Journal of Jilin University：Information Science Edition，2010，28(3)：264-269.(in Chinese)

[5] 肖峻，耿芳，杜柏均. 基于关联规则的城市电力负荷预测模型智能推荐[J]. 天津大学学报，2010，43 (12) ：1079-1085.
XIAO Jun，GENG Fang，DU Bojun. Intelligent Recommendation of Urban Power Load Forecasting Models Based on Association Rules [J]. Journal of Tianjin University，2010，43(12)：1079-1085.(in Chinese)

[6] 徐海鹏. 基于关联规则的股票预测方法[J]. 计算机与数字工程，2010，38 (3) ：150-153.
XU Haipeng. Forecasting Method of Stock Tend Based on Association Rules[J]. Computer and Digital Engineering，2010，38(3)：150-153.(in Chinese)

[7] 王红霞，李松. 关联规则挖掘在干旱预测中的研究与应用[J]. 微计算机信息，2010，26(4) ：264-269.
WANG Hongxia，LI Song. Research and Application about Association-Rules Algorithm in Drought Forecast System[J]. Control & Automation，2010，26(4)：264-269.

[8] AGRAMAL R，IMICLINSKI T，SWAM I. Data Mining：a Performance Performance[J]. IEEE Trans Knowledge and Data Engineering，1993，5：9142925.

[9] 陈工孟，须成忠.大数据导论[M]. 北京：清华大学出版社，2015.
CHEN Gongmeng，XU Chengzhong. Introduction to Big Data[M].Beijing：Tsinghua University Press，2015.(in Chinese)

[10] 周宝曜，刘伟，范承工. 大数据：战略·技术·实践[M].北京：电子工业出版社，2013.
ZHOU Baoyao，LIU Wei，FAN Chenggong. Big Data Strategy，Technology，Applications[M]. Beijing：Electronic Industry Press，2013.(in Chinese)

[11] 柏骏，夏靖波，钟赟，等. 网络运行态势感知技术及其模型[J]. 解放军理工大学学报：自然科学版，2015，16(1)：16-21.
BAI Jun，XIA Jingbo，ZHONG Yun，et al. Network running situation awareness technology and its model [J]. Journal of PLA University of Science and Technology ：Natural Science Edition，2016，16(1)：16-21.(in Chinese)

[12] 高波，张钦宇，梁永生，等. 基于 EMD 和 ARMA 的自相似网络流量预测[J]. 通信学报，2014，32(4)：47-56.
GAO Bo，ZHANG Qinyu ，LIANG Yongsheng，et al. Predicting Self-Similar Networking Traffic Based on EMD and ARMA[J]. Journal on Communications，2014，32(4)：47-56.(in Chinese)

[13] 卓莹，张强，龚正虎. 网络态势预测的广义回归神经网络模型[J]. 解放军理工大学学报：自然科学版，2012，13(2)：147-151.
ZHUO Ying，ZHANG Qiang，GONG Zhenghu. GRNN Model of Network Situation Forecast[J]. Journal of PLA University of Science and Technology：Natural Science Edition，2012，13(2)：147-151.(in Chinese)

[14] 钱叶魁，陈鸣. MOADA-SVR：一种基于支持向量机回归的多元在线异常检测方法[J]. 通信学报，2011，32(2)：106-112.
QIAN Yekui，CHEN Ming. MOADA-SVR：A Multivariate Online Anomaly Detection Algorithm Based on SVR[J]. Journal on Communications，2011，32(2)：106-112.(in Chinese)

[15] WANG Weiling　CHU Jianchong　XU Like. A Feature Selection Algorithm Based on Correlation [J]. Computer Applications and Software，2009，26(8)：259-261.

[16] WU Jianhua，SONG Qinbao，SHEN Junyi，et al. A Feature Selection Algorithm Based on Association Rules[J]. PR&AI，2009，22(2)：

[17] 陈莉，焦李成. 基于关系代数的关联规则挖掘算法[J]. 西北大学学报：自然科学版，2005，35(6)：692-694.
CHEN Li，JIAO Licheng. Association Rule Mining Algorithm Based on Relation Algebra Theory[J]. Journal of Northwest University：Natural Science Edition，2005，35(6)：692-694.(in Chinese)

[18] 李玲娟，张敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展，2011，21(2)：43-50.
LI Lingjuan，ZHANG Min. Research on Algorithms of Mining Association Rule under Cloud Computing Environment[J]. Computer Technology and Development，2011，21(2)：43-50.(in Chinese)

（编辑：徐楠楠）