

物联网数据特性 对建模和挖掘的挑战

关键词：物联网 数据属性 广义超图 图分解
子图挖掘 大规模可外推非参数模型

在本刊2010年第7期中,《海计算:物联网的新型计算模型》(作者:孙凝晖、徐志伟、李国杰)一文对面向物联网计算的概念和体系结构做了比较高层的畅想。海计算数据的特点是:(1)物物相联,相联关系非单一;(2)场景未知;(3)数据不完整;(4)信息与物理世界相连。这些特性对数据分析与建模有独特的要求。本文将从另一个层面,讨论从物联网的万千终端收集到数据后,对它们的管理、分析和使用所面临的一些基本挑战。首先讨论海计算环境中关于数据的特性、数据挖掘与分析的要求、数据模型及其基本操作等与数据相关的一些基本问题,然后探讨为之服务的体系结构和开发环境。

无论物联网在具体应用中有何种不同的表现形式,物联网数据挖掘分析应用通常都可以归纳为以下两大类:

预测 (Forecasting) 主要用于在(完全或部分)了解现状的情况下,推测系统在近期或者中远期的状态。例如,(1)在智能电网中,预测近期扰动的可能性和发生的地点;(2)在智能交通系统中,预测拥堵和事故在特定时间和地点可能发生的概率;(3)在环保体系中,根据不同地点的废物排放,预测将来发生生物化学反应产生污染的可能性。

寻证分析 (Provenance Analysis) 当系统出现问題或者达不到预期效果时,分析它在运行过程

中哪个环节出现了问題。例如,(1)在食品安全应用中,一旦发生质量问题,需要在食品供应链中寻找相应证据,明确原因和责任;(2)在环境监控中,当污染物水平超标时,需要在记录中寻找分析原因。

显然对于数据挖掘,预测和寻证分析都不是新问題。但在物联网环境中,由于数据的特性不同(特别是复杂的物物关联),导致建模方式和传统方式有很大差异。因此,为了提出合适的解决方案和正确思路,我们先分析物联网的数据特性,然后讨论合适的数学模型。简单来讲,与传统数据挖掘领域的情形相比,物联网数据的主要特点是:时空性、关联性、质量差、海量和非结构性。

物联网数据的特性和对建模的要求

在物联网应用中,原始数据通常是从一个时空网络的四维空间中收集上来的,图1是抽象示意图。每个点代表物联网中的一个个体,每条边代表物物相联关系。在某些应用中,相联关系是有向非对称的。随着时间的推移,物的数量和物物之间的关系也会发生变化。

数据的空间时效性

如图1所示,空间时效性是物联网数据的必要

范伟¹ 李晓明²

¹IBM美国华讯研究院
和IBM中国研究院

²北京大学

属性。所有原始数据在缺省状态下都具有时间、空间和设备戳（default time, space and device stamp），即表示在特定时间、地点在特定设备上收集的。例如，（1）在智能电网应用中，相位测量单元（Phase Measurement Unit, PMU）记录了特定输电线路在特定时间点的测量信息；（2）在智能交通系统中，车载GPS测量记录车辆记录了在特定时间的位置信息；（3）在食品安全应用中，每个数据包都包含在特定时间和地点的加工和处理信息；（4）在环境监控中，每个传感器都记录了在特定位置和时间的污染源测量数值。

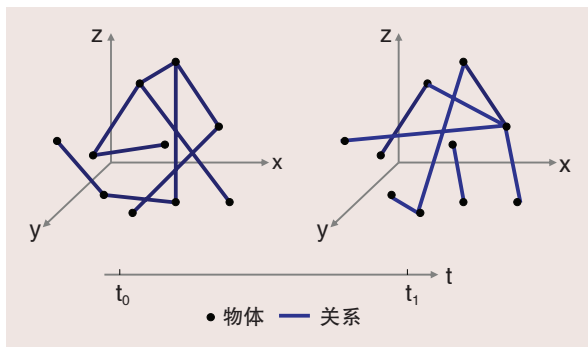


图1 时空相关的物联网

总之，在物联网应用中，时空设备信息是数据的固有属性。在数据的传输中，时空信息应当保留，不应当转换或者缺失。为了保证数据的完整性（data integrity）和安全性，应当研究有效的方案保证数据的时空完整性（例如，用附加码和共用算法的方式）。在为了减少带宽的操作中，应当用加入关键码的方式使我们能够从外存中找回这些基本信息。当数据在不同设备和应用中传输时，应当保证这些基本属性的完整正确性。在数据收集过程中，应当保证检测时间空间设备戳的完备性，防止张冠李戴和时序错乱，要确保数据的物理几何含义的正确。

数据的连接关系

在物联网应用中，各个物理对象不是独立存在的，它们之间存在复杂多样的关联。其中一些关联是直接的，一些是间接隐含的。在现有技术中，

通常的做法是忽略事物之间的关系，或者只是描述一种关系（例如，社会网络中的友谊关系等）。例如，（1）在智能电网中，不同用电户在物理电网上的相对位置会影响他们之间的关系和关联程度；（2）在食品安全应用中，不同供应方和使用方错综复杂的供求关系，会直接或间接地影响产品加工运输中的实施细节；（3）在环境监测中，不同污染源的相对位置和相对独立性，都会对监测系统的设计与实现带来影响。

在物联网应用建模时，应当充分考虑并表达物理个体之间的关系特别是直接的关系，间接的关系可以通过模型的办法（例如SVD、拉普拉斯变换等）推导出来。各个物理个体除以上论述的实时收集的时空数据之外，也应充分表达它们之间的连结关系。在一些应用中，这些连结关系也会随着时空的转换而发生变化（例如，智能交通中车辆之间的关系）。模型本身应有充分的能力来表达直接关系，以方便推理间接关系。

数据的质量问题

由于传输错误、传感器失效、停电等原因，物联网会发生数据成批成片或者部分丢失和错误。数据出错与丢失的原因可能是随机的也可能是系统的。例如，（1）在智能电网应用中，由于传感器电池没电了，可能在一段时间里完全得不到某些输电线路上的相位测量信息。在广域网传输中，由于设备出现故障，也会造成数据暂时丢失或者读写出现错误。（2）在车辆智能交通应用中，由于干扰、信号强度不够，有些GPS信息无法传到基站，造成随机或系统性数据丢失。有时也会出现“时序错乱”现象。

因此，在物联网应用中，数据挖掘建模分析应当充分考虑数据丢失和错误问题，解决方案应当能够容忍数据的丢失和错误。

数据的数量

对于物联网应用，数据的海量性是基本共识。总结起来，表现在数据表的行和列都很多，事物的

个体多，之间的关系多且复杂，场景变化多也很复杂，不可预测的因素多。与传统高性能数据挖掘相比，物联网应用侧重点在于后几方面：关系、变化、场景、不可测因素。

数据的非结构化

在传统的数据挖掘应用中，通常人们习惯的数据格式是由特征向量构成的， (f_1, f_2, \dots, f_k) 。其中， f_i 是数据分量。例如，对于描述学生的特征向量可以是：(姓名、性别、年龄、年纪、成绩、...)。大多数分类和聚类算法都是在特征向量上直接操作的。但是，大多物联网应用的原始数据都不是结构化的。例如，图结构、序列、读入连续测量值等。这样一来传统算法无法直接应用。在许多应用中，传感器读取的数据都是在特定时空中的连续值或者状态，物联网是个时空相联的图，而不是一个天然有特征项量的结构。从非结构化角度讲，物联网应用对于数据挖掘主要的挑战是：(1) 怎样自动抽取有用的特征，然后去适用已有的算法；(2) 或者提出直接在时空非向量空间中直接操作的算法。第一种办法相对简单直接，但是会有很多效率、方便性等方面的限制，而且会丢掉一些重要信息。第二种方法需要从数据表示到算法各个方面的创新。

以上谈了物联网的数据特性及其对数据挖掘的一些要求。下面着重讨论物联网数据建模。

基于超图的物联网数据模型——广义超图

鉴于以上认识，我们建议考虑超图和其扩展(广义超图, Generalized Hypergraph)来对物联网进行描述建模和挖掘。

超图是图概念的推广。在超图中，一个边可以和任意多的点联结。例如：图2。

在图2中，点集合 $X = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$

超边集合 $E = \{e_1, e_2, e_3, e_4\} = \{\{v_1, v_2, v_3\}, \{v_2, v_3\}, \{v_3, v_5, v_6\}, \{v_4\}\}$ 。

和普通图相比，超图最大的特点是超边可以把

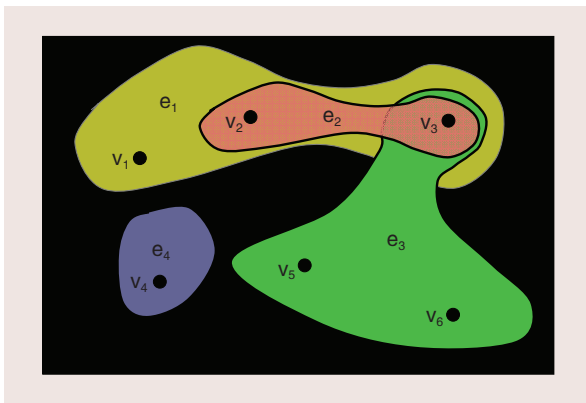


图2 超图示例

任意多的点结合起来。用超边可以代表物联网中错综复杂的关联关系。在物联网应用中，超图中的点可以代表每一个对象或者事物，超边代表多个物体之间的直接关联。为了表示扩充的语义，每个边上可以扩充包含信息描述超边代表多个对象之间的关联信息。例如，(1) 在智能交通系统中，每个点可以代表车，超边可以代表一起出行的团队，扩充的超边信息可以代表目的地信息等；(2) 在食品安全应用中，每个点可以代表成品，超边代表从同一个供应商来的原材料；(3) 在智能电网中，每个点可以代表一个供电源，超边可以代表太阳能车或者智能蓄电车。

在关联复杂的情况下，也可以用多个超图描述同一个物联网。其中，每一张图代表同一种关联关系的存在。另外，每一个点同时也包含刻画其自身的信息。为了简便，时间和位置信息可以记录在点信息中。超图可以用多矩阵(张量, tensor)的办法表示，也可以用集合的办法表示。这样可以引入许多图、集合和线性代数的操作。

基于广义超图的基本数据挖掘操作

子图挖掘

子图是由图的部分连通点和边构成的。在上面的例子中， $\{e_1, e_2, e_3\}$ 构成了子图。在常规图中，人们发现通过子图挖掘(subgraph mining)可以抽取有判

断能力的信息 (Discriminative Information)。在广义超图中,子图挖掘也应当能抽取重要的信息。由于在广义超图中我们扩展了超边和超点的信息,子图的包含和同构是模糊定义的。归纳起来,子图可以用来抽取图中的重要信息,从物联网中提取特征向量,这样一来很多已有的预测和寻证算法都可以使用。子图挖掘是NP难问题,而且大多数子图都不具有辨识信息。在“Direct Mining of Discriminative and Essential Frequent Patterns via Model-based Search Tree”

(SIGKDD'08)一文中,提出了一个称为MbT的贪婪算法来搜索有用子图。用传统的方法,搜索到的子图的频度 (support) 很低 (可以低到0.01%), 在没有搜索到之前虚存已经消耗完,所以根本无法找到此类子图。该算法的基本思想在图3中示例。类似的算法应当能用于物联网的数据挖掘,在时空网络中找到关键有效成分。

图分解

在大规模社会网络的应用中,图分解 (factorization) 对于发现网络中的问题和规律 (信息传播、兴趣爱好、好友等等) 都有不错效果。在电路设计中,图分解对于布线很有用。我们设想,在物联网应用中,广义超图分解应当能够发现大规模图中的规律性的核心成分,可以用来预测和寻证分析。

基于马尔可夫链的状态描述

在马尔可夫链模型下,预测将来只需要当前知识或信息,过去对于预测将来是无关的。许多物联网应用都属于这一类问题。例如,智能电网的稳定

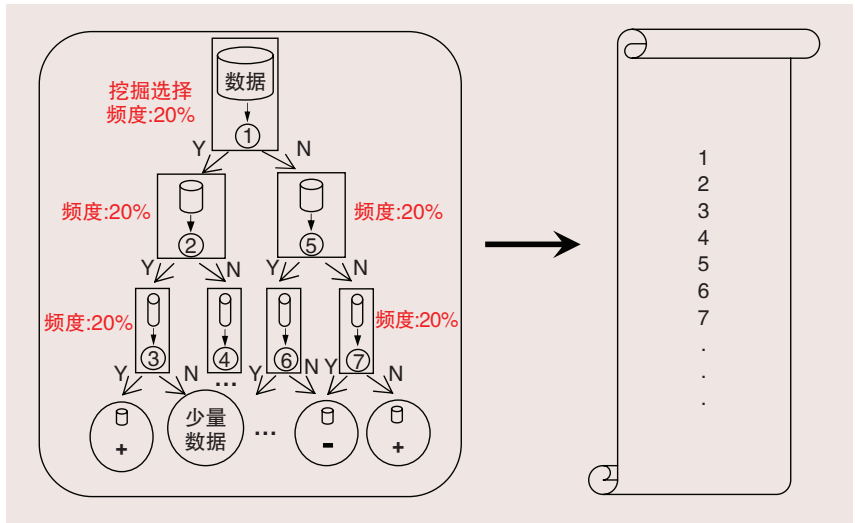


图3 MbT算法示例

通过建立基于模型的搜索树来直接挖掘有分辨信息和基本的模式。该算法通过子空间不完整枚举和信息增益等标准在子空间内找到次优特征,然后用递归的办法在更小的空间中搜索。最终得到少量的有用特征向量。

性估计,智能交通的拥阻估计,环境污染等。广义超图是物联网的状态表示,对应于马尔可夫链的状态。怎样在计算模型中快速地算出复杂状态 (广义图) 的状态迁移概率是值得探讨的问题。应当主要考虑怎样用广义超图去刻画描述状态,怎样刻画状态之间的差异,什么样的差异是根本性的。在应用场景中,当实际采集到的状态和模型记录状态不是100%相符时,怎样做状态匹配,等等。但是,从根本上,广义超图状态匹配还是子图比较问题。换言之是找出包含不同子图的过程。这样,上述KDD'08一文的MbT算法可能会有用场。

稳定的可外推非参数模型 (Numerically Stable Non-parametric Model that can extrapolate)

如上述,在物联网应用中数据质量差 (由于数据丢失、出错所造成)、数据复杂度高是很突出的问题,传统物理建模方法会有很大局限性。通常物理建模都首先需要了解事物间的关系,提出假设,建立数学模型描述数量上的关系,然后收集数

据，做实验和数据拟合，等等。在过去几百年中，这一基本方法解决了无数多人类社会生产和实践中的实际问题。但是，参数模型的根本问题是开发周期长，人力投入大，很难自动化。在物联网的应用中，我们主要的障碍是数据量太大、关系太复杂、可见和不可见的参数都太多。物联网的应用又非常广泛，场景多，不可测情况多。这样一来传统物理建模方法的多种弊端都会暴露出来。

近年来，在数据挖掘领域开发了一系列“非参数”模型（Non-parametric Method）。非参数模型的主要特点是，学习的过程本身就是发现模型或者函数的过程。通常是用自由格式函数（Free form Functions）对数据进行拟合。“决策树”（decision trees）是最常用的自由函数。在拟合的过程当中，树的结构和参数都是用数据和算法自动发现的。非参数模型最大的优点是对数据质量要求不高而且可以处理大规模数据。但是，非参数模型通常不能外推（当出入只在训练集合之外时，模型误差大），应当扩张非参数模型在这方面的优势，解决外推的局限。比较成功的非参数模型有，随机决策

树（Random Decision Trees）和随机森林（Random Forests）等。随机决策树的主要思想是在模型的构造过程中随机的选择可以用的变量，而不用信息增益等标准来做特征选择。它的主要好处是速度快，人为因素少，对数据质量低不敏感。随机决策树的主要步骤在图4中显示。

在众多应用中，特别是数据质量不高的场合，随机决策树和其他类似算法都体现出了很多优越性，在物联网应用中的着重点是怎样扩充外推功能。

物联网涉及信息技术的各个领域，在计算机领域内的许多方向都会与物联网开发和应用相关。《海计算：物联网的新型计算模型》一文中的一个要点是提倡以云计算设施来支撑物联网的计算需求。就计算模型而言，目前在云计算概念下主要是以MapReduce/Hadoop为核心的计算方式。这种计算方式的主要特点是“数据并行（Data Parallel）”，即对大量数据进行规范统一的处理与加工，不同数据之间的差别甚微。在物联网应用中，由于个体之间的区别，关系复杂，处理起来可能很难形式化为大规模数据的统一规整操作，而应当是很多对个体

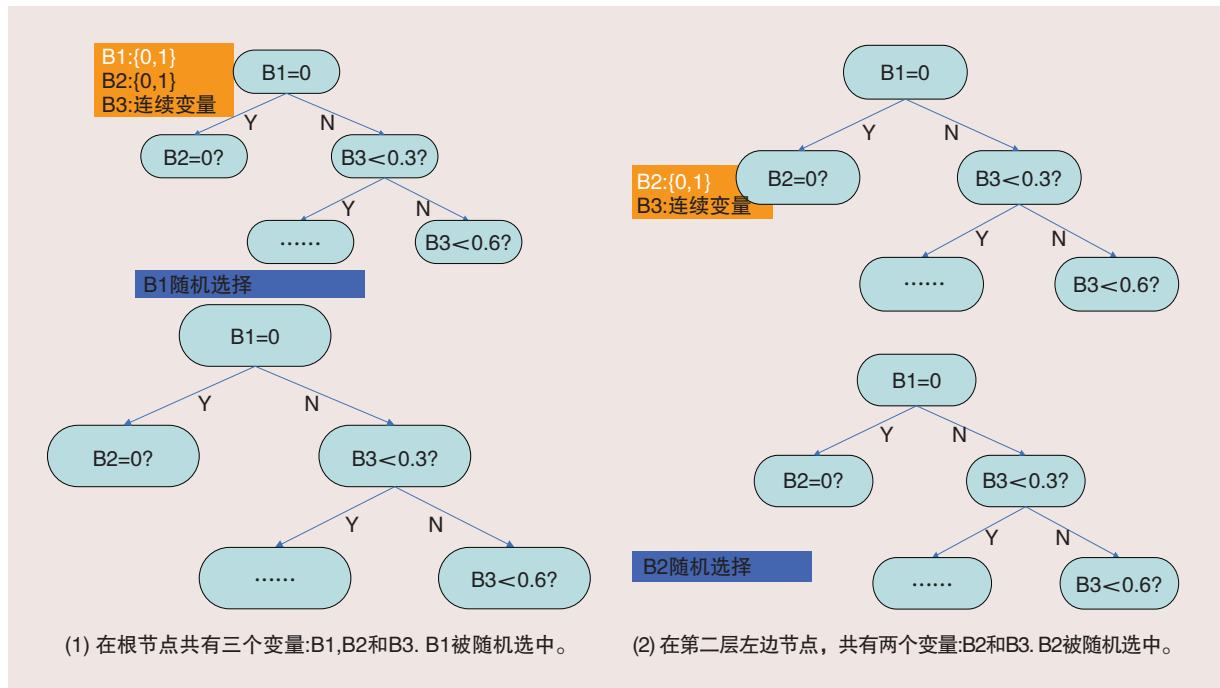


图4 随机决策树

或者小群体的操作，然后做集中处理。这样一来，传统云计算的模式/语言就不合适了。相比而言，类似“流计算”的结构/语言对于个体和小群体的加工灵活度大。所以，对于体系结构和编程环境的研究，也许应当集中于设计与个体及其复杂关系相适应的结构、环境和语言。

本文主要从数据挖掘与分析的角度讨论了物联网应用必须面对的若干问题。我们看到，物联网中产生的数据的确显示出与传统数据挖掘环境相当不同的特性，它们对建模和处理提出了新挑战。■



范 伟

IBM美国华讯研究院和IBM中国研究院研究员。主要研究方向为数据挖掘、机器学习。wei.fan@gmail.com



李晓明

CCF副理事长，本刊编委，北京大学教授，主要研究方向为网络信息搜索与挖掘、并行与分布处理。

览

中