



一种基于投影寻踪的电信网络数据处理方法

赵晋明

(山西省太原市文通电子有限公司, 山西 太原 030002)

摘要:在大数据时代,运营商采集和处理的数据出现了爆炸式增长。与此同时,采集数据维度增长使得从数据中提取信息的难度大大提高,导致了运营商新增的计算能力并没有带来与之匹配的信息处理能力提升。造成这一现象的主要原因在于并非所有维度上都有运营商关心的信息,过高的维度反而带来额外的噪音。因此,根据数据特性选择适当的维度十分必要。基于上述考虑,提出了一种基于投影寻踪的数据处理方法,并通过基于实际网络数据的对比验证了本文所述方案的有效性。

关键词:网络管理;降维;投影寻踪

中图分类号: TN915

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016322

Data processing method for telecom data based on projection pursuit

ZHAO Jinming

Wentong Company Limited, Taiyuan 030002, China

Abstract: In the big data era, the data which operator can collect and process is quickly growing. At the same time, due to the high dimension of the data, it is very difficult to extract and process information from them. This makes the waste of the computing capability of operators, even may lead to the decrease of performance due to the growth of the data dimension. The main reason of this phenomenon is that not all dimensions are interested with the information. Therefore, it is necessary to select the appropriate dimensions according to the data characteristics. Based on this, a data processing method based on projection pursuit was proposed. The result demonstrates the effectiveness of the scheme.

Key words: network management, dimension reduction, projection pursuit

1 引言

通信网络中存在着大量的复杂事物及现象,网络管理人员一直希望通过网络参数的提取来揭示隐藏在这些纷繁复杂的表象下的现象和客观规律。长久以来,网络管理者不断研制新的网络管理工具,发展新的性能测试技术。对于当前的网络状况,可以用来描述小区语音话务量、小

区语音质量、小区数据等效话务量、小区数据质量、掉线率、语音信道拥塞率、无线接入性、无线利用率等几十个指标。因此,对于每时每刻的网络状况,可以用多变量组成的数据细致地表示。这些数据,抽象出来就是高维数据。这些数据提供了网络中极其详细的信息。但是,数据维数的大幅度提高也给随后的数据处理和分析工作带来了巨大的困难。

网络管理系统作为一个数据处理系统,通常需要在输入数据的信息量较高的情况下才能有效工作。而当输入向量维数过高时,如何寻找有效信息将变得更加困难,导致“维度灾难”^[1,2]——系统计算量增加而效能无法提升。因此,必须采取一定的措施使系统能够有效地从数据中获得信息,发挥出计算能力的最大效能,这就是“降维”。

降维的核心思想就是要将输入数据的维数降低到一个合理的范围内,并尽可能多地保留数据中的原始信息。对于网络管理系统这样需要处理海量数据的系统,这种做法不仅是有用的甚至可能是必须的,图 1 描述了高维数据降维的过程。

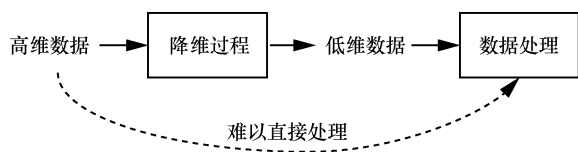


图 1 高维数据降维的过程

在大量的工程实践中,很多表面上看是高维、复杂的问题,实际可以通过很少的变量(特征向量)来描述,例如人脸识别、语音识别、基因测序等问题。然而,除了个别领域,目前国内外对降维的研究更多集中在理论探索,在工程中的具体实践还相对较少。英国谢菲尔德大学的Carreira^[3]在对各种降维方法综合分析研究的基础上,提出了一种抽象的模型;美国斯坦福大学的Donoh等人^[4]对于高维数据空间进行了比较深入的研究,获得了很有指导意义的结果。国内相关领域的研究还比较少。电信网管系统具有海量的数据,对这些数据的有效处理和准确分析具有十分重要的实际意义。本文从电信网管系统的具体需求出发,结合国内外相关研究成果的思想,提出了一种基于降维思想的数据处理方法,提升了数据处理过程的准确性和处理效率。

2 综合网管数据分析的现状和问题

随着计算机系统的普及,早在 1961 年,人们就已经认识“维度灾难”^[1]。具体而言,“维度灾难”描述了对给定问题估算的精度要求下,所需的样本点数随着变量个数的增加以指数形式增长。这样的要求给系统带来了巨大的负担。例如,对于一个目标光滑函数的拟合是基于观察数据的局部均值,但为了保证精度,必须找足够多的点,使得光滑函数在很多区域失去了局部性,带来不能接受的误差。

高维空间体现出许多在低维无法想象的特殊性质,最

显著的就是高维空间中的数据分布是“稀疏”的。具体而言,对于任意 d 维空间 R^d ,其中半径为 r 的球体的空间体积 $V_d(r)$ 可以表示为:

$$V_d(r) = \frac{2r^d \pi^{d/2}}{d \Gamma\left(\frac{d}{2}\right)} \quad (1)$$

这样的超球体存在一种低维下十分罕见的特性,即其“体积”的大部分存在于表面上,其“表面”与“整体”的体积比例可以写成:

$$f_d = \frac{V_d(r) - V_d(r-\varepsilon)}{V_d(r)} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d \quad (2)$$

于是:

$$\lim_{d \rightarrow \infty} f_d = 1 \quad (3)$$

在这种情况下,对于由网络参数构建的高维空间,传统的多元统计方法无法得到准确的结果,因此对采集到的数据进行适当的降维就十分必要。具体而言,降维的数学解释就是将高维数据投影到一个维数比原数据空间小得多的流形上。而降维操作的目标就是获得这一流形的低维坐标表示。

具体而言,降维问题的模型为 (T, F) ,其中, $T = \{t_i\}_{i=1}^N$ 是 D 维空间中的数据集合,寻找一个可以一一对应的映射 F 实现 $T \rightarrow X$,其中每一个元素的对应关系为 $t \rightarrow x = F(t)$,若 X 为 L 维空间,且 $L < D$,则认为 F 是数据集 T 到 X 的降维,其示意如图 2 所示。

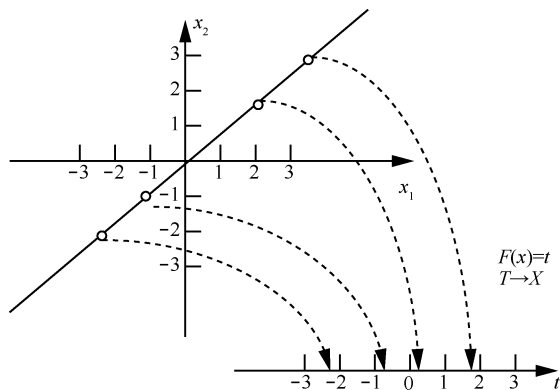


图 2 降维过程的数学含义

在通信网络中,由于各种因素的影响,如采样噪声、参数设计不合理、人为干扰因素等,使得网络采集数据中的信息被大量噪声所掩盖。通常,很多参数所携带的信息远小于其引入的噪声,因此,把它们“过滤”掉可以帮助网络管理人员更好地提取信息。从实用的角度来看,对于通信



网络参数的降维就是在保留网络结构信息的条件下尽可能减少所用参数的数量、降低数据中的噪音,使网络管理系统可以最大程度地实现对网络状态的客观评估。

3 基于投影追踪的数据分析方法

在降维的各种方法中,投影寻踪 (projection pursuit, PP) 是用来分析和处理高维数据,尤其是来自非正态总体的高维数据的一种有效方法。20 世纪 70 年代初, Kruskal 开创性地通过把数据投影到低维空间,以极大化某个指示参数的方式迭代出最佳投影结构的方法。投影追踪通过将原数据“投影”到某几个信息量最大的线性组合上,使得信息被保留的同时减少了数据的维度和数据中的噪音。目前,投影追踪的方法已经在多个领域进行了有效的推广,取得了较好的效果。

在电信网管系统中,可以采集到多种维度的数据。这些不同维度数据的线性组合通常是接近高斯分布的^[5]。从信息的角度来看,高斯分布几乎全部是噪音。在这种情况下,为了减少噪音,一方面需要减少输入参数的维度,另一方面需要通过线性变换,将原数据映射到信息含量更高的维度上。上述两方面可以通过采用投影跟踪来实现——通过选择信息量最高的少数几个投影方向进行坐标变换。

投影追踪就是实现上述目标的一种数学方法,即通过在原始数据的不同投影空间上寻找信息量最大的少数几个投影方向来降低数据的维度。具体来说,假设网络管理系统采集到的 d 维参数集为 X ,则降维的目标可以表述为找到最佳映射 $F \rightarrow R$,假设所采用的映射方法是线性的,则该映射可以写为: $F=AX$,其中 A 为转换矩阵。这样,降维的目标就是寻找最优转化矩阵 A ,使得 X 在新坐标下表现出更优的线性结构。

寻找最优转化矩阵的评价标准为 $Q(AX)$,为了简化运算电信网管系统的计算和管理负担,可以采用方差作为评价指标,即 $Q(a^T X) = Var(a^T X)$ 。这样,对于参数集 $X = \{x_1, x_2, \dots, x_n\}$,其方差可以写作:

$$Q(a^T X) = \frac{1}{N} \sum_{i=1}^N (a^T X_i - E(a^T X))^2 \quad (4)$$

这样,找到最优的 $a^T X$,得到的就是样本 a 散布最大的方向。多元正态分布线性投影依然是正态分布,因此如果在某个投影方向上找到与正态分布差别较大的数据集,那它就一定含有更多信息,因此在这个方向上进行数据分析可以更快、更精确地挖掘出其中所蕴含的信息^[6]。因

此,这样一种指标也成为信息散度。更精确的说,对于数据集 f 和 g ,其所对应的数据散度为:

$$d(f|g) = \int_{-\infty}^{\infty} g(x) \lg \frac{f(x)}{g(x)} dx \quad (5)$$

投影追踪就是基于信息的这一性质,即不断寻找能反映原高维数据信息的投影向量,通过对投影向量的分析来获得原高维数据中的有用信息。更具体地说,对于电信网络中的高维数据,可以采用如下步骤进行降维处理。

步骤 1 首先选定正态分布作为向低维映射的标准。

步骤 2 将原始数据在正态分布上进行投影,找到最远离正态分布的投影方向,这一投影方向上包含了最大的有用信息。

步骤 3 将上述投影过程中,与原正态分布相似度较大的方向上的数据删除,得到新的数据集。

步骤 4 对新数据集重复步骤 2、步骤 3,直到所有的投影方向上得到的映射都远离正态分布,即将原数据集中的全部有用信息提取完毕。

在实践中,这一方法可以快速找到最能反映网络状态的参数,提升网络管理人员对网络状态的理解。另外,投影寻踪的方法可以将高维数据投影到一维子空间,使得可以方便地找到影响网络某一性能的最主要因素。这一特性对于网络管理工作提供了极大的便利,在下文中将通过基于实际数据的案例对这一过程进行详细说明。

4 本文方案的验证与比较

为了验证本文所提出的电信网管系统中数据处理的方法,本文以运营商的实际数据为例,对比不同方案的数据分析过程得到的效果。原始数据为某省会城市 10 000 个小区的各项采集参数。将原始数据集进行可视化,得到的结果如图 3 所示。可以看到,由于原始数据中的相关性十分复杂,无法直接获得有关网络状态的直观信息。因此,需要对原始高维数据进行处理,以获得有关网络性能更加直观的关系。

例如,为了获得小区掉线数量的影响指标,可以按照本文所述方法对原始数据进行降维。根据上文所述过程,基于投影追踪的方式分析得到影响小区掉线的主要指标及影响因素(投影长度),得到结果分别为:小区数据流量(1.342)、小区载频数量(0.312)、小区 CPU 利用率(0.248)、小区语音总量(0.219)。可以看到,小区的数据流量是影响小区掉线率最重要的指标,其影响显著远超其他指标的影响。

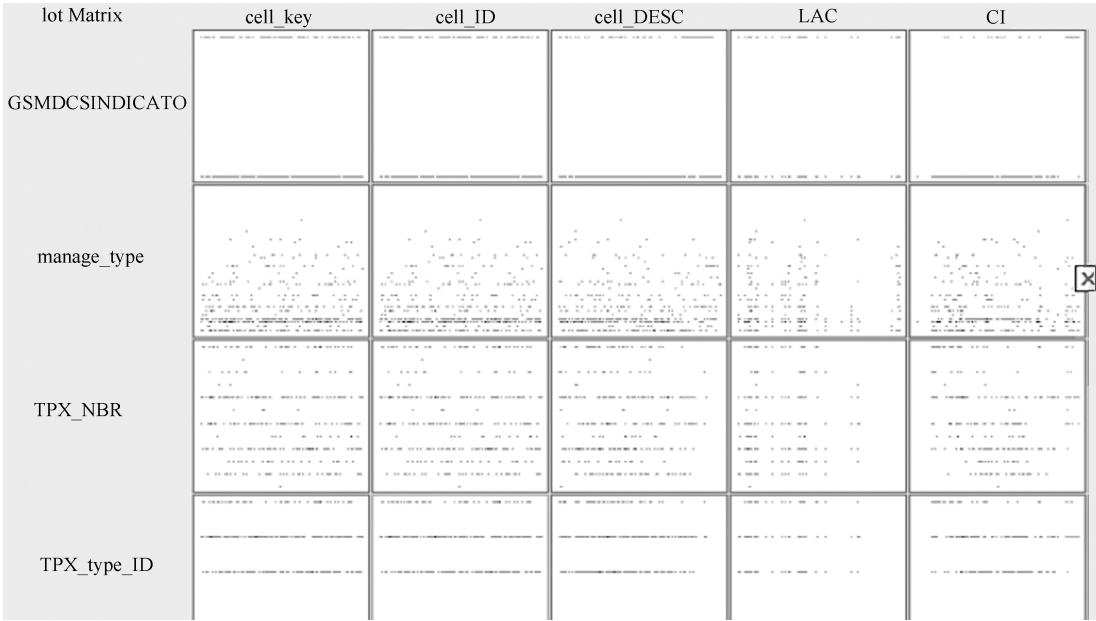


图3 电信网管系统采集到的原始高维数据集

为了验证这一结果的有效性,对两组样本进行分析,其结果如图4、图5所示。其中第一组样本是相邻两个月掉线次数变化与网络流量变化之间的关系;第二组样本为相邻两个月掉线次数变化与通话时长变化的关系。可以看到,对于所选取样本,掉线与数据流量之间存在明显关系而与语音通话时长关系并不明显,可见,掉线成因的先验判断可能会给网络状态的评估带来很大的不确定性。

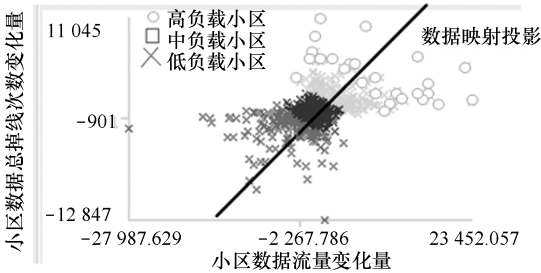


图4 无线小区掉线次数变化与数据流量变化的数据映射投影

与此同时,为了验证本文所提方法的实际性能,本文基于上述样本对比本文所提出算法与传统处理高维的联机分析处理(on-line analytical processing, OLAP)方法。其中样本为市中心高负载小区,且存在由于CPU负载过高引起过高掉话次数的现象。本文将所选小区分为10组,各自通过上述高维数据比较对掉线发生成因判断的准确性,结果如图6所示。

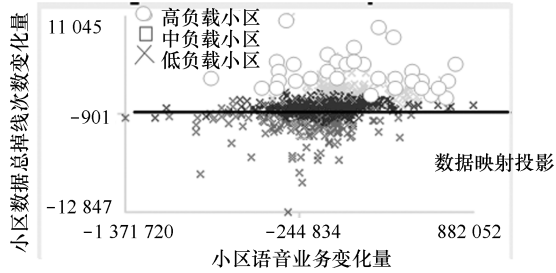


图5 无线小区掉线次数变化与语音业务时长变化的数据映射投影

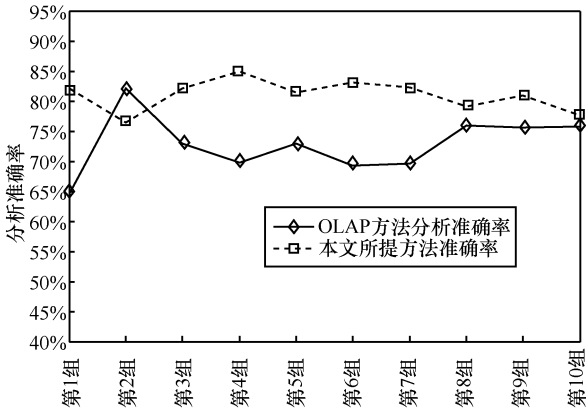


图6 本文所提方法与OLAP方法的性能比较

从结果可以看到,本文所提出的数据处理方法有效地提高了数据分析的准确性。对于某一业务而言,其质量并不完全取决于网络参数,但是,通过不断地对数据进行分



析,可以在很大程度上及时发现网络运行中存在的问题。因此,这样的准确率已经基本可以满足运营商对于网络管理的要求,同时也说明本文所提出的分析方法具有一定的适用价值。

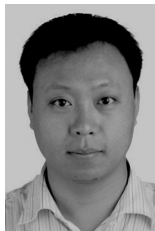
5 结束语

对数据的处理能力未来将成为运营商增强竞争力、提高网络管理效率的重要手段。因此,对数据的高效分析是运营商的重要资产。本文从提升运营商的数据处理能力的实际需求出发,提出了基于投影寻踪实现降低数据维度的有效方法。最后,通过基于实际数据对比验证证明了本文所提出方法的有效性。本文研究成果对未来运营商的数据使用和分析的相关研究有着较大的借鉴意义。

参考文献:

- [1] BELLMAN T E. Adaptive control processes [M]. Princeton: Princeton University Press, 1961.
- [2] MAATEN L J P V D, POSTMA E O, HERIK H J V D. Dimensionality reduction: a comparative review [J]. Journal of Machine Learning Research, 2007, 10(1).
- [3] CARREIRA-PERPINAN M A. Continuous latent variable models for dimensionality reduction and sequential data reconstruction [EB/OL]. (2001-01-29) [2016-11-02]. http://xueshu.baidu.com/s?wd=Continuous+Latent+Variable+Models+for+Dimensionality+Reduction+and+Sequential+Data+Reconstruction&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Dimensionality+reduction%3A+a+comparative+review&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_n=2.
- [4] DONOHO D L. High dimensional data analysis: the curses and blessings of dimensionality [EB/OL]. (2000-01-29) [2016-11-02]. http://xueshu.baidu.com/s?wd=High+dimensional+data+analysis%3A+the+curses+and+blessings+of+dimensionality&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Continuous+Latent+Variable+Models+for+Dimensionality+Reduction+and+Sequential+Data+Reconstruction&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_n=2.
- [5] DIACONIS P, FEREDMNA D. Asymptotics of graphical projection pursuit [J]. Annals of Statistics, 1984, 12(3): 793-815.
- [6] 张维明. 数据仓库原理与应用 [M]. 北京: 电子工业出版社, 2002.
- ZHANG W M. The principle and application of data warehouse [M]. Beijing: Publishing House of Electronics Industry, 2002.

[作者简介]



赵晋明(1973-),男,山西省太原市文通电子有限公司研发主管,主要研究方向为网络管理系统的架构和关键算法,具有丰富的理论和工程经验,为中国移动通信集团公司等运营商解决了大量运维难题,并实现多项关键技术突破。