

基于贝叶斯网络的电网事故预警模型

刘芸菲, 陆奎

(安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001)

摘要:近年来电力生产与作业过程中的安全问题频频发生,给人民生活带来不便,严重阻碍了电力行业的进一步发展。文中基于贝叶斯网络寻求一种电网事故预警模型。贝叶斯网络使用有向无环图来表示随机变量之间的条件独立关系。每个贝叶斯网络定义随机变量的联合实例化的联合概率分布。并且贝叶斯网络被认为是概率知识表示和推理领域中最有用的模型之一。为了构建模型,寻求一种新颖的贪婪算法。首先,统计电网事故致因数据属性的各个概率;其次,计算各个属性之间的互信息;最后,利用贪婪算法,每次选取最大的互信息属性,实现了贝叶斯网络的构造,进而完成电网事故预警模型的构建。使用 Adult 数据集进行模拟实验,结果表明构造的贝叶斯网络可以带来较好的分类效果。

关键词:电力生产;事故预警;贪婪算法;互信息;贝叶斯网络

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2019)08-0152-04

doi:10.3969/j.issn.1673-629X.2019.08.029

Early Warning Model of Power Grid Accident Based on Bayesian Network

LIU Yun-fei, LU Kui

(School of Computer Science & Engineering, Anhui University of Science & Technology,
Huainan 232001, China)

Abstract: In recent years, the frequent occurrence of safety in electric power production and operation has brought inconvenience to people's life and seriously hindered the further development of the electric power industry. We seek a grid accident warning model based on Bayesian network. Bayesian networks use directed acyclic graphs to represent conditional independent relationships between random variables. Each Bayesian network defines a joint probability distribution of joint instantiations of random variables, which is considered to be one of the most useful models in the field of probabilistic knowledge representation and reasoning. To build the model, we seek a novel greedy algorithm. First, the probability of the data attribute of the cause of the grid accident is calculated. Secondly, the mutual information between the attributes is computed. Finally, the greedy algorithm is utilized to select the largest mutual information attribute each time for construction of the Bayesian network, and then an early warning model for power grid accidents is constructed. We use the Adult dataset for simulation which shows that the Bayesian network constructed can bring better classification effect.

Key words: electric power production; accident warning; greedy algorithm; mutual information; Bayesian network

0 引言

目前,社会经济正处在高速发展的时期,电网规模持续扩大。电力的稳定供应不仅关系着企业的生存、效益和发展,更影响着人民生活、国家经济发展和社会稳定的大局^[1]。然而,近年来电力生产与作业过程中的安全问题频频发生,给人民生活带来不便,严重阻碍了电力行业的进一步发展。因此,为了降低电力行业事故的发生率,需要对可能发生危险的生产或检查工作进行预警,这样,才能更好地应对异常情况的

发生^[2-3]。

贝叶斯网络是应用广泛的预测模型之一,其使用非循环有向图编码随机变量之间的条件独立关系。非循环有向图可用于“读取”条件独立关系,从而提供对由贝叶斯网络表示的联合概率分布的结构洞察^[4]。

贝叶斯网络学习的一种标准方法是“搜索和得分”^[5]。选择得分以表示观察数据(和任何先前知识)支持任何候选 BN 的程度,然后进行搜索以找到具有最大得分的贝叶斯网络。因此,贝叶斯网络学习是一

收稿日期:2018-09-24

修回日期:2019-01-16

网络出版时间:2019-03-27

基金项目:国家自然科学基金(51274011,61772033)

作者简介:刘芸菲(1992-),女,硕士,研究方向为隐私保护、人工智能;陆奎,博士,教授,硕导,研究方向为人工智能、计算机网络。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1629.046.html>

个优化问题^[6-7]。不幸的是,即使非循环有向图中任何顶点的父亲数限制为2,这个优化问题对于任何合理的分数都是NP难的^[8]。

鉴于这种结果,大多数贝叶斯网络学习工作集中在启发式搜索上,其中不能保证找到最优贝叶斯网络。然而,关于精确贝叶斯网络结构学习的研究越来越多。一种方法是使用动态编程^[9-10],它已经成功地用于精确学习多达30个顶点。所以,文中集中于两个关键:首先,寻求一种新颖的贪婪算法,以互信息为衡量标准,旨在解决网络学习中的NP难问题;其次,将贝叶斯网络应用于电力事故的预警,根据典型事故中的致因因素来预测可能发生的事故等级。并根据等级的不同警示电网领域,做好事故的应对方案。

具体贡献在于:构造贝叶斯网络,根据互信息,为属性间的依赖关系提供了一种衡量方法;再者,通过互信息和贪婪算法的结合,优化了整个网络的构造过程。电网事故预警,在完成网络构建的同时,成功做到电网

预警的目的。

1 相关知识

电网的事故产生包含多种因素,如高空坠落、接地线不良和触电等。设 A 为电网事故的致因数据集属性, d 为 A 的大小, A_d 为最后电网事故的预测等级,分别为等级1、等级2、等级3、等级4和等级5。则 A 上的贝叶斯网络^[11]定义为一组属性对(AP),即 $(A_1, P_1), \dots, (A_d, P_d)$,则有:

- (1) 每个 A_i 都是 A 中的唯一属性;
- (2) 每个 P_i 是 A 中属性的子集。 P_i 是贝叶斯网络中 A_i 的父集合;
- (3) 对于任何 $1 \leq i < j \leq d$,有 $A_j \notin P_i$,即贝叶斯网络中没有从 A_j 到 A_i 的边缘,这确保了网络是非循环的。

图1显示了基于六个事故致因属性的贝叶斯网络。

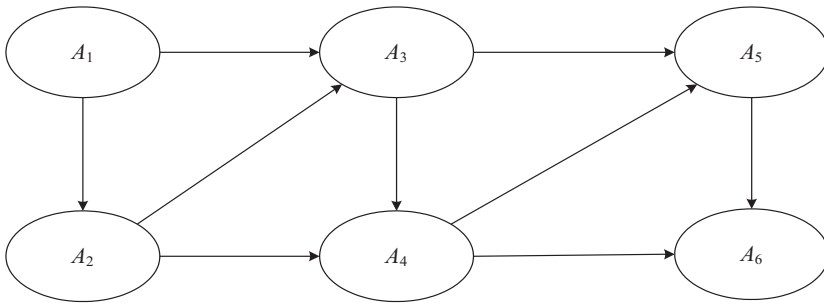


图1 六个电网致因属性的贝叶斯网络

2 贝叶斯网络的电网事故预警模型

2.1 私有方法

目的是构建一个 k 度贝叶斯网络。首先使用互信息来测量属性之间的相关性,域中的属性集 A 的熵 $H(A) = - \sum_{a \in \text{dom}(A)} \text{Pr}[A = a] \log \text{Pr}[A = a]$ 。 $I(A, P)$ 表示属性集 A 和 P 之间的互信息,如下:

$$I(A, P) = \sum_{a \in \text{dom}(A)} \sum_{p \in \text{dom}(P)} \text{Pr}[A = a, P = p] \log \frac{\text{Pr}[A = a, P = p]}{\text{Pr}[A = a] \text{Pr}[P = p]} \quad (1)$$

定义:最大AP对联合分布。给定AP对 (A, P) ,最大AP对联合分布 $M^*(A, P)$ 是最大化 A 和 P 之间的互信息的分布。

引理:假设 $|\text{dom}(A)| \leq |\text{dom}(P)|$, $M^*(A, P)$ 是最大AP对联合分布,如果:

- (1) 对于 $\forall a \in \text{dom}(A)$, $M^*(A = a) = 1/|\text{dom}(A)|$;
- (2) 对于 $\forall p \in \text{dom}(P)$,最多有一个 $a \in \text{dom}(A)$,其中 $M^*(A = a, P = p) > 0$ 。

证明:假设 $|\text{dom}(A)| \leq |\text{dom}(P)|$, $\log |\text{dom}(A)|$ 是互信息,则属性 A 和 P 之间的最大互信息是:
 $\max I(A, P) = \min \{ \max H(A), \max H(P) \} = \min \{ \log |\text{dom}(A)|, \log |\text{dom}(P)| \} = \log |\text{dom}(A)|$ (2)

假设 $M^*(A, P)$ 是满足引理中两个属性的联合分布。那么,给定信息论的基本结果,这两个属性相当于:

- (1) $H(A) = \log |\text{dom}(A)|$;
- (2) $H(A|P) = 0$ 。

因此, $M^*(A, P)$ 的互信息是:

$$I(A, P) = H(A) - H(A|P) = \log |\text{dom}(A)| \quad (3)$$

根据定义, $M^*(A, P)$ 是最大AP对联合分布。

另一方面,假设 $M^*(A, P)$ 是具有 $\log |\text{dom}(A_i)|$ 的最大联合分布,则互信息可表示为:

$$I(A, P) = \log |\text{dom}(A)| = H(A) - H(A|P) \quad (4)$$

其中 $H(A) \leq \log |\text{dom}(A)|$ 并且 $H(A|P) \geq 0$ 始终保持。因此,可以得出结论, $I(A, P)$ 在如下情况下

能实现最大化:

- (1) $H(A) = \log |\text{dom}(A)|$, 仅通过 $\text{dom}(A)$ 上的均匀分布实现;
- (2) $H(A|P) = 0$, 这意味着对于每个父集 p 存在属性 a , 使得 $\text{Pr}[A = a, P = p]$ 。

2.2 私有贝叶斯网络

为了构建 k 度贝叶斯网络, 对于 $k = 1$ 的情况, Chow 等^[12]提出, 根据最大互信息贪婪地选择下一个边缘是最佳的。Van 等^[13]指出, 当 $k > 1$ 时, 这个优化问题是 NP 难的。因此, 启发式算法^[14]经常在实践中使用。为了克服这个问题, 寻求一种新颖的贪婪算法。

- 算法: 贝叶斯网络构造。
- 输入: 数据集 DS, 限制参数 k ;
- 输出: 贝叶斯网络 BN。
- (1) 数据集 DS 清洗
 - (2) 初始化 $\text{BN} = \emptyset$ and $V = \emptyset$
 - (3) 从 A 中随机选择一个属性 A_i ; 添加 $(A_i, 0)$ 到 BN; 添加 A_i 到 V
 - (4) For $i = 2$ to d do
 - (5) 初始化 $\Omega = \emptyset$;
 - (6) for each $A_j \in A \setminus V$ do
 - (7) 找到 A 中所有最大父集
 - (8) for each $P_i \in V_k$ do
 - (9) 添加 (A_j, P_i) 到 Ω
 - (10) 从 Ω 中选择包含最大互信息 $I(A_j, P_i)$ 的一对 (A_i, P_i)
 - (11) Add (A_i, P_i) to BN; add A_i to V
 - (12) 结束
 - (13) 返回 BN

在算法(第 1~2 行)的开头, 对原始数据集 DS 进行清理操作(删除空值), 并将贝叶斯网络 BN 初始化为空的 AP 对列表。假设 V 是一个包含其父集已在 BN 的部分构造中被修复的所有属性的集合。作为下一步, 算法从 A 中随机选择一个属性(表示为 A_1)并将其父集 P_1 设置为 0(第 3 行)。算法的其余部分由 $d - 1$ 次迭代(第 4~11 行)组成, 每次迭代都贪婪地添加到具有大互信息的 AP 对中。具体而言, 在候选集中选择包含迭代期间所需的两个 AP 对:

- (1) $|P| \leq k$, 通过仅从 V_k 选择 P 来确保 BN 是 k

度贝叶斯网络, 其中 V_k 表示 V 的所有子集的集合, 并且其大小是 $\min(k, |V|)$ (第 4~10 行)。

- (2) 对于任何 $j < i$, BN 不包含从 A_i 到 A_j 的边, 这保证了 BN 是有向无环图。通过在迭代开始时请求 V 仅包含其父集(在先前迭代中已经确定的那个)(第 11 行)来保证该条件。

一旦确定了每个属性的父集合, 该算法就终止并返回贝叶斯网络 BN(第 12~13 行)。

2.3 时间复杂度分析

由于难以从具有低信噪比的小样本空间获得可靠的估计, 因此用户记录的数量需要足够大。

定理: 假设数据集 DS 的平均大小为 m , 贝叶斯网络节点属性集大小为 v , 则算法的时间复杂度是: $O(Nmkv + tNv^{2d})$ 。

证明: 算法将逐个扫描所有贝叶斯网络 BN 用户的记录, 其长度为 $k * m$, 因此时间复杂度估计为 $O(N(km)v)$ 。此外, 在 t 次迭代中, 在观察每个比特串时计算每个组合的后验概率将产生 $O(tNv^{2d})$ 的时间复杂度。因此, 总时间复杂度为 $O[N(km)v + tNv^{2d}]$ 。

3 实验

3.1 实验设置

为了模拟电网事故等级的预测问题, 采取美国 1994 年人口普查数据库抽取而来的成年人(Adult^[15-16])数据集进行实验。成年人数据集包含了 32 561 条成人数据, 并同时具有连续属性和分类属性。为了更好地衡量实验的效果, 选取种族, 国家, 教育, 阶级, 工作和收入六个属性。将文中的私有贝叶斯网络与 K-近邻算法和决策树算法进行对比, 并同时验证在收入是否超过 50K 情况下的分类效果。对于每个分类任务, 使用数据中 70% 的元组作为训练集, 另外 30% 作为测试集。

3.2 实验分析

为了构造贝叶斯网络 BN, 计算每个属性之间的互信息, 并且每次将具有最大互信息值的 AP 对添加到贝叶斯网络 BN。构造的贝叶斯网络如图 2 所示。

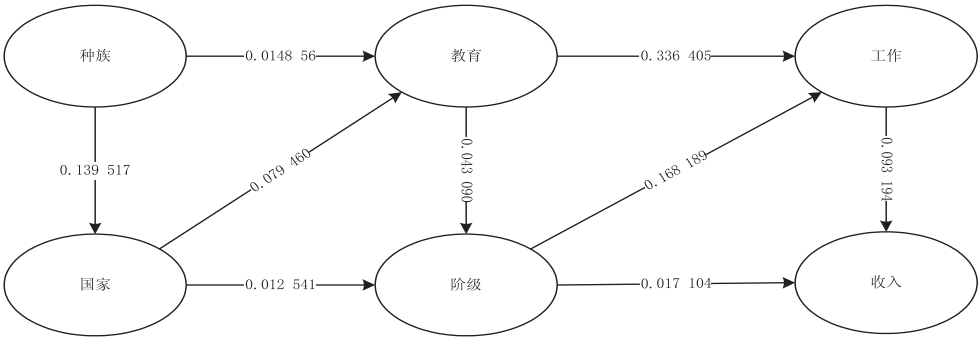


图 2 DS 中六个属性的贝叶斯网络

对于贝叶斯网络 BN 的训练集设置为 7 722 条记录。为更好地衡量 BN 的分类能力,现将 BN 划分为六个数量级进行分类预测,数据集划分为 $L=(1K, 2K, 4K, 8K, 16K, 30K)$ ($1K=1\,000$)。在图 3 中观察到的总体趋势是,随着实验数据集的增加,分类错误率呈现不同程度的波动。不同数据集的确切行为有所不同,在某些情况下,其数据集渐近趋近,而在另一些情况下,具有更多的 S 形。可以看出,这种行为主要是由于贝叶斯网络具有不同元数据捕捉数据分布的先天能力。在多组数据划分的数据集里,文中的私有贝叶斯网络的分类错误率普遍低于 K-近邻和决策树分类的效果。这是因为,文中的私有贝叶斯网络不仅成功实现了分类的目的,而且考虑了属性间的关系。

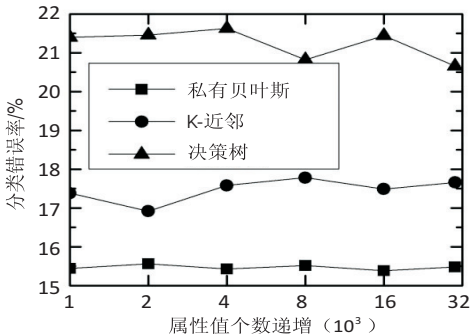


图 3 分类误差率对比

4 结束语

电网安全问题事关民生之本,迫切需要一种能够早发现、早预警的事故模型。基于以上情景,提出了一种基于贪婪算法和互信息的故事预警模型,可以高效、直观地对事故等级进行划分,这有利于电网日常的风险防范。特别地,文中提出了贪婪构造贝叶斯网络的方法。

贝叶斯网络模型已被证明是近似表示相关数据的有效方法,以这种模型发布的数据不仅高效而且准确,并比分类机制能提供更高的准确性。文中提出的私有贝叶斯网络的构造关键部分是利用互信息来构造相关性模型,这不仅提供了属性间的关联情况,更提高了发布数据的质量。实验结果表明,构建的私有贝叶斯网络具有较好的分类效果。

参考文献:

[1] 张 烈,王德林,刘亚东,等. 国家电网 220 kV 及以上交流保护十年运行分析[J]. 电网技术,2017,41(5):1654-1659.

[2] 唐震宇,蒲风霞. 电力信息网络安全管理探析[J]. 华东科技,2017(7):221.

[3] 沈 亮,王 栋,玄佳兴. 电力信息系统云安全风险分析与评估技术[J]. 电信科学,2018,34(2):153-160.

[4] 兰 杰,袁宏杰,夏 静. 基于离散时间贝叶斯网络的动态故障树分析的改良方法[J]. 系统工程与电子技术,2018,40(4):948-953.

[5] 霍利民,朱永利,张立国,等. 用于电力系统可靠性评估的贝叶斯网络时序模拟推理算法[J]. 电工技术学报,2008,23(6):89-95.

[6] 李硕豪,张 军. 贝叶斯网络结构学习综述[J]. 计算机应用研究,2015,32(3):641-646.

[7] 张 平,刘三阳,朱明敏. 基于人工蜂群算法的贝叶斯网络结构学习[J]. 智能系统学报,2014,9(3):325-329.

[8] 吴 欣,郭创新. 基于贝叶斯网络的电力系统故障诊断方法[J]. 电力系统及其自动化学报,2005,17(4):11-15.

[9] 张振海,王晓明,党建武,等. 基于专家知识融合的贝叶斯网络结构学习方法[J]. 计算机工程与应用,2014,50(2):1-4.

[10] 姚宏亮,张一鸣,李俊照,等. 动态贝叶斯网络的灵敏性分析研究[J]. 计算机研究与发展,2014,51(3):536-547.

[11] NOYES N, CHO K C, RAVEL J, et al. Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis[J]. PLOS One, 2018, 13(1):e0191625.

[12] STACHNISS C, KRETZSCHMAR H. Pose graph compression for laser-based SLAM [M]//Robotics research. [s. l.]:Springer, 2017:271-287.

[13] VAN BEEK P, HOFFMANN H F. Machine learning of Bayesian networks using constraint programming [C]//International conference on principles and practice of constraint programming. [s. l.]:Springer, 2015:429-445.

[14] NAZARI-HERIS M, MOHAMMADI-IVATLOO B, GHAREHPETIAN G B. Short-term scheduling of hydro-based power plants considering application of heuristic algorithms: a comprehensive review [J]. Renewable and Sustainable Energy Reviews, 2017, 74:116-129.

[15] KIM S, LEE H, CHUNG Y D. Privacy-preserving data cube for electronic medical records: An experimental evaluation [J]. International Journal of Medical Informatics, 2017, 97:33-42.

[16] YUAN C, MALONE B. Learning optimal Bayesian networks: a shortest path perspective [J]. Journal of Artificial Intelligence Research, 2013, 48:23-65.