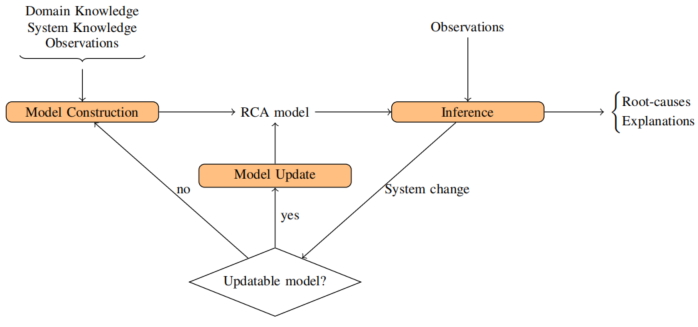
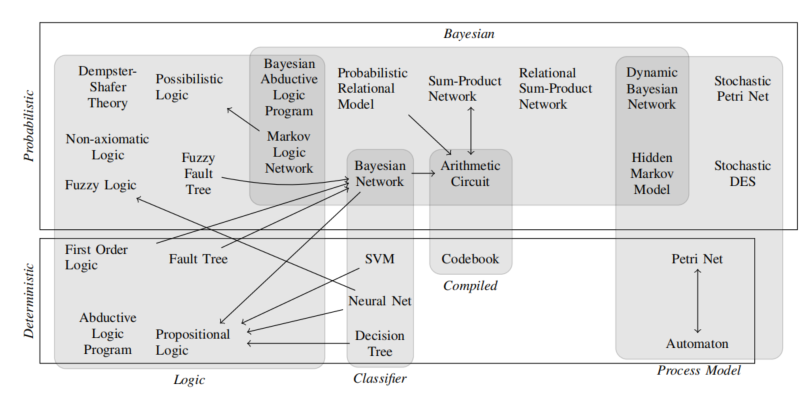
# RCA Survey

1. Background
   1. Terminology
      1. RCA: 也被称为故障定位、故障隔离或警报/事件相关性，是推断产生一组给定症状的一组故障的过程。RCA的核心概念即causality和explanation.
      2. Event: 在系统运行过程中出现的异常情况
      3. Faults/Problems/Root Causes: 造成其他事件的根本原因
         1. permanent
         2. intermittent
         3. transient
      4. Error: 错误是由一个或多个故障引起的，是系统的一个条件与其理论上正确的条件之间的差异。
      5. Failure: 从系统外部可以观察到的错误。
      6. Symptom: Failure的外部表征。
   2. RCA的评价指标
      1. 时间
         1. real-time diagnosis
         2. post-mortem diagnosis
      2. 复杂度
         1. system size: 需要诊断的组件数
         2. data size: 诊断中必须要处理的数据量
         3. inference length: 从symptom到fault必须要遍历的最大组件数
         4. effect propagation time: 对一个组件进行修改传播到临近组件的时间
         5. evolution rate: 在诊断系统中变化的速度和程度。
         6. domain knowledge required
         7. system knowledge required
   3. RCA 的工作流

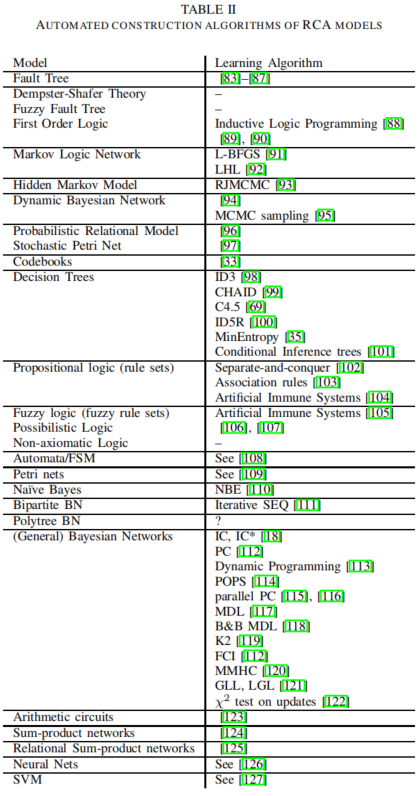


* 1. RCA的模型分类



* 1. 影响RCA表现的指标
     1. size：变量、规则、节点的数量
     2. inference structure: 不同模块相互关联的模式，主要指标为inference length
  2. RCA的生成模式
     1. manual generation:基于完全的领域知识
     2. assisted generation: 基于部分领域知识构造子模块，基于系统知识组织子模块
     3. automated generation:完全基于数据

1. 基于数据的RCA模型生成
   1. 基于机器学习的分类模型与对应学习算法



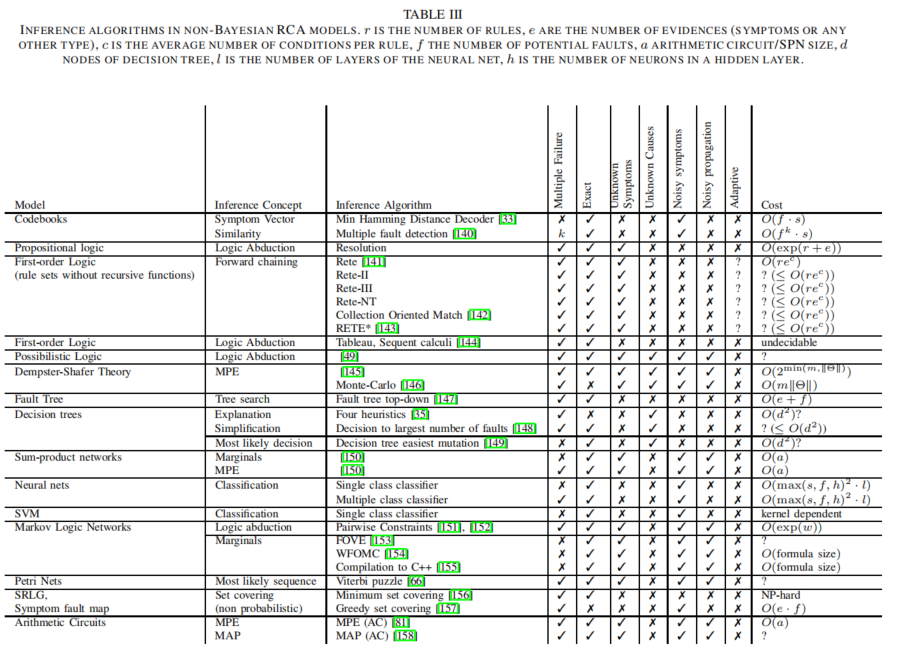
* + 1. 基于过程的模型
       1. automata
       2. Petri nets
    2. 非基于过程的模型
       1. decision tree
          1. Hoeffding Tree

Adaptive Hoeffding Tree

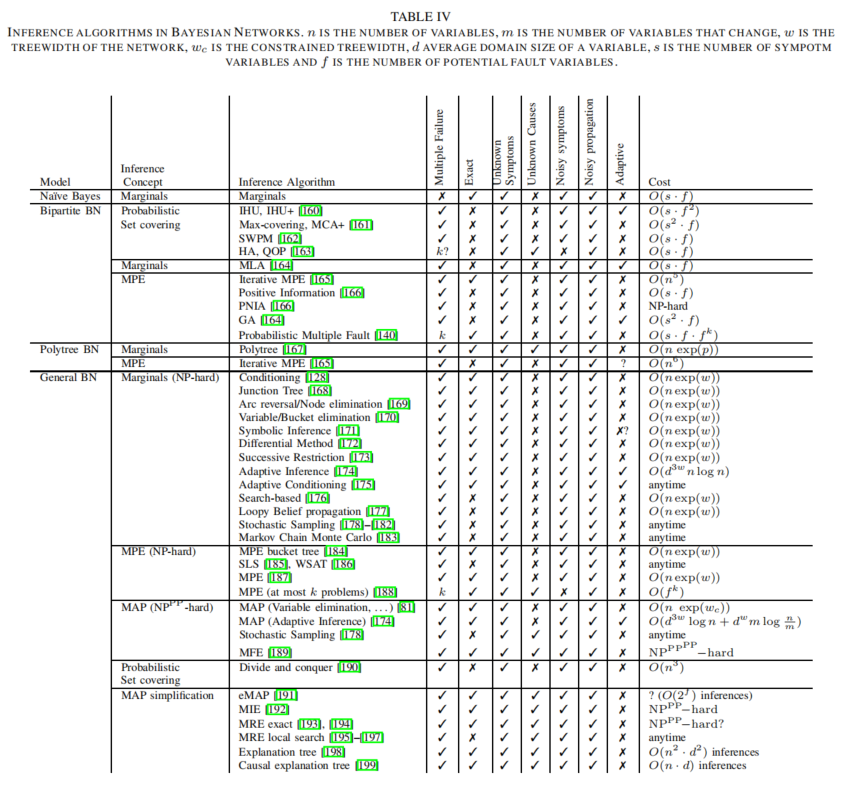
* + - * 1. Very Fast Decision Tree

CVFDT

* 1. 非贝叶斯RCA模型



* 1. 贝叶斯网络



1. RCA的学习模型
   1. 其中最快的学习算法是针对非过程模型的学习规则（决策树或直接规则集）
      1. 贪婪方法：MinEntropy
   2. 基于流数据包括突变检测的自适应决策支持树，该方法假设流数据非常大从而可视为连续的,但不能实时读取数据
      1. 基于Very Fast Decision Tree的CVFDT
      2. 基于Hoeffding Tree的Adaptive Hoeffding
   3. 基于分类的RCA模型，训练复杂度差异很大
      1. NB，SVM，NN
   4. Bayesian Networks是最受欢迎的根因分析模型，但学习BN模型是NP-Complete问题
      1. LGL，从局部因果生成全局因果
   5. 对于evolution rate较高的诊断系统，须允许增量式学习，并不适用分类算法
      1. DT的ID5R
      2. BN的\chi^2检验
2. RCA模型的推断
   1. 基于解释性
      1. 基于规则的模型
      2. 贝叶斯网络：考虑了所有与观测相容的世界所有可能的聚合，但真实世界只能是其中一个状态。
      3. Most Probable Explanation: 只输出最可能的相容世界，即模型中所有变量的最高概率的指派，在模型symptom值全部已知且仅包含causes和symptom时很有意义。
      4. Maximum A Posteriori:在选择最有可能的变量之前，一些变量可以被抽象出来，从而聚合了一些世界。这通常是在中间变量不代表潜在causes或没有相关信息的symptom的情况下实现的，因此只考虑所有潜在原因的集合。
      5. Most Reasonable Explanation/Most Inforbable Explanation: 根据一些最佳标准，它们都试图找到哪个变量子集最适合用户来解释观察结果，而不是用户必须像在MAP中那样预先提供该集合。
   2. 选择标准
      1. 在边际变量、MEP或MAP之间的特定选择取决于是否必须抽象化与故障无关的变量，以及可能同时发生的故障的数量。Marigin 适用于单故障检测，MPE和MAP适用于多故障检测。如果变量仅仅代表不同的可能配置选项，并且没有一个是错误的，后者将尝试自动归纳哪些变量对观察到的symptom有最大的影响。