

文章编号: 1005—8893 (2001) 04—0042—03

数据挖掘技术在故障诊断中的应用^{*}

王洪元¹, 史国栋¹, 符彦惟¹, 夏德深²

(1. 江苏石油化工学院 计算机科学与工程系, 江苏 常州 213016; 2 南京理工大学 计算机系, 江苏 南京 210094)

摘要: 数据挖掘技术是当今计算机信息技术的一项新兴技术, 它综合运用了人工智能、计算智能、模式识别、数理统计等先进技术, 从大量数据中挖掘和发现价值和隐含的知识。研究了数据挖掘的主要方法和它在故障诊断中的应用。

关键词: 数据挖掘; 故障诊断; 人工神经网络; 遗传算法

中图分类号: TP 301.6

文献标识码: B

随着信息技术的高速发展, 数据库应用的规格、范围和深度不断扩大, 传统的依靠数据库的查询检索机制和统计学方法已经远远不能满足现实需要, 它迫切要求自动地和智能地将待处理的数据转化为有用的信息和知识, 从而为决策服务。数据挖掘正是为迎合这种需要而产生并迅速发展起来的用于开发信息资源的一种新的数据处理技术^[1~4]。

一种较为认可的定义是: 数据挖掘就是从大型数据库的数据中提取人们感兴趣的知识。这些知识是隐含的, 事先未知的潜在有用信息, 提取的知识表示为概念, 规则, 规律, 模式等形式。这种定义把数据挖掘的对象定义为数据库, 而更广义的说法是: 数据挖掘是在一些事实或观察数据的集合中寻找模式的决策支持过程, 数据挖掘的对象不仅是数据库, 也可以是文件系统, 或其它任何组织在一起的数据集合。

数据挖掘技术是当今智能系统理论和技术的重要研究内容, 它综合了人工智能, 计算智能(人工神经网络, 遗传算法), 模式识别, 数理统计等先进技术, 它已用于工业, 商业, 金融, 医学, 行政管理等领域。如模糊控制器的建模, 故障诊断的建模, 金融数据的预测等。本文将介绍数据挖掘技术的主要方法以及在故障诊断中的应用。

1 数据挖掘技术的主要方法^[1~4]

从不同的角度看, 数据挖掘技术有多种分类方法, 如根据发现的知识的种类, 挖掘的数据库类型, 挖掘方法, 挖掘的途径等进行分类。目前, 常用的数据挖掘技术方法包括:

(1) 决策树方法。利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的字段, 建立决策树的一个结点, 再根据字段的不同取值建立树的分支; 在每个分支集中重复建立树的下层结点和分支的过程, 即可建立决策树, 从而可从中生成分类规则, 并利用规则和决策树生成复杂知识结构的方法。

(2) 神经网络方法。神经网络方法是基于生物神经系统的结构和功能而建立起来的模拟人脑神经元方法, 以MP模型和HEBB学习规则为基础, 可以建立三大类神经网络模型: ①前馈式网络。它以感知机, 反向传播模型, 函数型网络为代表, 可用于预测, 模式识别等方面。②反馈式网络。它以Hopfield的离散模型和连续模型为代表, 分别用于联想记忆和优化计算。③自组织网络。它以ART模型, Koholon模型为代表, 用于聚类。基于神经网络的数据分类通常具有较小的分类误差和对噪声

* 收稿日期: 2001—10—22

基金项目: 江苏省高校自然科学基金资助(01KJB520004)

作者简介: 王洪元(1960—), 男, 江苏常熟人, 高级工程师, 在职博士生, 感兴趣方向: 模式识别与智能控制。

较强的鲁棒性, 其应用过程的关键性问题为: 网络的构建和训练, 根据属性的数目和类型的树目可以确定相应的网络输入输出模式, 并形成合适网络结构; 网络删减, 在不影响分类误差的前提下, 去除多余的网络节点和链接。经过删减的网络可以提供简洁和有意义的分类规则; 规则提取, 即从经过删减的网络中提取分类规则。

(3) 粗集 (Rough Set) 方法。在数据库中, 将行元素看成对象, 列元素看成属性 (分为条件属性和决策属性)。等价关系 R 定义为不同对象在某几个 (或几个) 属性上取值相同, 这些满足等价关系的对象组成的集合称为等价关系 R 的等价类。条件属性上的等价类 E 与决策属性上的等价类之间有 3 种情况: 下近似: Y 包含 E ; 上近似: Y 和 E 的交非空; 无关: Y 和 E 的交为空。对下近似建立确定性规则, 对上近似建立不确定性规则 (含可信度), 对无关情况不存在规则。

(4) 遗传算法。这是模拟生物进化过程的算法, 由 3 个基本算子组成: ①繁殖 (选择)。是从 1 个旧种群 (父代) 选出生命力强的个体, 产生新种群 (后代) 的过程。②交叉 (重组)。选择两个不同个体 (染色体) 的部分 (基因) 进行交换, 形成新个体。③变异 (突变)。对某些个体的某些基因进行变异 (1 变 0, 0 变 1)。这种遗传算法可以起到产生优良后代的作用。这些后代需满足适应度值, 经过基于若干代的遗传, 将得到满足要求的后代 (问题的解)。遗传算法已在优化计算和分类机器学习方面显示了明显的优势。

2 数据挖掘技术用于故障诊断的建模

故障诊断的建模是故障诊断的关键技术之一。被诊断对象的知识模型由对象运行模型, 诊断经验规则, 诊断神经网络模型, 诊断案例 4 种方式表示, 它们按诊断知识由一般到特殊的次序分三层来组织。第一层知识描述了最一般的诊断知识, 对象运行模型描述了正常工作形态, 用于预测对象的正常输出值以确定异常特征, 供诊断或真值维护和解释。第二层知识描述了一般的诊断知识, 用于基于规则的诊断, 诊断规则由 IF ... THEN ... 形式表达, 其中规则前件表示被诊断对象的症状, 规则后件表示被诊断对象的故障诊断结果。第三层知识由根据相似诊断事例经训练构造而成的人工神经网络模型, 它描述了类似于这些诊断事例的一般的诊断

知识, 用于神经网络模型的诊断。第四层知识由基于二叉树的案例库组成, 它描述了关于个别的特殊诊断案例的诊断知识, 用于基于案例的诊断。

2.1 故障诊断知识模型的自学习建模

首先对于故障诊断数据库按被诊断对象以及被诊断对象的故障诊断结果进行整理排序。对于各诊断对象的各故障诊断结果分别进行知识模型的自学习建模。建模策略是: 首先采用粗集方法从数据集中学习症状与诊断结果之间的诊断规则。然后对该数据集进行测试, 凡能被所挖掘的诊断规则解释的数据样本从数据集中删除, 这表明被删除的大量数据所隐含的故障诊断知识可由这些诊断规则来表达。将该数据集作为训练样本集, 用于构造诊断神经网络模型; 神经网络模型经训练学习构造好后, 对该数据集进行测试, 凡诊断结果误差小于某阈值的数据样本从该数据集中删除, 这表明被删除的大量数据所隐含的故障诊断知识可由该诊断神经网络模型来表达。而该数据集中未被删除的少量数据, 经用户交互检验去掉其中的数据噪声后存放在案例库中。这些案例表达了不同于一般诊断知识模型的特殊故障的诊断知识, 这些案例根据其异常特征进行二叉树组织以便于案例检索时进行基于案例的故障诊断。

2.2 基于粗集的诊断规则挖掘

基于粗集的规则挖掘方法的基本思路是: 首先进行知识约减 (即计算数据对象中不可省略的核心知识部分), 然后根据知识约减挖掘规则。约减的启发式搜索策略为: 首先计算判别集合 A , 把属性在 A 中出现的次数与属性权重的乘积作为启发函数。在属性删减时, 先删除启发函数值小的属性以保证所得的约减具有少的属性个数和大的权重。

2.3 基于人工神经网络的故障诊断建模

由诊断对象的故障诊断事例集经训练构成故障诊断模型, 模型的形式取决于模型的结构和其学习规则, 它们可根据诊断对象的情况来选择。以三层前馈网络和 BP 学习算法神经网络模型为例, 它是一个具有输入层、隐含层和输出层的网络。输入层节点表示训练事例中诊断对象的症状描述, 节点数与症状数相同。隐含层节点数由学习算法自动选择。输出层节点表示训练事例中诊断结果的描述, 节点个数与描述诊断结果的元素相同。训练学习过

程中, 根据误差反向传播算法以及输出层节点实际输出与期望输出的差异来调整输出层和隐含层的权值, 直至累积误差小于设定的值为止。训练完成后, 就能根据诊断对象的实际症状, 应用已建立的诊断模型计算出对应的诊断结果。

2.4 基于混合知识模型的故障诊断推理^[5,6]

当被诊断对象的故障诊断知识模型经自主学习建立以后, 可运用该模型对象的混合知识模型(运行模型, 规则, 神经网络模型, 案例)对该诊断对象进行故障诊断推理, 其策略为: ①首先运用表示大范围故障诊断的知识模型诊断推理出粗的故障范围, 然后再运用表示小范围故障诊断的知识模型进一步诊断出细的故障范围。②依据故障诊断知识模型由一般到特殊的次序进行诊断推理。首先根据被诊断对象的运行模型确定症状, 匹配相应的故障诊断规则进行诊断推理。若无故障诊断的前提与该症状相匹配, 则运用诊断神经网络模型进行故障诊断。若神经网络模型的诊断结果的精度低于设定阈值, 则根据被诊断对象的症状匹配案例库中的诊断案例进行基于案例的故障诊断。

3 结束语

数据挖掘技术是近年出现的综合了人工智能,

计算智能, 模式识别, 数理统计等先进技术的一门新技术, 本文对其主要方法作了介绍, 并对其在故障诊断方面的应用作了探讨。可以预计, 随着数据挖掘技术的进一步发展和完善, 其在许多领域都有着广泛的应用价值和理论研究前景。

参考文献:

- [1] Han J, Kamber M. 数据挖掘—概念与技术(影印版)[M]. 北京: 高等教育出版社, 2001.
- [2] Alahakoon D, Halgamuge S K, Srinivasan B. Dynamic Self Organizing Maps with Controlled Growth for Knowledge Discovery [J]. IEEE Trans Neural Networks, 2000, 11 (3): 601—613.
- [3] Ziarko W. Rough Sets, Fuzzy Sets and Knowledge Discovery [M]. New York: Springer-Verlag, 1994.
- [4] Lu H, Setiono R, Liu H. Effective Data Mining Using Neural Networks [J]. IEEE Trans on Knowledge and Data Engineering, 1996, 8 (6): 957—961.
- [5] Shin C K, Yu S J, Yun U T, et al. A Hybrid Approach of Neural Network and Memory Based Learning to Data Mining [J]. IEEE Trans Neural Networks, 2000, 11 (3): 637—646.
- [6] Feret M P, Glasgow J L. Combining Case-Based and Model-Based Reasoning for the Diagnosis of Complex Devices [J]. J Applied Intelligence, 1997 (1): 30—36.

Data Mining Technique and its Application in Fault Diagnosis

WANG Hong-yuan¹, SHI Gou-dong¹, FU Yan-wei¹, XIA De-sheng²

(1. Department of Computer Science and Engineering, Jiangsu Institute of Petrochemical Technology, Changzhou 213016, China; 2. Department of Computer, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Data mining, a rising technique nowadays, combines many advanced techniques, such as artificial intelligence, computational intelligence, pattern recognition, statistics etc., to mine and discover valuable and hidden knowledge from databases. The primary methods in data mining and its application in fault diagnosis are presented in this paper.

Key words: data mining; fault diagnosis; artificial neural network; genetic algorithm