

# 本科生毕业设计中期检查汇报

基于工业大数据的生产设备故障诊断

题 目：

|  |  |
| --- | --- |
| 院 系 | 机械科学与工程学院 |
| 专业班级 | 机械1401班 |
| 姓 名 | 张照博 |
| 学 号 | U201410606 |
| 指导教师 | 金海 吴波 |

2018 年 4 月

**一、课题研究进度安排**

表1 课题研究进度安排表

|  |  |  |
| --- | --- | --- |
| 学期 | 周次 | 工作任务 |
| 2017-2018  第一学期 | 18周——19周 | 接受任务，翻译参考文献，完成开题报告，对课题有初步掌握，完成开题答辩 |
| 20周——21周 | 查询资料，完成文献综述等任务 |
| 2017-2018  第二学期 | 1 周——3 周 | 完成课程设计之余，搜集资料 |
| 5 周——6 周 | 接受学院检查进度，完成总体方案设计 |
| 7周——8周 | 完成方案1，基于决策树的模型构建 |
| 9周——10周 | 完成方案2，基于支持向量机的模型 |
| 第11周 | 撰写毕业论文，完善资料 |
| 第12周 | 完善论文，并且进行论文查重 |
| 13周——15周 | 论文答辩，并且评定成绩 |

**二、毕业设计内容概述**

**三、毕业设计已完成部分**

**1. 预期达到的目标**

1.1 获取足量数据、实现基于数据驱动的故障模型的建立

1.2 能基于故障时的异常数据完成对故障的推理与诊断

1.3 能够建立故障数据数据库，不断丰富故障模型

1.4 人机交互接口，提供生产人员与故障模型的交互界面

**2. 关键理论和技术**

**2.1 数据挖掘**

对采集到的数据进行清理，挖掘，形成有价值的知识，赋予相对应的故障信息，使得最后形成可以被人理解的相关信息，以此为基础构建故障模型；

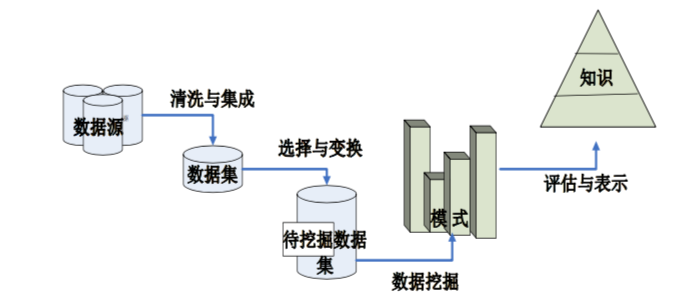


图1 数据挖掘的基本过程

数据挖掘常采用的算法有很多，下面介绍一下我准备采用的一种数据挖掘算法---决策树算法：

决策树算法是数据挖掘中的一个常用的方法，是一种利用树形规则对数据集进行分类的过程算法。通俗讲就是由样本数据集生成树状决策模型并用于分类的数据挖掘算法，建立一种完全基于数据驱动的故障树模型。

决策树能够较为简单直白的描述出多个对象的描述属性与对象最终的分类间的关系。决策树中的每个节点表示一个对象属性,每个节点的分枝路径则代表的该对象不同的属性值对数据集的划分。叶节点为路径所对应的最终分类。一般情况下决策树仅有单一决策结果输出，若需要多个输出，则可以通过建立相互独立但存在嵌套关系的一系列的决策树来获得多个输出的处理能力。

每棵决策树都是一种通过样本数据建立起来的树状决策模型,并通过其分支来对实际数据中的数据元组或对象按照他们的属性进行分类。决策树可以依靠对样本数据集的划分进行节点的选择和树结构的构建，并且在构建过程中可以递归的对树进行“剪枝”处理，直到不能再对类继续划分，或只有单独的类为止。

**2.2故障数据的处理**

一个设备具有很多的参数，如风机，它的参数可能包括电机电流、电机线圈温度、轴承温度、振动值、进出口介质温度和流量等，这些参数间是有复杂的关联关系的。当我们构建知识库的过程中，必须要通过计算参数间的关联度这种手段剔除一些对设备运行状态影响不大的测点，从而提高整体的诊断精度水平。下面是两种常用的关联度计算算法：

2.2.1.Apriori算法：使用候选项集找频繁项集

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。然后使用第1步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项，这里采用的是中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集，使用了递推的方法。

可能产生大量的候选集,以及可能需要重复扫描数据库，是Apriori算法的两大缺点。

2.2.2. FP-树频集算法

针对Apriori算法的固有缺陷，J. Han等提出了不产生候选挖掘频繁项集的方法：FP-树频集算法。采用分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息，随后再将FP-tree分化成一些条件库，每个库和一个长度为1的频集相关，然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法,使得一个FP-tree可以放入主存中。实验表明，FP-growth对不同长度的规则都有很好的适应性，同时在效率上较之Apriori算法有巨大的提高。

**2.3设备运行数据的获取**

工业大数据需要海量的生产设备历史数据和实时运行数据，这些都需要通过一定的数据采集手段才能得到，这也是本课题的一个重要问题，即如何获取足量的数据来训练模型，使其达到理想的性能与精度。

**3. 主要研究内容**

3.1研究数据挖掘算法，采集数据进行模型构建，并且根据新的数据进行模型改进、重构；

3.2 研究关联度计算算法，对故障数据进行适当处理，提高最后得到的故障模型的精度；

3.3 研究机器设备的属性之间的联系，有效的剔除一些对模型精度无益的内容，提高运行效率；

3.4 研究模型的改进方案，如决策树中的“剪枝”方法，对模型进行精简，减少模型构建所消耗的资源。

**4. 完成课题的方案**

**4. 1 构建故障树（基于决策树）**

采用决策树算法，结合大量历史数据，构建一个故障树模型。决策树模型是一种树形结构，在构建过程中需要用到两个很重要的概念：

4.1.1 信息熵

假设当前样本集D中有N个样本，而整个样本有k个分类，每个分类对应的样本数量为那么对于每个分类，他们各自占据的信息量为：

则此样本总体的信息熵为

4.1.2 信息增益

假设当前样本集D中有N 个样本，每个样本都有一些属性，假设我们目前取属性A作为我们计算信息增益的属性。

根据属性A，我们可以属性A的不同取值（假设有v种），将整个样本集D分为v个子样本集，每个样本子集的样本数为N，那么每一个样本子集的频率为：

那么该样本集的A属性的信息增益即为：

下面是利用这两个概念获得故障数的过程：

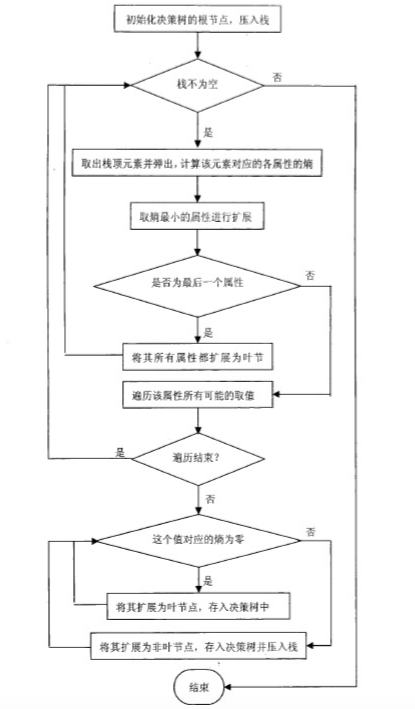


图2 ID3算法生成决策树流程图

4.2 SVM分类方案

基于支持向量机（SVM）的故障诊断方法将诊断问题看成样本分类的问题，即根据历史数据训练出分类器。将数据空间划分成不同的区域，每个区域对应一种运行状态，然后将测试数据投影至数据空间。通过定位其所在区域，推测出测试数据对应的运行状态。下图是一个三分类样例。

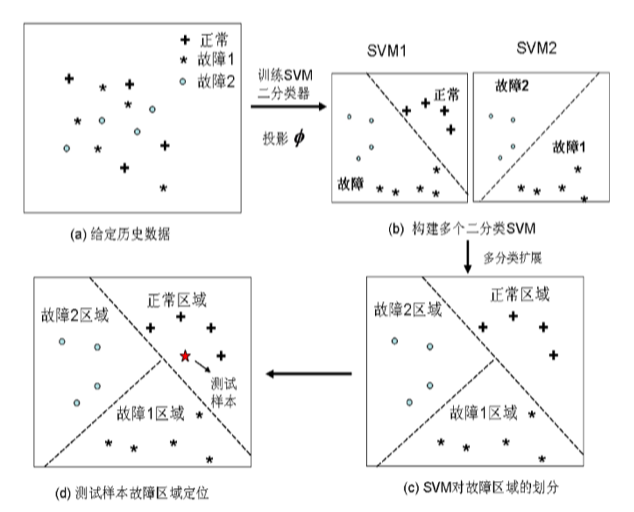


图3 三分类分类样例

在图中，针对目标系统，该方法首先采集各种运行状况下的数据。构建出训练样本集: 然后对训练样本集进行数据预处理，包括去量纲化、特征选择等。并采用SVM对处理过的数据进行学习，生成整个二分类器。见图(b); 由于故障诊断、尤其是故障隔离通常需要面对多分类问题，所以还需要用特定的多分类扩展策略。将多个: 二分类器组合成一个整分类器，使得SVM能够区分多个故障区域，见图(c) 最后，当故障区域被有效划分时。我们只需要对测试样本进行投影确定该样本所属区域，即可实现故障隔离。

SVM的成功运用到实际主要得益于两大技术: 其一，依据SRM准则设计间距最大的分类超平面: 在高维空间里计算出线性最佳分类面。其二，根据核函数媒介得出输入空间里非线性学习算法。关于核函数的方法是当前比较活跃的研究领域。核函数的方式就是通过非线性变换将非线性空间里数据样本集映射至高维线性空间里，在高维空间里导求线性学习算法，要是各个坐标分量之间的相互影响仅仅局限于内积时，那就不用知晓具体非线性变化地形式，只需要把满足Merce条件的核函数代替线性算法里的内积，即可得出原先输入空间里的非线性算法心调。

常用到符合Mercer规定地核函数包括多项式函数、径向基函数以及Sigmoid函数。选取一个核函数就会得出一个SVM。**四、后期时间安排**

**五、近期参考文献**

[1] 盛博, 邓超, 熊尧等. 基于图论的数控机床故障诊断方法[J]. 计算机集成制造系统, 2015, 06: 1559-1570.

[2] 李晗, 萧德云. 基于数据驱动的故障诊断方法综述[J]. 控制与决策, 2011, 26(1): 1-9+16.

[3] 刘强, 柴天佑, 秦泗钊. 基于数据和知识的工业过程监视及故障诊断综述[J]. 控制与决策, 2010, 25(6): 801-807+813.

[4] Zhang, Liangwei. Big Data Analytics for Fault Detection and its Application in Maintenance, 2016

[5] Jay Lee, Hung-An Kao, Shanhu Yang. Service innovation and smart analytics gor Industry 4.0 and big data environment[J]. Percedia CTRP, 2014, 16:3-8.

[6] 邳文君,宫秀军.基于Hadoop架构的数据驱动的SVM并行增量学习算法[J].计算机应用,2016,36(11):3044-3049.

[7] 赵华,苏东,乔文生.TBM主变速箱的状态监测与故障诊断[J].建筑机械化,2003(06):44-45+43.

[8] 徐牧. 基于SVM的变压器故障诊断研究[D].安徽理工大学,2017

[9] 罗雨滋,付兴宏.数据挖掘ID3决策树分类算法及其改进算法[J].计算机系统应用,2013,22(10):136-138+187.

[10] 张媛.采用数据挖掘技术中ID3决策树算法分析学生成绩[J].科技信息,2009(06):537.

[11] 张睿. ID3决策树算法分析与改进[D].兰州大学,2010.

[12] 钟福磊. 工业大数据环境下的混合故障诊断模型研究[D].西安电子科技大学,2015.

[13]朱霄珣. 基于支持向量机的旋转机械故障诊断与预测方法研究[D].华北电力大学,2013.

[14]易辉. 基于支持向量机的故障诊断及应用研究[D].南京航空航天大学,2011.

[15]王振华,杜宇波.基于ESMD和SVM的滚动轴承故障诊断[J].现代制造技术与装备,2018(01):122+124.