附件 1

密级:（涉密论文填写密级，公开论文不填写）

硕士学位论文

基于随机森林算法的故障诊断研究

作者姓名：张晓丹

指导教师：焦建彬　教授　中国科学院大学

学位类别：工程硕士

学科专业：物流工程

研究所　：中国科学院大学 工程管理与信息技术学院

二零一四年四月

附件 2

Research of Fault Diagnosis Based on Random Forest

By

Xiaodan Zhang

A Dissertation Submitted to
University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Master of Logisticis Engineering

College of Engineering & Information Technology

April， 2014

中国科学院大学直属院系

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学直属院系

学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘　要

现代社会中人类从物质生活到文化艺术、娱乐等精神生活都离不开化工产品。随着科学技术的发展，化工生产过程正在不断地朝着智能化与复杂化方向发展，在带来了更高的生产效率和经济利益的同时，其安全性也显得越来越重要。复杂的自动化生产过程和化工产品的特殊性决定了这类系统一旦发生故障，给人类的生命、环境和经济带来的都将是灾难性破坏。及时高效的故障诊断方法或技术能给化工生产过程提供一定的保障，最大程度的减小损失，这对于整个人类社会来说无疑是一种重大贡献。

基于数据驱动的故障诊断是当前这一领域的研究热点，本文以田纳西-伊斯曼化学品公司仿真数据为研究对象，通过随机森林算法的数据分类研究来达到故障诊断的目的。本文的主要研究内容和贡献如下：

1、研究了随机森林算法在本文采用数据上的高效性。详细研究了集成算法、决策树和随机森林算法的来源、算法原理和优缺点，并通过实验验证了随机森林在运行速度、分类准确率、过拟合程度、噪声容忍性等方面的优越性能。

2、提出了一种提取数据动态特征的方法。通过对数据的横向和纵向可视化研究，分析本文所研究数据的时序特征，然后采用窗口遍历的形式，提取动态方差特征和均值特征，有效的加大了不同故障数据之间的差异性，极大的提升了后期采用算法进行分类的准确率。

3、本文提出的方法在 TEP 数据上得出了最优分类结果。结合本文提出的数据动态特征提取方法和随机森林算法，通过实验进行参数调整和算法验证，与其它分类算法（C4.5、CART、SVM、Adboost）相比，本文提出的方法和算法得出了最优分类结果，直接增大了故障诊断的可靠性。

关键词：故障诊断，田纳西-伊斯曼过程 ， 随机森林

Abstract

- **Xiaodan Zhang** (Logisticis Engineering)

Directed by: **Jianbin Jiao** (Professor)

People cannot live without chemical products any more in modern society. With the huge breakthrough of technology and science, chemical production process is becoming increasingly intelligent and complex. The complicacy and particularity of automatic chemical process would cause devastating damages to human life, environment and economy once this kind of system breaks down. Efficient fault diagnosis technique can minimize the losses to a great extent, which means a significant devotion for the whole human society.

Data driven based fault diagnosis is the focus of current research in this field. Simulated data of Tennessee Eastman process is used in this thesis. Random Forest algorithm is adopted to accomplish the purpose of data discrimination and fault diagnosis. The main contents and dedications of this thesis is described as follows:

Firstly, the efficiency of Random Forest algorithm is proved. Detailed information of ensemble algorithm, Decision Tree and Random Forest is explored and compared. It is verified that Random Forest algorithm exceeds others in running speed, discrimination accuracy, the degree of overfitting, and the tolerance of noises.

Secondly, a type of dynamic feature extraction method is proposed. The time-serial feature of TEP datasets is found via visualization analysis. The difference of different data types is augmented through extracting dynamic variance and mean features of the datasets, which promotes the discrimination accuracy dramatically.

Finally, the method proposed in this thesis accomplishes the state-of-the-art accuracy. Through combining the dynamic feature extraction method and Random Forest algorithm, our method completes the highest performance, which increases the reliability of fault diagnosis directly.

KEY WORDS：Fault Diagnosis, Tennessee Eastman process, Random Forest

目　录

图目录

表目录

第一章 绪论

1.1 课题背景和研究意义

化工产品在人类现代生活中占据着越来越重要的地位，从日常的衣、食、住、行到高层次的航空航天研究等方方面面，都需要化工产品为之服务，化工工业也已经成为国民经济发展中的重要支柱产业。而随着工业自动化理论、计算机技术的迅速发展与重大突破，化工生产过程也越来越智能化、大规模化与复杂化。这些变化一方面显著的提高了化工生产效率，满足了人类生活方方面面的需求；另一方面，化工生产过程的安全性也显得越来越重要，复杂的自动化生产过程意味着其发生故障或者失效的潜在可能性也越来越大，而这类系统一旦发生故障，极易导致生产中断、爆炸、毒气泄露等，不仅会严重威胁到人身安全，还会对生态环境造成不可逆转的破坏。

国内外因化工系统故障而引起的重大灾难几乎没有停止过：1986 年 4 月 26 日前苏联切尔诺贝利核工厂发生重大事故，该电站第 4 发电机组爆炸，核反应堆全部炸毁，大量放射性物质泄漏，成为核电时代以来最大的事故。辐射危害严重，导致事故后前 3 个月内有 31 人死亡，之后 15 年内有 6-8 万人死亡，13.4 万人遭受各种程度的辐射疾病折磨，方圆 30 公里地区的 11.5 万多民众被迫疏散[1]；2008 年 9 月 17 日 15 时 35 分左右，云南南磷集团寻甸磷电有限公司液氯充装车间发生氯气泄漏事故，发生泄漏的是液氯充装车间充装装置液氯液下泵零部件，一个直径为 25 毫米的球阀垫圈因气体压力过大，导致被损坏，氯气泄漏出来，导致厂区 71 名工人出现中毒反应；2010 年 1 月 7 日 17 时 30 分左右，中国石油天然气集团公司兰州石化公司 303 厂 316 烃类罐区一裂解碳四储罐阀门处突然发生泄漏，现场可燃气体浓度达到极限，在当班操作人员紧急处理时发生爆炸，爆炸事故造成了 6 人遇难，1 人重伤，5 人轻伤[2]。

这些大大小小的事故给了我们沉痛的教训，化工工业生产过程的安全性已成为社会发展中一个亟待解决的问题。美国、西欧等发达国家已经越来越重视这个问题，近年来已投入大量的人力和物力，加强对该领域的资助，以期望通过严格控制生产设备、传感测量设备的制造工艺，以及充分利用生产数据，为提高产品质量和故障诊断提供有用信息，从而为化工生产过程提供保障。

本论文受到国家重大基础研究计划“973”项目“事故致灾过程和事故致因理论”课题资助，该课题以探索危险化学品事故的超量与触发因素以及事故致因分析为目标，旨在提出更加快速有效的化工过程系统故障诊断方法，及时的发现系统中存在的异常，对事故进行预警。

1.2 故障诊断国内外研究现状

所谓故障是指被定义为系统中至少一个特性或参数出现较大偏差，超出了可接受的范围[3]。工业系统中发生的故障类型包括过程参数的变化、干扰参数的变化、执行器故障和传感器故障等[4]。系统故障诊断是对系统运行状态和异常情况作出判断，并根据诊断作出判断为系统故障恢复提供依据。目前故障诊断研究主要集中于故障检测与诊断两方面。国际故障诊断领域的理论权威 P.M.Frank 教授在 1990 年将故障诊断方法分为三类：基于解析模型的方法、基于信号处理的方法和基于知识的方法[5]。2003 年，Venkatasubramania 将故障诊断方法分为基于数学模型的方法、基于知识的方法和基于数据驱动的方法[6-7]。本文将基于此进行故障诊断研究，有关故障诊断方法分类如图 1-2 所示。

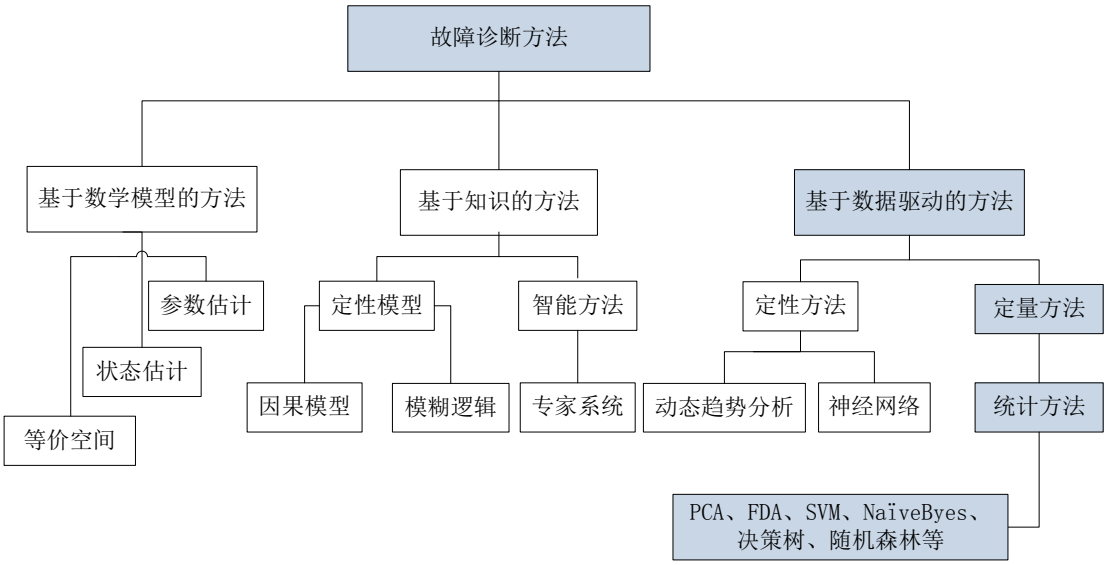


图 1 - 1 故障诊断方法分类

1.2.1. 基于定量数学模型的方法

基于数学模型的方法通常利用控制理论领域的研究成果，如参数估计、状态估计和等价空间等方法。它的机理是建立过程的数学模型，重构过程的参数和状态，然后与相对应的可测信息进行比较产生残差，对残差进行分析和处理进而实现故障诊断的技术。异常情况的发生会引起过程残差发生变化（正常情况下,残差等于零或近似为零),从而就能够检测并诊断过程中的对应异常信息。

在应用这类方法时，必须知道过程的异常反应与模型参数之间的关系，即该工业过程可以建立比较精确的数学模型。但是对于复杂的工业生产过程来说，准确详细的数学模型往往很难得到，即使能够得到，这些理论上的等式也只能

描述系统中一小部分的变量特征，这就限制了该类方法的应用。

1.2.2. 基于知识的方法

该方法主要是利用人工智能的方法，如模糊逻辑、因果分析和专家系统等，模仿人类的思维和行为，构造自动判别系统以完成故障诊断。基于知识的方法不需要对工业过程的精确的数学模型，但需要有大量生产实践经验和故障信息知识储备，当故障发生时，系统行为自动与知识库中的信息想比对，推理出故障模式。这样的知识推理模式比较简单，但是这些知识的获取和故障规则的建立是非常困难的，并且当推理规则比较多时其匹配过程比价费时。另外当系统过程发生知识库中没有的故障时，基于知识的方法由于缺少对应的知识储备而无法判别系统状况。

1.2.3. 基于数据驱动的方法

随着传感设备、计算机技术和数据库技术的飞速发展，在化工工业过程中大量的过程数据很容易被采集并存储下来。但是，这些数据可能包含几百或上千个测量变量和控制变量，工厂操作人员很难同时对这些信息进行人工监控及处理。怎样利用现有的丰富的工业生产过程数据，通过合适的方法或算法变为有用的信息，使之服务于生产安全和故障诊断，已经得到了工业界及学术界的高度重视以及广泛的研究。

基于数据驱动的方法以采集的过程数据为基础，通过各种数据处理与分析方法挖掘出数据中隐含的信息，从而进行智能化故障诊断。这类方法通常采用机器学习、模式识别及统计学理论及算法对过程数据进行分类、聚类、特征提取等，以达到故障检测和故障诊断的目的。其中经典的方法如主成分分析法（Principal Components Analysis, PCA）、Fisher 判别分析（Fisher Discriminant Analysis, FDA）、贝叶斯、神经网络（neural network）、支持向量机（support vector machine, SVM）、决策树、随机森林（Random Forests, RF） 等已经得到了众多机构和学者的研究和重视。Kano [8]以主元子空间的差异来检测和区分故障的产生。Chiang 等[9]提出了一种 FDA 和 SVM 相结合的故障诊断方法。该方法首先通过 FDA 降维方法进行特征提取，剔除了不相关特征，然后通过特征扩展将系统的动态特性加入其中，最后采用 SVM 分类器进行故障诊断。Zhang 等[10]介绍了基于模糊神经网络的在线故障诊断方法并将该方法成功的应用于罐式搅拌器。Li 等[11]介绍了基于 SVM 的在线故障诊断方法，并将其应用于化学流程工业当中。

1.3 本文的研究内容

本文采用基于数据驱动的故障诊断方法，以田纳西-伊斯曼化学品公司仿真数据为基础，对其进行故障诊断研究。在故障类型和数据一一对应的基础上，通过数据分类研究达到故障诊断的目的。本文对分类算法的研究主要为决策树和随机森林算法，通过原理分析、理论证明、特征提取及实验验证，使得本文提出的模型达到了最优分类效果。具体研究内容如下：

提出了一种提取数据动态特征的方法。通过对数据的横向和纵向分析，得出了本文所研究数据具有特殊的时间序列特征，通过提取动态方差信息和移动平均法对数据进行特征提取，有效的加大了不同故障数据之间的差异性，极大的提升了后期采用算法进行分类的效果。

研究了基于决策树的集成算法随机森林。详细研究其原理和改进理论，并通过实验验证了其在算法原理、运行速度、分类准确率、过拟合程度、噪声容忍性等方面的优越性能。

结合本文提出的数据动态特征提取方法和随机森林算法，通过实验进行参数调整，与其它分类算法（C4.5、CART、SVM、Adboost）相比，得出了最优分类效果。

1.4 本文的组织结构

第一章，绪论。主要论述故障诊断的研究背景和意义，分析国内外的研究现状以及发展趋势，最后说明了本文的主要研究内容和组织结构。

第二章， 随机森林算法理论基础。主要论述集成算法、决策树算法和随机森林算法原理。

第三章，基于随机森林的故障诊断。首先介绍 TEP 数据背景、数据特征及其物理含义；然后，采用决策树算法和随机森林算法对数据进行分类分析。

第四章，基于动态特征提取方法的故障诊断。提出一种提取数据动态特征的方法，并用决策树算法和随机森林算法分别对处理后数据进行分类分析。最后再对多种算法实验结果进行对比分析，体现出我们提出的方法的优越性。

最后是结束语，总结本文的主要工作，展望了未来工作的方向，以及对如何进一步提高故障诊断的准确性进行了探讨。

第二章 随机森林算法基础研究

随机森林[11](Random Forest)是 Leo Breiman 于 2001 年提出的分类和预测模型，是一种基于决策树的集成算法。而在此之前，已经有相关概念的产生。

1994 年，Amit 和 Geman [12,13]首先提出了随机树的概念，采用随机生成节点测试训练决策树，然后将决策树算法集成起来，采用测量概率均值来作为集成树的输出，并用于手写体识别中。1995 年，Tin Kam Ho[14]提出了随机决策森林（Random Decision Forests）的概念，基于树的分类器模型，通过引入多棵树，通过对特征空间的选择，以互补的方式构造不同的树，后来又基于特征空间的研究，提出了一种 Random Subspace 方法[15]。取得了理想的实验效果。

2001 年，Leo Breiman [16]对 Random Forest 进行了明确的概念定义，通过数学推导公式证明了随机森林不易过拟合的理论，从而奠定了随机森林理论研究的基石。

在介绍随机森林算法之前，本章将首先介绍随机森林算法的两个重要组成部分：集成算法与决策树算法。

2.1 集成算法

在机器学习算法的实际应用中，很多单分类器的模型精度往往达不到理想的效果。因此，越来越多的研究希望通过集成的方法来提高模型精度，尝试将各分类器进行集成整合，优势互补，规避劣势。这些方法被称为分类器集成方法(Ensemble method)。该类方法旨在构建一组基分类器，通过一定的方式组合到一起，最后进行联合投票获取最终的预测值。这种集成分类器算法不仅可以提高学习的泛化能力，而且可以有效提高分类器的鲁棒性，能够解决许多单分类器无法解决或难以取得理想结果的问题，因而在诸多领域得到了广泛的关注与应用，随机森林即属于这类集成学习方法的一个典型。

集成方法的基分类器可以相同，也可以不同。目前的研究多数集中在对相同基分类器的研究上。另外 Breiman 指出，当选择的学习算法是不稳定的算法时，通过集成得到的集成分类器分类性能才会有显著的提高。因此具有不稳定性的神经网络和决策树算法常被选为集成学习中弱学习算法[16]。本文研究的随机森林算法即属于以决策树算法为弱分类器的集成算法。

当前，最流行的集成学习方法是 Bagging 和 Boosting，下面将详细介绍这两种方法。

2.1.1. Bagging

Bagging[17]是 Breiman 在 1996 年提出的，Bagging 即 Bootstrap aggregating，是一种利用 Bootstrap 抽样方法构造训练样本集，用不同的 Bootstrap 抽样样本集基于某种算法分别训练 t 个不同的模型，最后投票决定类别的集合算法。aggregating 是聚集的意思，代表的是最后决策时的投票法则。

Bootstrap 本来是用于估计统计量的一种重采样方法，通过重采样，构造不同的样本集，通过计算不同样本集的统计量（如方差），从而可得到统计量的估计。该方法在统计学上经常与 Jackknife（在一批样本点中，每次删除一个或者几个样本点，用剩下的样本和同样的估计量公式去重新计算估计值）相比较。

在 Bagging 中，采用自助抽样法（Bootstrap Sampling）进行样本集构建，对于一个含有 n 个样本的训练样本集 D，给定一个基元学习算法，采用以下步骤：

1、采用 Bootstrap 抽样方法，从原始训练样本集 D 中有放回地随机抽取 n 次，得到和原样本集容量相同的样本集 Bootstrap 1；

2、重复步骤（1）t 次，得到 t 个新的样本集 Bootstrap 1，Bootstrap 2，…，Bootstrap t；

3、用这 t 个新的样本集分别训练 t 个算法模型 $(h_1, h_2, \cdots, h_t)$；

对于一个新的测试样本 x，将其分别放入 t 个算法模型，最后按投票的方式决定其类别。

对于以上步骤（1）中 Bootstrap 抽样方法，会导致某些初始样本多次被抽到或者没有被抽到。这样一方面保证 Bootstrap 抽样样本集各不相同，可以用来训练不同的模型，另一方面导致某些初始样本不在其中，D 中每个样本最终未被抽取的概率为 $(1-1/n)^n$，其中 n 为原始样本集 D 中样本的个数。当 n 足够大时，$(1-1/n)^n \approx 0.368$，这表明原始样本集 D 中接近 36.8%的样本不会出现在 bootstrap 抽样样本集中。

Bagging 方法通过随机抽取的方式构造不同的训练样本集，从而增加了分类器之间的差异，进而提高组合分类器的泛化能力。其中各分类器组合采用的是并行组合方式，拓扑结构如下所示：

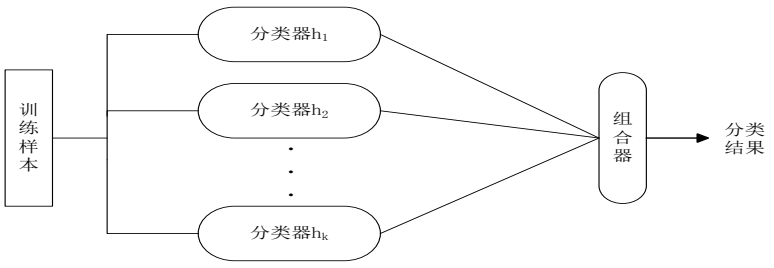


图 2 - 1 分类器组合并行拓扑结构

2.1.2. Boosting

Boosting[18]方法是 Schapire 于 1990 年提出的：首先每个样本都被赋予一个相同的初始权重，第一轮训练后，对正确分类的样本降低权重，对错误分类的样本加大权重，然后利用调整过权重的样本集进行第二轮的训练，依次迭代 t 次得到 t 个模型。这 t 个模型即 t 个分类器，其权重取决于各自的表现，最后综合投票确定最终分类结果。

Boosting 中各分类器组合采用的是串行组合方式，其拓扑结构图如图 2-2 所示：

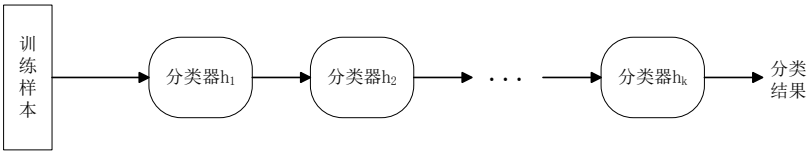


图 2 - 2 分类器组合串行拓扑结构

基于 Boosting 的典型算法是 Adaboost，该算法由 Yoav Freund 和 Robert Schapire[49]提出于 1995 年。 Boosting 方法能够增强学习器的泛化能力，但是该方法对噪声点敏感，其串行结构也导致其运算速度较慢。

2.1.3. Bagging 和 Boosting 的异同

Bagging 和 Boosting 组织方式上的区别：

1、 Bagging 采取 Bootstrap 的方式进行样本集抽样选取，Boosting 初始样本集为所有原始数据；Bagging 是各个分类器独立训练的，Boosting 是迭代训练的（当前分类器模型是在上一个分类器基础上建立的）；

2、 Bagging 没有权重，最后投票公平竞争，Boosting 每轮迭代的训练样本是有权重并不断调整的，最后形成的分类器也是有权重的；

Bagging 和 Boosting 区别的数学区别：

1、 Bagging 这种组织方式偏重于降低方差（variance），通过训练样本重采样并独立训练模型，考虑了大部分样本的准确性，牺牲掉部分噪声点的利益，这样构建的模型就比较聚拢，结果会相对稳定，但是可能会有偏差；

2、 Boosting 偏重于降低偏差（bias），尽可能的减小训练数据的偏差，这样在存在噪声点的情况下会造成过拟合。一般而言，在训练样本没有噪声或噪声很小的情况下，Boosting 的准确率要高于 Bagging，但是当训练样本噪声比较大时，Boosting 会出现严重的退化，而 Bagging 几乎不受影响。

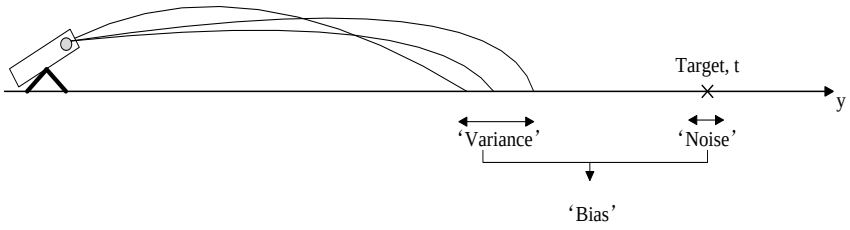


图 2 - 3 偏差与方差效果示意图

Breiman 也从偏差和方差的角度对 Bagging 的泛化误差进行了分析[13]。他指出，Bagging 对不稳定的基元算法能显著提高预测的准确率，这是由于不稳定算法的偏差较小，方差较大，Bagging 通过减小方差可以很好的减小泛化误差。决策树是这一类型不稳定基元算法的典型，其与 Bagging 方法结合能有效减小方差，从而减小集成分类器的泛化误差。

2.2 决策树算法

决策树方法的起源是概念学习系统(Concept Learning System，CLS)，1966 年，Hunt[19]等人提出了第一个构造决策树的算法 CLS，它采用自顶向下的递归方式，在决策树的内部结点进行属性值的比较，叶子结点是要学习的类别，后来的许多决策树算法都可以看作是对 CLS 算法的改进与更新。Quinlan 于 1986 年提出了 ID3[20](Iterative Diehotomize)算法，以信息熵作为选取分裂属性的度量标准。1993 年，Quiulan 又提出了 C4.5 算法[21]，主要改进是使用信息增益率替代信息增益来选择属性，并且可以直接处理连续型属性。

分类和回归树（CART, Classification and Regression Trees）[22]算法是由 Breiman 等于 1984 年提出的一种经典决策树算法，不仅可以用于传统的分类，还可以用作回归。CART 算法生成的决策树是二叉树，在每个内部结点处选择具有最小 Gini 指标值的属性作为分裂属性。

决策树是一个树状模型，训练时树的根结点是整个数据集合空间，从根结点开始自上而下进行结点分裂，在当前结点对所有属性进行测试，选取最优属性将数据集合空间分割成 2 块或更多块，然后对子结点递归调用以上方法，最后构建出决策树模型，每个叶结点是带有分类标签的数据分割。

决策树模型如图 2-4 所示。

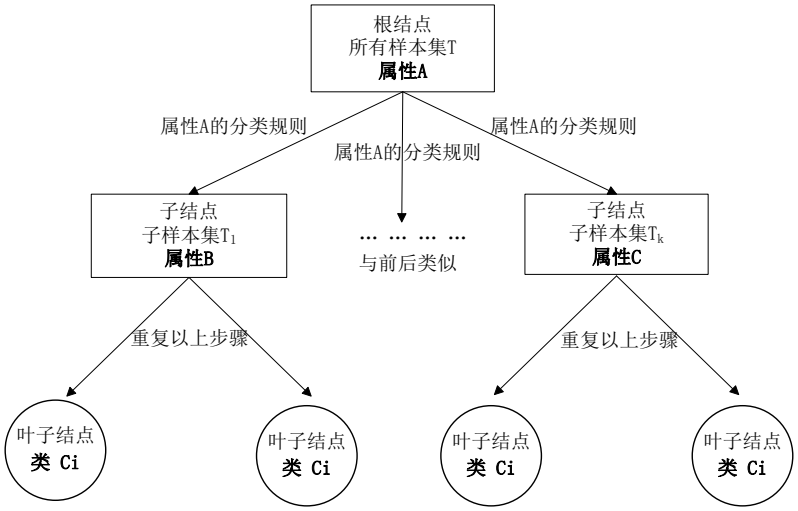


图 2 - 4 决策树模型

决策树每个结点表示一个属性，其每个分支表示按照该属性进行分类的规则，叶子结点表示最终结果。如何选取属性是 ID3、C4.5、CART 算法的主要不同之处，以下我们将着重介绍一下 C4.5 算法和 CART 算法。

2.2.1. C4.5 算法

由于 C4.5 算法是 ID3 算法的后继改进，其原理和计算过程都非常相似，我们先讨论一下 ID3 算法属性选择过程。

1、计算样本集 *T* 的经验熵。设训练样本集 *T* 的类标号属性为 *Ci* (i= 1,…,n)，则 *T* 的经验熵如公式（2-1）所示：

$$Entrop(T) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2-1}$$

其中， p_i 为属于类 *Ci* 的样本在 *T* 中所占的比例。

2、计算各个属性变量的经验条件熵。设属性 *A* 有 *K* 个不同的取值{*A1, A2,…, Ak*}，将数据集 *T* 划分为{T1, T2,…, Tk}，则 属性 *A* 对应的经验条件熵为：

$$Entrop_A(T) = \sum_{j=1}^{k} \frac{|T_j|}{|T|} Entrop(T_j) \tag{2-2}$$

其中，$|T_j|/|T|$为第 *j* 个划分的权重。

3、计算属性的信息增益。属性 *A* 对 *T* 划分得到的信息增益为：

$$Gain(A) = Entrop(T) - Entrop_A(T) \tag{2-3}$$

4、计算所有属性的信息增益，信息增益最高的属性选作当前结点的分裂属性。

5、按被选择属性的类别对当前结点样本进行划分，形成两个或多个子结点，然后对子结点进行以上步骤的递归调用，最后得到整棵树模型。

C4.5 算法在原理上类似 ID3 算法，唯一的不同是 C4.5 算法利用信息增益比代替信息增益来选择属性：

$$\begin{cases} GainRatio(A) = \dfrac{Gain(A)}{Split\ \inf(A)} \\ Split\ \inf(A) = -\sum_{i=1}^{k} \dfrac{|T_i|}{|T|} \times \log_2(\dfrac{|T_i|}{|T|}) \end{cases} \tag{2-4}$$

其中，*Gain(A)*是前面的信息增益，*Split inf (A)*表示拆分信息，实际上就是将每个样本视为等可能情形下的熵，克服了用信息增益选择属性时偏向选择取值多的属性的不足。

和 ID3 算法相比，C4.5 算法在效率上有了很大的提高。不仅可以直接处理连续型属性，还允许训练样本集中的样本出现属性空缺，生成的决策树的分枝也较少。但是，C4.5 算法在选择属性、分割样本集上所采用的技术仍然没有脱离信息熵原理，因此生成的决策树仍然是多叉树。

2.2.2. CART 算法

CART 算法利用基尼指数（*Gini index*）作为尺度来选择属性。同时决定该属性的最优二值切分点。在某一给定结点，CART 算法选择结点属性的步骤如下：

1、计算不纯度。首先，针对某个属性 *A*，训练样本集 *T* 的类标号属性为 *Ci*(i=1,…,n)，属性 *A* 有 *K* 个不同的取值{*A1, A2,⋯, Ak*}。 *T* 的不纯度如公式（2-5）所示

$$Gini(T) = 1 - \sum_{i=1}^{n} P_i^2 \tag{2-5}$$

其中，P_i 为属于类 *Ci* 的样本在 *T* 中所占的比例。

2、计算属性 *A* 的基尼指数。基尼指数在此用来计算属性划分样本集的差异程度。差异程度越小，划分的效果越好。假设属性 *A* 的一个二元划分将 *T* 分成了 *T1*，*T2* 两部分，则对应于该划分的 *Gini* 指标如公式（2-6）所示：

$$Gini_A(T) = \frac{|T_1|}{|T|} Gini(T_1) + \frac{|T_2|}{|T|} Gini(T_2) \tag{2-6}$$

其中，*T1*，*T2* 表示 *A* 对 *T* 的一个二元分裂。

由于属性 *A* 有 *K* 个不同的取值，故有 *2K-2* 种划分，需要计算 *2K-2* 次基尼指数。然后将基尼指数最小的二元划分作为该属性的分裂方法。

3、重复 1、2 步骤，计算所有属性的基尼指数，最后基尼指数最小的属性作为该结点的分裂属性。该属性和它的分裂子集（离散值属性）或分裂点（连续值属性）一起形成分裂准则。

CART 的二叉树形式简化了模型的结构，提高了构建模型的速度其可以用于分类和回归，本文研究的随机森林算法采用了 CART 分类器作为基元算法。

2.2.3. 决策树的优缺点

决策树算法自出现以来就受到极大关注和广泛研究，已经成为工业界和学术界频繁应用的典型算法，总结而言，决策树算法有以下几个显著的优点：决策树方法的应用范围非常广泛，样本集数据可以是连续或离散的，包括离散的语义数据，另外部分属性的确实也不影响使用；决策树模型简单直观，容易理解，其非线性判别结构适合处理多类问题，同时计算效率非常高；决策树方法能够有效的抑制训练样本噪音。

但是决策树也具有与生俱来的缺点：自上而下的逐层判别可能会过于适应噪声，从而导致过度拟合的问题。因此如何改进决策树算法，保持其优点的同时克服缺点成为广大学者研究的方向。

2.3 随机森林算法理论

2.3.1. 随机森林的定义

随机森林中最基本的组成元素是决策树。为了产生 *t* 个决策树，需要生成 *t* 个独立同分布的随机向量$\Theta 1,\cdots\Theta t$，使用训练集 *D* 和Θ生成第 *t* 棵树 *h（X，*Θt*）*，其中 *X* 为输入向量。

定义 1 [16] 随机森林是由一组树结构分类器 {*h（X，*Θt*），t=1，...*}组成的分类器，其中Θt 是相互独立且同分布的随机向量。由所有决策树投票决定输入向量 X 的输出。

2.3.2. 随机森林模型

随机森林是一种将 Bagging 方法和决策树算法相结合的集成算法，与传统的决策树相比，有更强的泛化能力和更好的分类效果。一般情况下，随机森林里单棵树用的就是 CART，即树结点采用 *Gini* 分裂方式，

设训练集有 *n* 个样本，每个样本为 *m* 维。利用 Bootstrap 抽样方法建立 *t* 棵决策树模型，随机森林的随机性体现在两个方面（1）自助样本选择的随机性，保证每个决策树构建时所用数据不同；（2）结点划分时特征选择的随机性。

随机森林模型建立如图 2-5 所示：

1、首先，利用 Bootstrap 抽样从原始训练集抽取 *t* 个样本集，分别命名为 Bootstrap 1、Bootstrap 2. . . . . Bootstrap t，且每个样本集的样本容量都与原始训练集一样；

2、其次，对 *t* 个样本集分别建立 *t* 个决策树模型，每棵决策树结点分裂时随机选取 *v*（*v*<*m*）个特征进行信息增益（或信息增益比、基尼指数）计算，在 *v* 个特征中选取最优值作为结点属性；

3、每个结点属性和分叉规则确定，最后得到 t 个决策树模型。

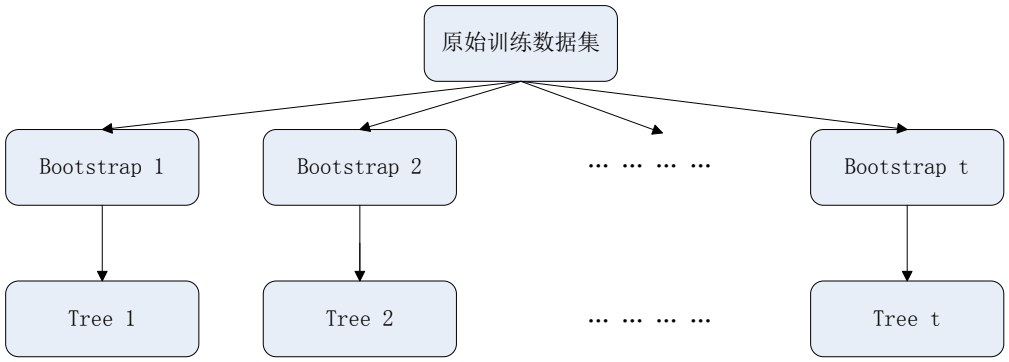


图 2 - 5 随机森林模型

随机森林测试过程如图 2-6 所示：

1、对于一个新的样本 *V*，将其分别输入 *t* 个决策树模型，得到 *t* 种分类结果；

2、按投票的方式决定其最终分类结果。

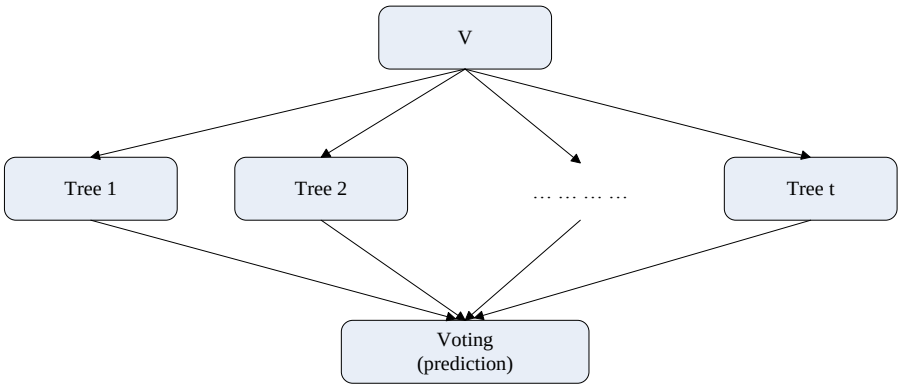


图 2 - 6 随机森林测试过程

2.3.3. 随机森林的理论基础

给定 t 个分类器集合$\{h_1(X),h_2(X),\cdots,h_t(X)\}$，输入向量 X 和输出向量 Y（分类结果），定义边缘函数（Margin Function）：

$$mg(X,Y)=av_tI(h_t(X)=Y)-\max_{j\neq Y}av_tI(h_t(X)=j) \tag{2-7}$$

其中 $I(\cdot)$ 为指示函数，$av_t(\cdot)$ 为取平均值。$mg(X,Y)$ 描述了将输入向量 X 正确分类为 Y 的平均得票数超过将 X 分错为其它任何类的得票数的程度，显然，$mg(X,Y)$ 值越大，则表明分类效果越好，正确分类的置信度（confidence）越高。

定义分类器的泛化误差：

$$PE^*=P_{X,Y}(mg(X,Y)<0) \tag{2-8}$$

其中下标 X，Y 表明了概率的定义空间。

在随机森林中，$h_t(X)=h(X,\Theta_t)$。当森林中树的数目很大时，它服从于大数定律。

定理 1：随着树的数目的增加，对于所有随机向量$\Theta_1,\Theta_2,...,\Theta_t$，$PE^*$ 几乎处处收敛于

$$P_{X,Y}(P_\Theta(h(X,\Theta)=Y)-\max_{j\neq Y}P_\Theta(h(X,\Theta)=j)<0) \tag{2-9}$$

其中Θ 表示随机森林参数向量，$h(X,\Theta)$ 表示分类器输出。

这个结果表示了随着树的增加，泛化误差 PE^* 将趋向某一上界，这也解释了为什么随机森林不会出现过拟合。

定义随机森林的边缘函数：

$$mr(X,Y)=P_\Theta(h(X,\Theta)=Y)-\max_{j\neq Y}P_\Theta(h(X,\Theta)=j) \tag{2-10}$$

同上 $mg(X,Y)$ 的意义，该边缘函数表示正确分类的平均投票率超过错误分类最大投票率的程度。

定义 s 为分类器 $h(X,\Theta)$ 的强度：

$$s=E_{X,Y}mr(X,Y) \tag{2-11}$$

假设 $s\geq 0$，根据切比雪夫不等式可得：

$$PE^* \le \mathrm{var}(mr)/s^2 \tag{2-12}$$

经由一系列的推导可得定理 2：

随机森林的泛化误差上界的为：

$$PE^* \le \bar{\rho}(1-s^2)/s^2 \tag{2-13}$$

其中 ρ 是相关系数的均值，s 是树的分类强度。

该上界并不是那么精确，但是为我们提供了一个对随机森林性能评价的参考依据，即泛化误差只和两个参数有关，ρ 越小，s 越大，则泛化误差越小。

2.3.4. 随机森林的优缺点

随机森林作为决策树基础上的提升算法，具备很多优越性：

1、 适合多分类。随机森林自身就是一种多分类算法，可以一次性对多种类型数据进行建模和分类测试，而 SVM、FDA 等算法是二分类算法，当用于多分类是，需要建立多个一对多模型或者多对多模型，而随机森林算法的建模和测试过程都可以一次性完成，原理和过程都非常简单；

2、 不必担心过度拟合。Breiman 通过大数定理证明了随机森林这一特征，克服了决策树容易过拟合的缺点；

3、 运行速度快。能高效处理大样本数据，并能够并行处理。

4、 适用于数据集中存在大量未知特征。当数据集中存在部分变量值缺失时也不影响随机森林算法的使用；

5、 能够估计哪个特征在分类中更重要。通过随机森林的袋外数据估计可以进行变量重要性测量；

6、 对噪声不敏感。当数据集中存在大量的噪声时同样可以取得很好的预测性能。

这些优点使得随机森林在近些年获得了广泛关注和研究，但同时也发现一些不完善的地方，和不同算法相比优劣势也不相同：

1、 如和单个决策树相比，随机森林速度比较慢，但模型精度和分类准确率要优于决策树；

2、 和 Boosting 相比，随机森林的执行速度要快很多，当样本噪声很大时，随机森林模型优于 Boosting，但当样本噪声很小或可以忽略不计时，随机森林的分类准确率低于 Boosting[23]；

3、 和 SVM 相比，随机森林的速度比较快，但当样本数比较少时，随机森林模型建立过程中可供选择的样本很少，不能很好的发挥 Bagging 的优势，而

SVM 则比较适合小样本数据集。

2.3.5. 随机森林在其它领域应用现状

由于随机森林的众多优点，近 10 年来随机森林的理论和方法在许多领域都有了比较迅速的发展：Pall Oskar Gislason[24]利用随机森林方法对土地的覆盖面积进行了研究，并发现随机森林与其它组合算法相比训练更快；S.L.A.Lee [25] 利用随机森林和聚类建立了一种混合随机森林算法，对助肺 CT 图像进行肺结节的自动检测，同时还在随机森林中加入了聚类方法；Diaz-Uriate R and Andres S A D[26]将随机森林用于对微阵列数据进行分类，并且提出了一种基于随机森林的基因选择分类方法；A Bosch 等[27]将随机森林用于图像分类；Lan Guo[28]将随机森林用于故障预测；Bo-Suk Yang 等[29]研究了随机森林算法应用于机器故障诊断的可能性，并且将遗传算法引入随机森林用来提高分类准确率。A Criminisi 等[30,31]从分类、回归、密度估计、流行学习和半监督学习、动态学习多个方面研究随机森林建立统一模型，并将其用于 Kinect 人体姿态识别中。

2.4 本章小结

本章对随机森林算法进行了详细的介绍。首先，介绍了集成算法的概念和来源，详细介绍了 Bagging 和 Boosting 算法及其异同；然后，介绍了决策树算法的起源与背景，并详细介绍了三种典型的决策树算法 ID3、C4.5 和 CART 的数学原理及优缺点；最后，详细介绍了随机森林算法模型、理论基础、优缺点及其应用现状。

第三章 基于随机森林的故障诊断

3.1 TEP 简介

3.1.1. TEP 背景介绍

田纳西-伊斯曼过程（Tennessee Eastman process，TEP）是由美国 Eastman 化学公司 1993 年创建的一个真实工业过程仿真平台[32]。该过程以实际的复杂工业化工过程为基础，其仿真数据已被众多研究者广泛采用，作为衡量过程监控和故障诊断等研究方法优劣的标准测试集。

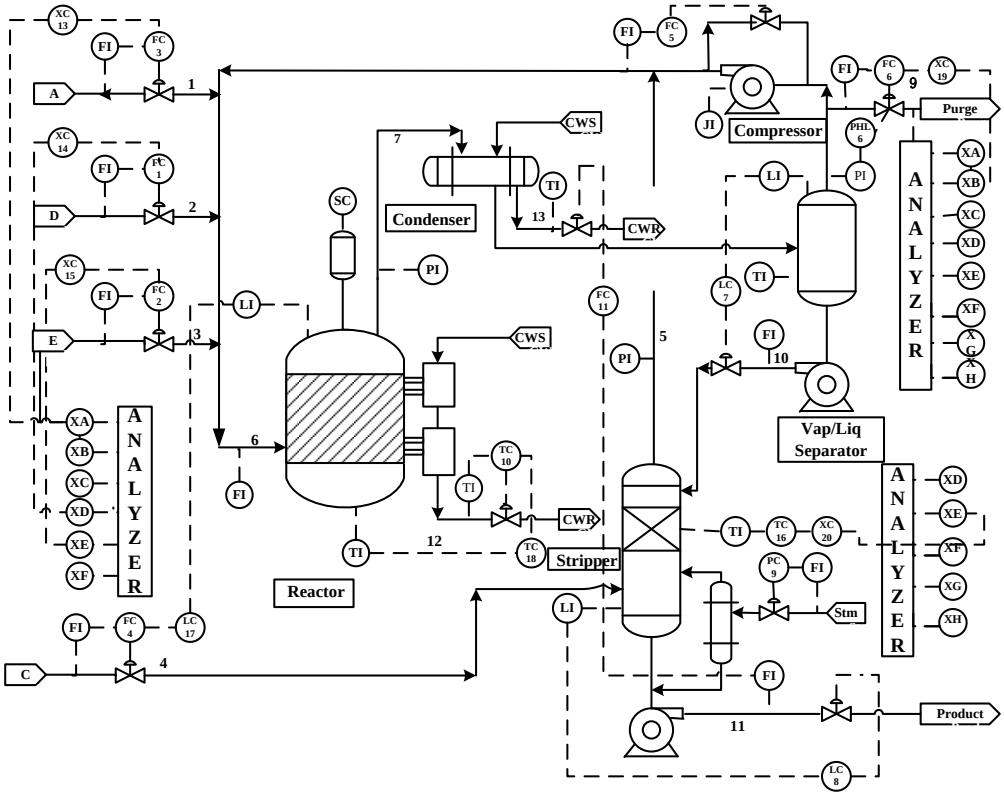


图 3 - 1 TEP 过程工艺流程图

TEP 过程流程图如图 3-1 所示，该过程由五个反应单元组成：反应器（Reactor）、冷凝器(Condenser)、分离器(Separator)、汽提塔(Stripper)和压缩机(Compressor)；共有八种成分：A、B、C、D、E、F、G 和 H，其中 A、C、D、E 为四种气态反应物，B 为催化剂，F 为副产物，G 和 H 为最终产物。在反应器中进行的化学反应如下：

$$A(g)+C(g)+D(g)\rightarrow G(liq)\quad ,\qquad \text{产品 1}$$

$$A(g)+C(g)+E(g)\rightarrow H(liq)\ ,\qquad \text{产品 2}$$

$$A(g)+E(g)\rightarrow F(liq)\ ,\qquad \text{副产品}$$

$$3D(g)\rightarrow 2F(liq)\ ,\qquad \text{副产品}$$

整个过程包含四个输入流，一个产品流和一个净化流。气体成分 A、C 和 E 以及惰性成分 B 被加入反应器，液态产物 G 和 H 在反应器中形成。反应器的产品流通过冷凝器冷却，然后送入到汽/液分离器。从分离器出来的蒸汽通过压缩机再循环送入反应器。为了防止过程中惰性成分和反应副产品的积累，必须排放一部分再循环流。来自分离器的冷凝成分（流 10）被泵入汽提塔。流 4 用于汽提流 10 中的剩余反应物，这些剩余反应物通过流 5 与再循环流结合。从汽提塔底部出来的产品 G 和 H 被送到下游过程。

3.1.2. TEP 数据描述

TEP 过程包括 12 个控制变量和 41 个测量变量，41 个测量变量又包含 19 个成份测量变量和 22 个连续变量，其中 19 个非连续测量成份变量由 3 个成份分析仪获得。

表 3-2 为 TEP 数据的 53 个过程变量描述表，其中 22 个连续测量变量记为 V（1）～V（22），分别表示反应单元的物料流量、温度、液位、压力等连续参数；19 个成分测量变量记为 V（23）～V（41）为值，分别表示不同流的成分含量；12 个控制变量用 U（1）～U（12）来表示，分别为不同进料口及装置内阀门等控制参数。

本文实验中采用的数据集由 MATLAB 仿真程序获得，训练集和测试集中的数据包含了 41 个测量变量和 11 个控制变量（反应器中搅拌器的搅拌速度除外），即每个样本都为 52 维变量构成，采样间隔取 3 分钟。

表 3 - 1 连续测量变量

22 个连续测量变量		19 个成分测量变量		12 个控制变量	
V1	A 物料流量	V23	流 6 成分 A	U1	D 进料量
V2	D 物料流量	V24	流 6 成分 B	U2	E 进料量
V3	E 物料流量	V25	流 6 成分 C	U3	A 进料量
V4	A、C 混合物料流量	V26	流 6 成分 D	U4	总进料量
V5	回收流量	V27	流 6 成分 E	U5	压缩机阀
V6	反应器进料率	V28	流 6 成分 F	U6	排放阀
V7	反应器压力	V29	流 9 成分 A	U7	分离器灌液流量
V8	反应器液位	V30	流 9 成分 B	U8	汽提器产品流量
V9	反应器温度	V31	流 9 成分 C	U9	汽提器水流阀
V10	放空率	V32	流 9 成分 D	U10	反应器冷却水流量
V11	产品分离器温度	V33	流 9 成分 E	U11	冷凝器冷却水流量
V12	产品分离器液位	V34	流 9 成分 F	U12	搅拌速度
V13	产品分离器压力	V35	流 9 成分 G		
V14	产品分离器出口流	V36	流 9 成分 H		
V15	汽提塔液位	V37	流 11 成分 D		
V16	汽提塔压力	V38	流 11 成分 E		
V17	汽提塔出口流量	V39	流 11 成分 F		
V18	汽提塔温度	V40	流 11 成分 G		
V19	汽提塔蒸汽流量	V41	流 11 成分 H		
V20	压缩机工作功率				
V21	反应器冷却水出口				
V22	分离器冷却水出口				

TEP 仿真模型包含 21 个预设定的故障。这些故障中，16 个是已知的，5 个是未知的。故障 1~故障 7 与过程变量的阶跃变化有关，如冷却水的入口温度或者进料成分的变化；故障 8-故障 12 与一些过程变量的可变性增大有关；故障 13 是反应动力学中的缓慢漂移；故障 14、15 和 21 与粘滞阀有关，故障 16~20 是未知的，详细故障描述如表 3-2 所示。

训练集中正常模式数据中有 500 个样本数据，其余 21 类故障模式具有 480 个样本数据；测试集中每种模式的样本数据个数为 960，正常模式所有样本为正常数据，故障模式中前 160 个样本数据为正常数据，后 800 个数据为故障数据，这是因为仿真程序进行仿真时，各种故障是在 8 小时后引入的。

表 3 - 2 TEP 21 类故障描述

编号	故障描述	分类
1	A/C 进料流量比变化，组分 B 含量不变（流 4）	阶跃
2	组分 B 含量发生变化，A/C 进料流量比不变（流 4）	阶跃
3	物料 D 的温度发生变化（流 2）	阶跃
4	反应器冷却水入口温度发生变化	阶跃
5	冷凝器冷却水入口温度发生变化	阶跃
6	物料 A 损失（流 1）	阶跃
7	物料 C 压力损失（流 4）	阶跃
8	物料 A，B，C 的组成发生变化（流 4）	随机变量
9	物料 D 的温度发生变化（流 2）	随机变量
10	物料 C 的温度发生变化（流 2）	随机变量
11	反应器冷却水入口温度发生变化	随机变量
12	冷凝器冷却水入口温度发生变化	随机变量
13	反应动力学特性发生变化	慢漂移
14	反应器冷却水阀门	黏住
15	冷凝器冷却水阀门	黏住
16	未知	未知
17	未知	未知
18	未知	未知
19	未知	未知
20	未知	未知
21	流 4 的阀门固定在稳态位置	恒定位置

3.2 决策树算法分类实验

3.2.1. 实验方法

本文实验采用 21 类数据进行同时分类，这 21 类数据包括一类正常数据和表 3-1 中的前 20 类故障数据，由于第 21 类故障数据特性比较差，在本文实验中暂不予考虑。

我们采用数据挖掘软件 weka 中 C4.5 和 CART 算法同时对 21 类 TEP 数据进行分类。本文实验训练样本为 500+480*20=10100 个（正常数据训练样本为 500 个，其它 20 类故障训练样本为 480 个），测试样本为 800*21=16800 个（每类故障测试样本都为 800 个）。

3.2.2. 实验结果及分析

采用 C4.5 算法和 CART 算法对 21 类故障同时分类，结果如下：

表 3 - 3 决策树分类结果

	分类准确率（%）	
算法	训练模型	测试结果
C4.5	96.8	57.7
CART	89.51	57.99

由表 3-3 可得，C4.5 算法和 CART 算法的测试结果相差不大，但测试准确率要远远低于训练模型准确率，可见 TEP 数据特性和决策树算法的双重作用导致了比较严重的过拟合，即在训练数据集上模型表现良好，但在测试数据集上应用结果很差。

表 3 - 4 采用 CART 算法对 TEP 数据进行测试的混淆矩阵结果

Confusion matrix of test result of original dataset under CART

Actual		Predicted																			
	Total	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Normal	800	128	1	0	143	0	28	0	0	10	146	37	15	9	2	0	134	34	11	13	47	42
Fault 1	800	2	778	0	0	0	0	0	0	3	0	0	0	3	10	0	2	2	0	0	0	0
Fault 2	800	4	0	775	5	0	0	0	0	0	2	1	0	4	0	0	8	0	0	0	1	0
Fault 3	800	129	0	11	137	0	19	0	0	3	116	26	18	0	3	0	94	64	13	21	97	49
Fault 4	800	8	0	0	3	711	1	0	0	0	3	0	61	4	1	1	0	0	0	0	6	1
Fault 5	800	34	1	0	42	0	468	0	0	38	50	16	1	19	15	0	37	14	5	18	6	36
Fault 6	800	0	3	0	0	0	0	794	0	0	0	0	0	0	0	0	0	0	0	1	0	2
Fault 7	800	0	0	0	0	0	0	0	795	0	1	0	0	0	0	0	0	0	0	2	2	0
Fault 8	800	13	9	24	10	0	28	0	2	305	26	63	4	169	46	0	4	24	7	41	10	15
Fault 9	800	142	1	2	156	0	7	0	0	3	166	17	27	3	3	0	119	41	12	11	65	25
Fault 10	800	37	0	2	57	0	17	0	0	64	68	302	7	31	22	0	40	45	6	16	40	46
Fault 11	800	42	0	2	51	47	5	0	0	0	89	20	406	2	0	24	26	10	23	5	43	5
Fault 12	800	8	0	0	16	0	58	0	1	125	14	52	4	356	35	12	4	18	6	73	5	13
Fault 13	800	7	0	10	9	0	45	0	2	119	8	75	3	146	180	4	14	2	14	85	10	67
Fault 14	800	7	0	0	9	0	1	0	0	2	10	0	12	5	0	707	5	3	32	2	4	1
Fault 15	800	142	0	3	118	0	22	0	0	6	166	43	8	10	6	0	134	19	19	30	45	29
Fault 16	800	17	0	5	50	0	9	0	0	15	30	153	1	8	7	0	24	372	9	10	44	46
Fault 17	800	6	0	0	20	0	6	0	0	2	9	1	8	9	0	27	20	15	633	7	8	29
Fault 18	800	9	0	0	21	0	4	0	3	6	16	1	2	31	4	0	11	6	2	660	5	19
Fault 19	800	67	0	1	42	0	5	0	0	2	55	5	10	3	1	0	25	29	6	5	517	27
Fault 20	800	60	0	1	54	0	9	0	0	5	74	18	5	22	6	0	44	23	9	18	34	418

利用 CART 算法对 21 类 TEP 数据同时分类的测试结果如表 3 - 4 所示，Normal 行代表了正常数据总共有 800 个样本，其中分正确的个数为 128 个，错分为第一类故障的个数为 1 个，错分为第三类故障的个数为 143 个，依次类推。不难看出，从左上角到右下角的对角线上的值代表了每一类分类正确的个数，

而正常数据、故障 3、故障 9、故障 13、故障 15 等类型分类效果极差，正是这些故障类型的错分率导致了整体分类准确率的下降。

3.3 随机森林算法分类实验

3.3.1. 实验方法

同 3.2.1 小节，在随机森林算法分类实验中，仍对 21 类数据模式进行同时分类。其中训练样本为 500+480*20=10100 个（正常数据训练样本为 500 个，其它 20 类故障训练样本为 480 个），测试样本为 800*21=16800 个（每类故障训练样本都为 800 个）。

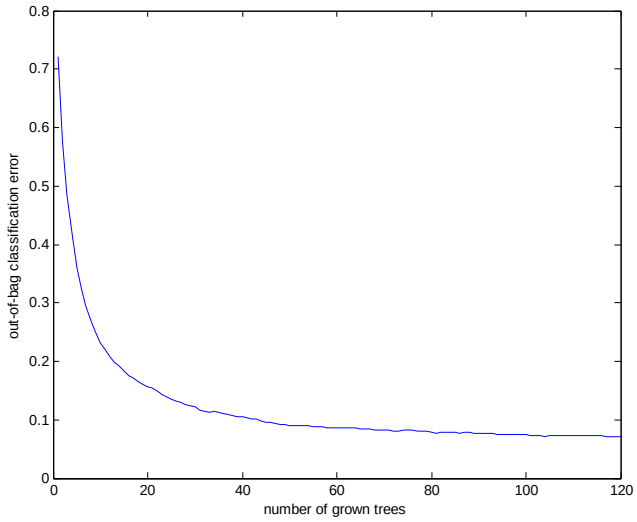


图 3 - 2 令 v=7，t=0~120，随机森林袋外误差趋势图

首先利用随机森林对 TEP 数据进行袋外误差估计，在随机特征选择数固定为某一值的同时，随机森林中树的个数 t 则从 0 递增到 120，结果如图 3-2 所示。

由图 3-2 可以看出，随机森林模型的袋外误差随着随机森林中树的个数 t 的增加而减小，最后趋向于平缓，这一结果也验证了第二章随机森林的理论基础研究结果：随着树的增加，泛化误差 PE^* 将趋向某一上界，这也解释了为什么随机森林不易出现过拟合。

对 TEP 数据进行随机森林模型建和测试。由于随机森林算法有两个参数需要设置，森林中树的个数 t 和随机特征选择数 v。在本实验中，采用自动化测试方法寻找最优参数，设置树的个数 t 从 10-100 以 10 个单位递增，随机特征选择数 v 从 5-30 以每次 5 个单位递增。

3.3.2. 实验结果及分析

采用上述实验方法，测试准确率输出结果如下图表所示，表 3-5 为定量结果，图 3-3 为定性趋势展示。

表 3 - 5 不同参数下随机森林分类结果

Test precision of original dataset under different RF paramaters(%)

number of tree	number of random feature selected					
	5	10	15	20	25	30
10	60.48	61.18	62.46	62.59	61.89	62.01
20	62.19	63.77	63.87	64.19	63.61	63.08
30	63.09	6456	64.78	63.97	64.18	63.81
40	63.12	64.76	64.87	64.39	64.08	64.35
50	63.79	65.11	65.01	64.39	64.41	64.02
60	64.24	65.24	**65.68**	65.11	64.89	64.59
70	64.19	65.25	65.48	64.74	64.62	64.52
80	64.56	65.24	65.49	65.14	64.93	64.57
90	64.71	65.23	65.53	65.19	64.71	64.35
100	64.29	65.49	65.38	65.21	65.16	64.39

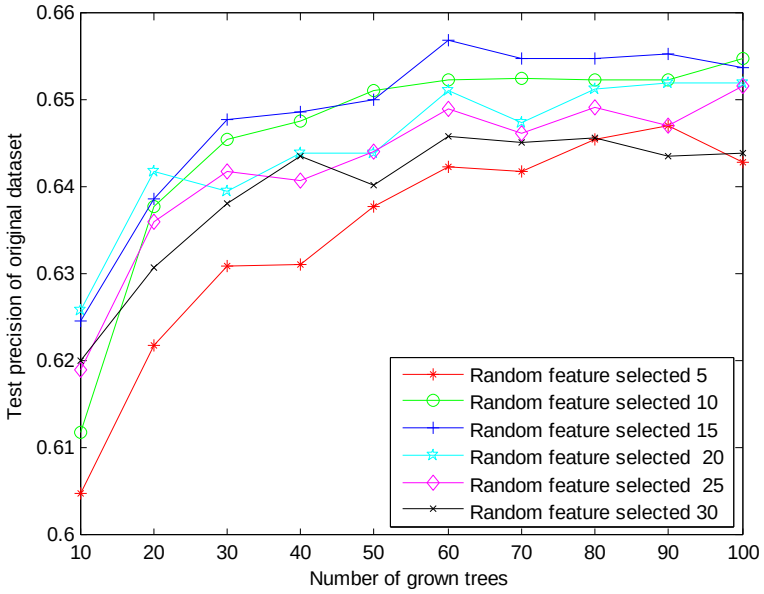


图 3 - 3 不同参数下随机森林分类结果图示

由以上图表可知随着随机森林树的个数的增加，分类准确率基本呈上升趋势，随机特征数的选择则呈波动趋势。在本实验中，当 T=60，F=15 时，分类准确率达到最高值：65.68%。

通过对比，可得随机森林算法优于决策树算法，其分类准确率高出 7%左右。与决策树算法相比，随机森林模型的泛化能力有了显著提高。

3.4 本章小结

本章主要研究了决策树算法（C4.5、CART）和随机森林算法应用于 TEP 数据的分类效果。首先介绍 TEP 过程工艺流程以及数据来源，并对其变量和故

障类型进行详细的数据描述；然后分别采用决策树算法和随机森林算法对 21 类数据进行分类实验，并着重对随机森林参数选取进行试验描述，结果表明随机森林算法比决策树算法具有明显的优越性，分类准确率提高了 7%左右。

第四章 基于动态特征提取方法的故障诊断

为了进一步提高数据分类及故障诊断准确率，本节将通过详细分析 TEP 数据特征，根据其数据表现提取关键特征，以保持相同数据类型间的一致性，增大不同数据类型之间的差异性，达到预处理方法和分类算法相结合达到分类最优的目的。

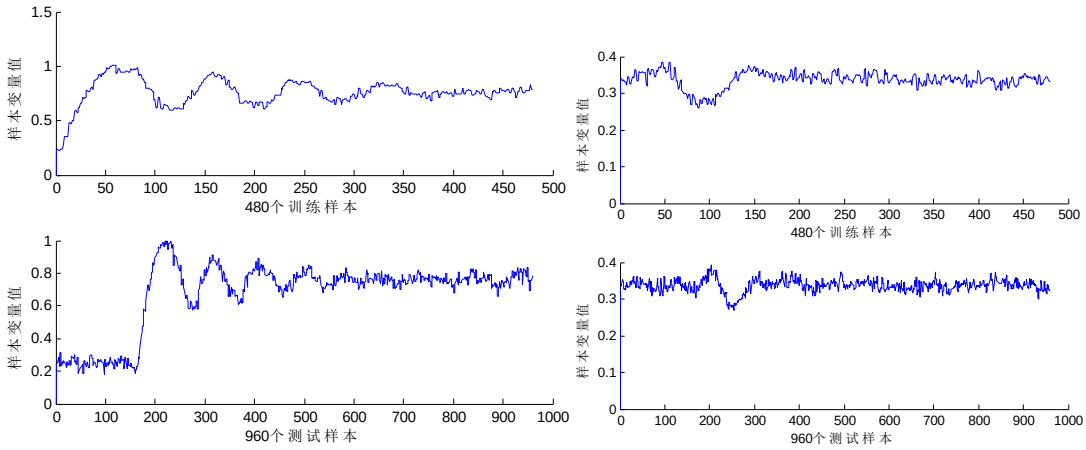
4.1 TEP 数据特征

本节对 TEP 数据进行可视化展示，以发现其数据特征，为提出高效的动态特征提取方法做铺垫，方便后期深入挖掘信息。

4.1.1. 对某类故障某个变量进行训练和测试样本的时序分析

本小节我们将某一故障所有样本（训练样本为 480 个，测试样本为 960 个）按顺序排列作为横坐标，画出针对某一变量的走势图。由于样本变量值的获取是每三分钟抽样所得，故这种按照样本抽样顺序排列的横坐标等价于时间上的一种度量，本文中都将此类型横坐标直接视为时间。

对于训练样本和测试样本的时间走势图，本小节以第一类故障第一个变量和第五类故障第十个变量作为代表展示效果。



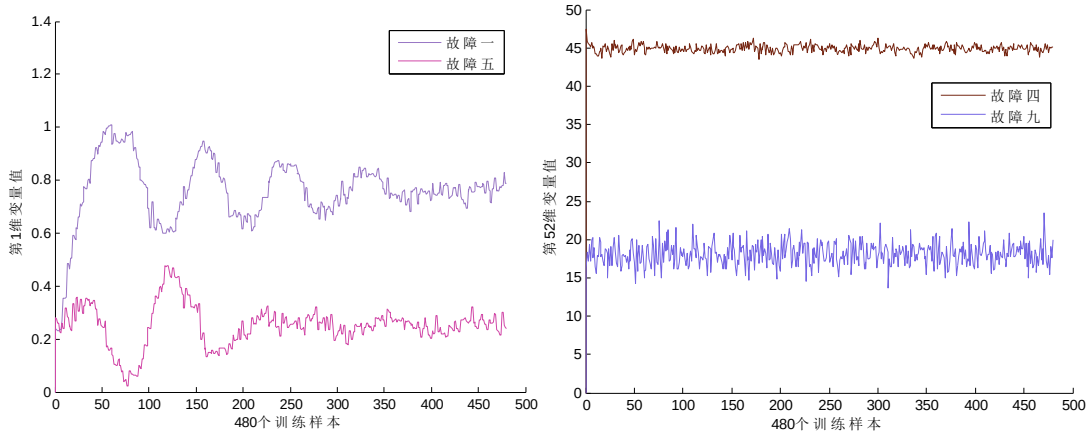
（a） （b）

图 4-1 （a）第一类故障第 1 个变量 （b）第五类故障第 10 个变量

由图 4-1 可知 TEP 数据某一变量值的变化具有很明显的动态信号特征，如第一类故障第一个变量随时间变化趋势类似阻尼振荡。在故障诊断中该时序特征应该被考虑到，而不应该将每个样本数据作为一个单一的数据看待。由于测试样本的前 160 个数据为正常数据，故在本文实验中将其剔除，只取后 800 个数据进行测试，由图 4-1 也可以看出，训练样本和测试样本（后 800 个）的变

化趋势在时间上也是相合相应的。

4.1.2. 对不同故障同一变量进行幅值比较分析



（a） （b）

图 4 - 2 训练样本走势图（a）故障一和故障五的变量 1（b）故障四和故障九第 52 个变量

由图 4-2 可知，不同故障同一变量的值有着明显的不同，这和化工过程故障的物理原因也是一致的：不同故障会造成对某一变量值的变化，或者某一变量值发生异常变化会导致某种故障。而这种变量值的差异性特征正是我们采用数据挖掘算法，尤其是统计学方法进行有监督分类的依据。

根据数据可视化分析，采用一定的方法对数据进行预处理，以提取其动态特征，加大不同类型之间的差异性，对于提高故障分类的准确率将会有重大突破。

4.2 动态特征提取方法介绍

针对上节对 TEP 数据的详细分析，考虑根据其时序特征，对数据进行特征提取，以增大不同类型数据之间的可分性。

我们考虑提取原数据的均值和方差特征代替原数据进行分类，以增大不同故障数据间的差异性；另一方面，考虑用时间窗口的方法提取其均值和方差特征，这样使得数据呈现一定的动态特征，即当前数据能够代表其前面一定时间内数据的特征。

由此我们提出了动态均值与方差处理（Dynamic mean and variance processing，DMVP）的方法：采用移动平均法以窗口 n 提取某一时间点的均值和方差特征，替代原变量用于分类和故障诊断。

移动平均法的计算公式如下：

$$x_t = \frac{1}{n}(x_{t-1} + x_{t-2} + \cdots + x_{t-n}) \tag{4-1}$$

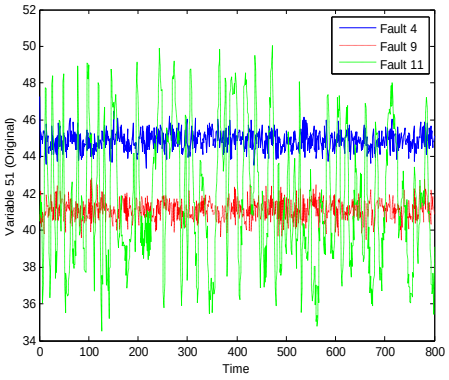
式中，x_t 表示对当前的预测值；x_{t-1} 表示 t-1 时刻的值，则 x_{t-n} 表示 t-n 时刻的值；n 为移动平均的窗口大小。

在 DMVP 中，假设 t 时刻，某一 TEP 样本数据为 $s_t=(v_{t,1},v_{t,2},\cdots,v_{t,52})$，其前 n 个样本依次为 $s_{t-1},s_{t-2},\cdots,s_{t-n}$。当前 52 维样本数据 s_t 经 DMVP 处理后扩展成 104 维：$s_t^{new}=(mean_{t,1},mean_{t,2},\cdots,mean_{t,52},\mathrm{var}_{t,1},\mathrm{var}_{t,2},\cdots,\mathrm{var}_{t,52})$。计算公式如下公式所示：

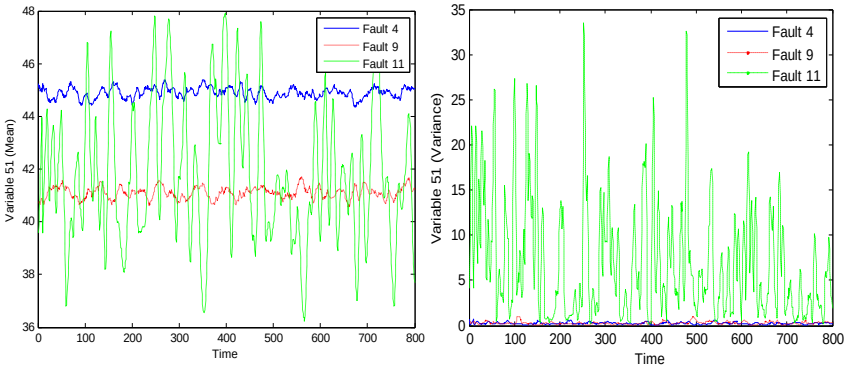
$$mean_{t,i}=\frac{1}{n}(v_{t-1,i}+v_{t-2,i},\cdots v_{t-n,i}) \tag{4-2}$$

$$\mathrm{var}_{t,i}=(v_{t-1,i}-mean_i)^2+(v_{t-2,i}-mean_i)^2+\cdots+(v_{t-n,i}-mean_i)^2 \tag{4-3}$$

其中 $mean_{t,i}$ 和 $\mathrm{var}_{t,i}$ 分别代表该样本经处理后第 i 个变量的均值和方差。为了展示以上数据表达方法的有效性，我们对第四，九，十一类故障进行分类实验。由于数据之间的交叉重叠比较多，这三类故障经常被用以检验方法效果。其中第四类故障代表反应器冷却水入口温度发生变化，第 9 类故障表示物料 D 的温度发生变化（流 2），第十一类故障表示反应器冷却水入口温度发生变化。



(a)



(b) (c)

图 4 - 3 故障 4、9 和 11 的第 51 维变量的(a)原始值(b)动态均值(c)动态方差

由图 4 - 3 可以看出，在第 51 维变量上，故障 11 和另外两种故障类型有明显交叉重叠，而通过动态方差处理，故障 11 很好的分离了出来。

我们采用决策树 C4.5 算法对这三类数据进行分类，以验证 DMVP 方法的有效性。

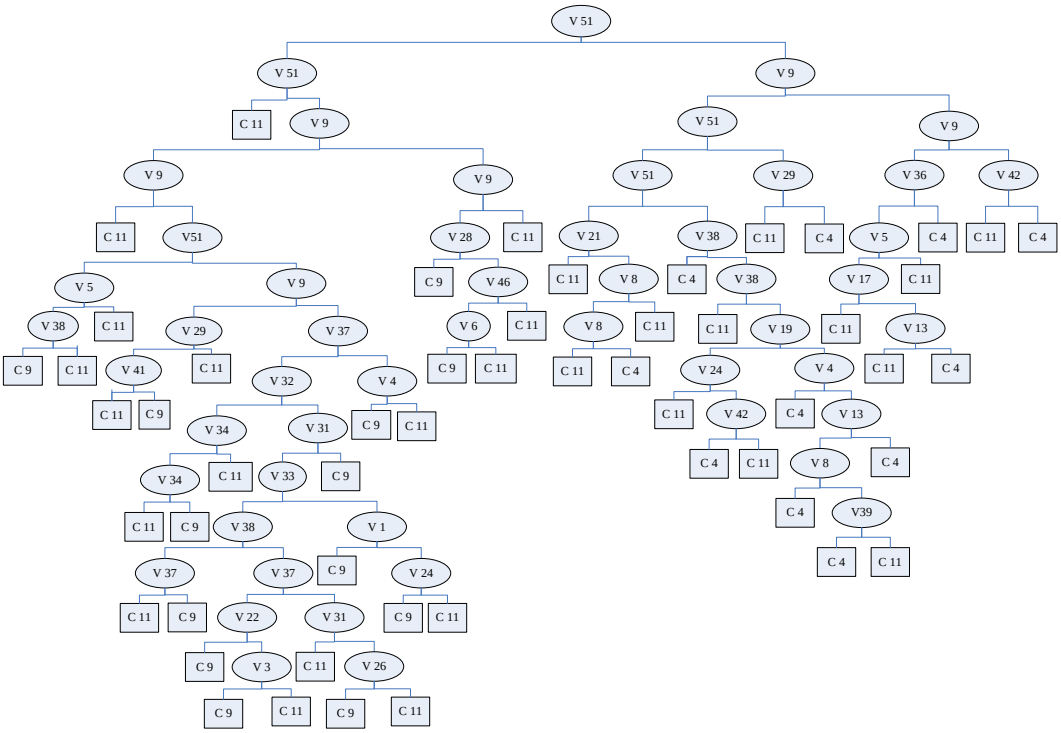


图 4 - 4 原数据的决策树可视化模型

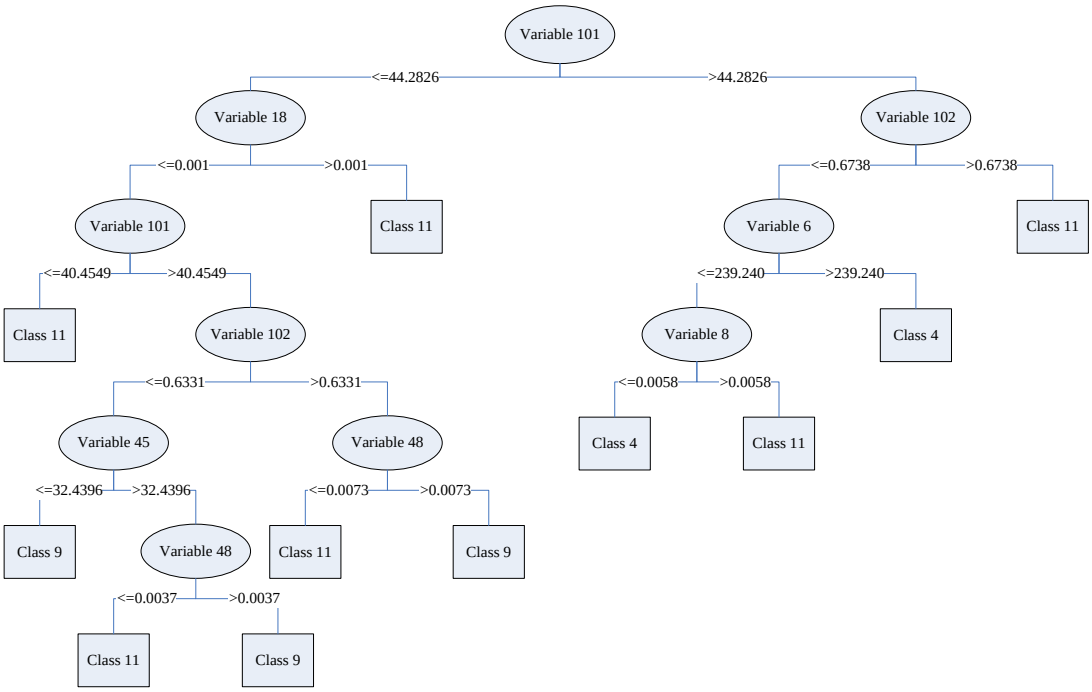


图 4 - 5 经过 DMVP 处理后的数据的决策树可视化模型

上图中根结点和每个内部结点代表某一特定变量，叶子结点代表故障类型。

由图 4-4 和 4-5 可以看出，原数据的决策树模型比较复杂，而经过 DMVP 处理的数据的决策树模型相对简单，深度和广度都大大缩减。其中 Variable 101 和 Variable 102 分别代表第 51 维变量的动态均值和方差信息。从可视化树模型可以看出，第 51 维变量对于四、九、十一三类故障起着极其重要的作用。

表 4 - 1 DMVP 处理前后数据的分类结果

实际类别		预测类别					
	总样本数	IDV(4)		IDV(9)		IDV(11)	
		原数据	DMVP	原数据	DMVP	原数据	DMVP
IDV(4)	800	702	**796**	0	**0**	98	**4**
IDV(9)	800	0	**0**	653	**778**	147	**22**
IDV(11)	800	70	**21**	95	**10**	635	**769**

表 4 – 1 中数据表示将实际类别分为某一类别的样本个数，如 702 表示对于原数据的训练和测试，将第四类故障测试数据分类正确的有 702 个，而将其误分为第九类和第十一类故障的样本数分别为 0 和 98 个;而 778 表示对于经过 DMVP 处理（此处取 w=10）的数据的训练和测试，将第九类故障测试数据分类正确的样本有 778 个，其它 22 个样本都分错到第十一类。

由表可以容易得出，经过 DMVP 处理的数据分类准确率明显提高。更精确的分类准确率结果见表 4-2：

表 4 - 2 测试准确率对比

算法	测试准确率（%）	
	原始数据	DMVP 10
C4.5	82.92	97.63

实验结果证明了 DMVP 方法的高效性，当然图中只是展示了第 51 维变量的特征，更多的变量区分会进一步加大此类重叠数据可分性。对于 DMVP 方法的意义，均值特征的加入是为了让每个样本点具有一定时间内的时序信息，其能够代表某一时间范围内的数据特征，并且能够保留数据原来的特征；加入方差特征，更好的提取了某一时间范围内的数据变化特征，加大了原本均值接近但是方差相差很大的数据之间的差异性，而 DMVP 方法对此种类型数据有极强的表达性能。

4.3 决策树算法实验分析

4.3.1. 实验方法

对 TEP 训练数据和测试数据分别进行 DMVP 处理，在此我们选择窗口数分别为 10、15 和 20，分别用 DMVP 10、DMVP 15、DMVP 20 来代表，结果得到 104 维的训练数据（共 500+480*20=10100 个样本数据）和测试数据集（共 800*21=16800 个样本数据）。然后用 C4.5 和 CART 算法分别对处理后数据集进行分类实验。

4.3.2. 实验结果及分析

对 21 类故障数据同时分类，训练和测试结果如下：

表 4 - 3 针对不同数据的决策树测试准确率

算法	测试准确率（%）			
	原始数据	DMVP10	DMVP15	DMVP20
C4.5	57.70	65.76	67.79	69.45
CART	57.99	68.17	70.70	71.02

由表 4-3 实验结果可知，经 DMVP 处理后的 21 类数据由决策树算法分类结果显示，分类准确率能够提升 10%左右。并且随着 DMVP 处理窗口数的增大，测试准确率也呈上升趋势。

4.4 随机森林算法实验分析

4.4.1. 实验方法

同上节实验方法，对 TEP 训练数据和测试数据分别进行 DMVP 处理，结果得到 104 维的训练数据（共 500+480*20=10100 个样本数据）和测试数据集（共 800*21=16800 个样本数据），然后用随机森林算法进行分类分析。DMVP 窗口选择和随机森林参数选取在下一小节根据实验结果进行描述分析。

4.4.2. 实验结果及分析

对 TEP 数据 21 类故障同时分类，结果如图 4 - 6 所示：

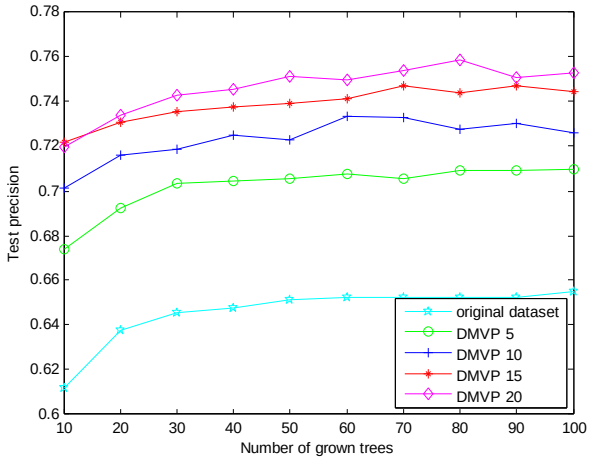


图 4 - 6 不同数据集下随机森林模型测试准确率

图 4-6 中横坐标代表树的个数 t =10~100（步长为 10），不同颜色线条代表不同数据的测试准确率，original dataset 为原始数据，DMVP 5、DMVP 10、DMVP 15、DMVP 20 分别代表 DMVP 处理窗口为 5、10、15、20。

由图 4-6 可得以下结论：

随着树的数目的增加，测试准确率也基本呈上升趋势，但同树的数目越大，随机森林模型的构建和测试也就越慢，故在选择随机森林树的个数 t 时，应该在可接受范围内尽量大。

DMVP 处理窗口长度越大，对应的分类准确率越高，但是同时应该考虑到窗口越大则预处理越耗时，并且窗口过大时，会消除时序数据的差异性，将不太平缓的数据集平滑化，从而导致过滤掉重要特征，极限情况下（w>样本数）则，将样本数据集归一为一个数据，这样对于分类识别与诊断是没有意义的。

综上所述，我们选择 DMVP 窗口长度为 15，在 TEP 化工环境中，每一变量值代表之前 45 分钟内的数据动态趋势，而这一值可根据实际需要更改。我们在此选择 DMVP 15 作为基准，对随机森林算法进行不同参数下的分类实验，结果如下图表所示。

表 4 - 4 经 DMVP 15 处理后数据在随机森林模型不同参数下分类准确率

Test precision of DMVP 15 dataset under different RF paramaters(%)

Number of tree	Number of random feature selected					
	5	10	15	20	25	30
10	69.01	72.19	72.32	72.63	73.26	73.42
20	71.42	73.11	73.57	73.73	74.06	73.3
30	73.53	73.53	73.65	74.63	74.49	74.38
40	73.46	73.77	74.73	74.72	74.89	75.05
50	73.09	73.94	74.41	74.69	74.85	75.07
60	72.88	74.11	74.28	74.65	74.99	74.53
70	73.67	74.71	74.79	75.07	75.09	75.14
80	73.28	74.39	74.69	74.72	74.93	75.18
90	73.26	74.71	74.86	74.91	74.91	74.79
100	73.59	74.46	74.87	74.55	74.77	**75.17**

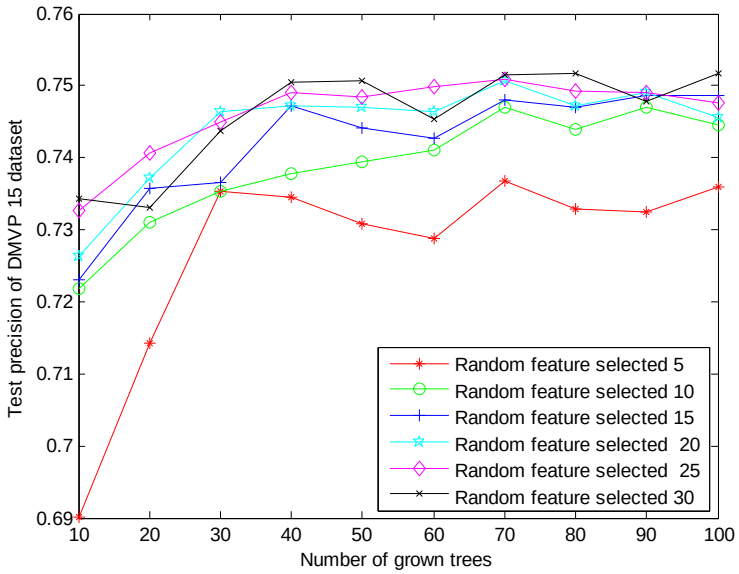


图 4 - 7 经 DMVP 15 处理后数据在随机森林不同参数下的分类准确率

图 4-7 中横坐标代表森林中树的数目 t =10~100（步长为 10），不同线条代表不同随机变量选择数 v =5~30（步长为 5）—从 5 以 5 为步长递增到 30。

由图可得，在测试数据和训练数据固定的情况下（如 DMVP 15 数据），随机森林对其分类准确率随树的个数 t 和随机特征选择数 v 的变化而变化，总体而言，分类准确率随 t 和 v 的增大而增大，当 t=100，v=30 时，达到相对最优值 75.17%。

4.5 实验汇总

4.5.1. 实验方法

本节中对原始数据 Original data 和经 DMVP 处理后数据（DMVP 15）分别进行分类模型建立和测试，采用综合评价指标 F1 Index 对 C4.5、CART 和随机森林进行横向比较，同时加入最常用的模式分类算法 SVM 和 AdaBoost 的实验结果进行对照。F1 Index 的定义为：

$$F1(f)=\frac{2*Precision*Recall}{Precision+Recall} \tag{4-4}$$

其中 *Precision* 和 *Recall* 分别代表准确率和召回率，本文中分类准确率皆表示 *Precision*，在此为了更加公平的进行分类结果对比，采用综合评价指标 *F1 Index* 对多种算法进行对比。

4.5.2. 实验结果及分析

对原始数据 Original data 和经 DMVP 处理后数据（DMVP 15），分别采用不同算法分类结果（*F1 Index*）如表 4-5 所示。

由表 4-5 结果可知，在 21 类 TEP 数据同时分类问题上，本文主要研究的随机森林算法平均测试结果优于 C4.5、CART、SVM、Adboost 等算法，其中第三、四、六、八、九、十、十一、十二、十三、十四、十六、十七、十八、十九、二十这 15 类故障和正常数据（Normal）的 *F1 Index* 都优于其它算法；同时根据表 4-5 所示，从单个算法的角度来看，DMVP 处理方法使得采用任何一种算法的性能都得到了大幅提升（大约 10%左右），随机森林在 DMVP 处理后数据上的分类 *F1 Index* 则达到了最高值 74.75%，这对于故障诊断的意义无疑是非常重大的。

表 4 - 5 *F1 Index* 实验汇总

F1 Index										
	C4.5		CART		Random Forest		SVM		AdaBoost	
编号	Original	DMVP15	Original	DMVP15	Original	DMVP15	Original	DMVP15	Original	DMVP15
Normal	19.3	19.2	15.4	22.9	18.7	**21.1**	13.1	21.0	18	22.7
1	96.7	95.6	97.7	96.4	99.3	99.1	97.7	98.0	98.6	99.4
2	94	97.8	94.7	93.5	97.7	99.8	95.8	100.0	96.9	98.2
3	19.4	19.9	15.7	28.4	20.2	**30.9**	17.3	15.2	15.7	25.7
4	88.5	94.9	91.3	100	94.5	**100**	85.8	75.7	91.2	99.9
5	51.2	98.3	61.1	97.4	69.3	97.3	92.6	88.1	66.6	98.3
6	99.4	99.8	99.6	100	99.6	**100**	100.0	100.0	99.7	100
7	98.4	96.2	99.2	100	99.9	99.2	100.0	100.0	99.6	99.8
8	35.9	61.6	40.5	68	62.8	**75.5**	50.0	67.8	51	74.5
9	14.9	12.8	18	12.6	16.2	**20.7**	15.0	18.6	16.6	17.1
10	29.4	45.2	37.1	41.9	41.1	**52.6**	19.6	28.1	42.2	52.1
11	63.5	92.7	58.3	93.9	74.9	**98.9**	13.3	78.8	69.9	99.3
12	43	79.5	43.6	79.1	56.7	**90**	57.5	76.7	55.3	88.6
13	26.4	30.4	31.6	48.5	34.4	**47.9**	39.8	55.6	37.6	51
14	95.2	99.4	89.8	99.4	96.7	**100**	73.6	100.0	96.6	100
15	16.3	17	17.3	10.3	19.6	15.5	16.1	17.7	17.3	17.2
16	39.4	42	48.9	48.6	50.4	**60.2**	26.7	29.2	52.4	57.5
17	79.6	93.4	78.8	96.9	91.3	**99.5**	79.2	86.8	88.4	98.5
18	78.5	76.4	72.6	79.3	83.6	**89**	89.2	80.2	82.7	88.5
19	65.9	97.2	57.8	98.1	72.4	**100**	55.3	100.0	71.4	98.9
20	48.7	67.2	50.1	61.5	60.7	**72.7**	52.7	57.3	58.6	72
平均	57.31	68.4	58.05	70.31	64.76	**74.75**	56.68	66.41	63.2	74.2

随机森林之所以在 TEP 数据上取得这么好的性能，有以下几个原因：

1、作为基于决策树的集成算法，随机森林自身有着比较好的分类能力、不容易过拟合、对噪声不敏感等性质，这些导致随机森林对不同故障类型数据有着比较强的分类能力。而 DMVP 提取数据动态特征的方法直接加大了不同类型数据之间的差异性，对于进一步提升分类准确率有着直接意义；

2、随机森林的非线性树形结构也是其重要优势，对于 TEP 数据来说，由于其变量值在时间上的动态特性，导致同一故障类型同一变量值的幅度有一定的起伏，即同一类型数据本身蕴含了两种或多种呈现方式，如果想通过某个分类超平面将这种数据分为同一类是比价困难的;随机森林不依赖与某个超平面，而是根据统计学原理对同一类型数据进行统计，然后根据统计结果进行决策分类，再加上多棵树的综合决策，这种决策分类的方式拓展了随机森林的分类精度，也使其比较适合 TEP 数据分类。

除了在分类准确率上，随机森林体现出比较好的性能之外，在运行速度方面，随机森林也具备一定的优势。决策树自身是一种运行速度非常快的算法，这是由它的树形决策方式决定的，由于随机森林由于是多棵树的集合，所以其速度要低于决策树，并且树的数目的多少直接决定了随机森林的运行速度，另外随机森林引入了随机特征选择，即每个结点分裂时，只需要计算选取特征的不纯度信息即可，这样可以极大的节省运算时间，对于高维数据而言，这个特征对于提升运算速度的效果会比较显著。和 SVM、Adaboost 等算法相比较，随机森林仍然具备很大的速度优势。并且随机森林的并行结果决定了其面对海量数据的优越性，即可以利用多种资源进行并行计算。

4.6 本章小结

本章主要研究了经 DMVP 动态特征提取后的 TEP 数据的不同算法分类效果。首先对 TEP 数据进行可视化分析，从横向纵向不同角度分析该数据的数学特征；然后根据可视化分析结果提出了 DMVP 方法，并在第四类、第九类、第十一类这三类故障进行分类试验，通过实验结果对比，验证了该方法的显著效果；之后分别采用决策树算法和随机森林算法对 DMVP 处理前后的 21 类数据分别进行分类实验，并和其他几种算法（SVM、Adaboost）结果进行对比分析，结果表明本文提出的结合 DMVP 方法和随机森林算法对 TEP 数据进行分类的方法优于其它算法；最后，本文分析了随机森林性能优于其它算法的原因。

总结与展望

故障诊断是工程领域的重要研究内容，基于数据驱动的故障诊断是当前这一领域的研究热点，本文以田纳西伊斯曼化学品公司仿真数据为研究对象，对其进行故障诊断研究。

本文首先简述了故障诊断的研究背景和意义、国内外研究现状和目前已存在的研究方法，然后着重研究基于数据驱动的方法中的随机森林算法，在故障类型和数据一一对应的基础上，通过数据分类达到故障诊断的目的。总结全文的主要工作，具体内容如下：

1、研究了决策树算法，以及基于决策树的集成算法随机森林。详细研究其算法来源、算法原理和改进理论，并通过实验验证了随机森林在分类准确率、过拟合程度、噪声容忍性等方面的优越性能。

2、提出了一种提取数据动态特征的方法（DMVP）。本文通过对数据的横向和纵向可视化分析，得出了 TEP 数据具有特殊的时间序列特征，通过提取动态方差和均值特征提取，有效的加大了不同故障数据之间的差异性，极大的提升了后期采用算法进行分类的效果。

3、结合本文提出的数据动态特征提取方法（DMVP）和随机森林算法，通过实验进行选参，与其它分类算法（C4.5、CART、SVM、Adboost）相比，得出了最优分类结果。

本文所提出算法虽然取得了一定的成功，但是仍然一些不足。例如，本文提出的 DMVP 方法虽然有效的提取了数据特征，提高了分类准确率，但对动态时序数据的特征挖掘还不够充分。在将来的工作中，将考虑从以下三个方面改进：

1、针对单个故障数据集进行分段研究。由于 TEP 数据在时间上的动态特性，导致同一故障类型同一变量值的幅度有一定的起伏，即同一类型数据本身蕴含了两种或多种呈现方式，虽然随机森林的非线性树形结构在一定程度上适合于此类数据，但如果通过聚类算法或根据可视化数据信息进行人工设定阈值对时序数据进行分段，并分别进行建模和测试，对于分类效果会有更大的突破；

2、通过变量值变化直接定位故障。对数据进行变量重要性测量，即通过相关算法（如 PCA、随机森林等）对 TEP 数据进行变量重要性测量，将某一故障和某些变量直接对应起来，当某些变量发生显著变化时，可以直接对应诊断出某一故障；更进一步，可以考虑将本文采用的 21 类故障同时分类策略和变量重

要性测量结合使用，相辅相成，以达到更好的故障诊断效果；

3、更多的实验平台和数据类型研究。化工生产过程是个非常复杂的系统，其物理系统和实际应用环境也各不相同，TEP 数据只是一种仿真数据，虽然具有一定的代表性，但其蕴含的信息毕竟还很有限，更深入的故障诊断研究需要更多的化工数据做支撑，通过对更多的数据进行试验比较分析，将会得到更加可靠的结果，对于研究算法的实际故障诊断应用也会有更加充分的依据。

参考文献

[1] http://baike.baidu.com/link?url=1gU9o2Y1kJsZVnt6UKWJputaSXdv0Nmdk98EZK3e-A2pqs7YGrLbVYgYm_yVJJmk.
[2] 安全文化网.化工事故案例[Z].2013. http://www.anquan.com.cn.
[3] Isermann, R., Ball, P. Trends in the application of model based fault detectionand diagnosis of technical processes. In Proc. of the 13th IFAC World Congress, Piscataway, New Jersey: IEEE Press, 1996, N: 1-12.
[4] Kesavan, P., Lee, J.H. Diagnostic tools for multivariable model-based controlsystems. Ind. Eng. Chem. Res. 1997, 36(7): 2725-2738.
[5] Frank P M. Fault diagnosis in dynamics systems using analytical and knowledge based redundancy: a survey and some new results [J]. Automatic, 1990, 26(3):459-474.
[6] Venkatasubramanian V, Rengaswamy R, Kavuri S N. A review of process fault detection and diagnosis Part II: qualitative models and search strategies [J]. Computers and Chemical Engineering.2003, 27(3):313-326.
[7] Venkatasubramanian V, Rengaswamy R, Kavuri S N, Yin K. A review of process fault detection and diagnosis Part III: process history based methods [J]. Computers and Chemical Engineering.2003,27(3):327-346
[8] Kano M, Hasebe S and Hashimoto I. A new multivariate statistical process monitoring method using principal component analysis [J]. Computers and Chemical Engineering, 2001, 25:1103-1113.
[9] Leo H. Chiang, Mark E. Kotanchek, Arthur K. Kordon. Fault diagnosis based on Fisher discriminant analysis and support vector machines IJ].Computers & Chemical Engineering, 2004, 28(8):1389-1401.
[10] J. Zhang, A. J. Morris. On-line process fault diagnosis using fuzzy neural networks [J]. Intelligent Systems Engineering, 1994, 3(l):37-47.
[11] Yunfeng Li, Zhifeng Wang, Jingqi Yuan. On-line Fault Detection Using SVM-based Dynamic MPLS for Batch Processes [J]. Chinese Journal of Chemical Engineering, 2006, 14(6):754-758.
[12] Amit Y, Geman D. Randomized inquiries about shape: an application to handwritten digit recognition. Technical Report 401, Dept of Statistics, University of Chicago, IL, Nov 1994.
[13] Amit Y, Geman D. Shape quantization and recognition with randomized trees [J]. Neural Comput, 1997, 9(7).
[14] T. K. Ho. Random Decision Forest. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 1995,

8:278-282.
[15] T. K. Ho, The Random Subspace Method for Constructing Decision Forests, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(8):832-844.
[16] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1):5-32.
[17] Breiman L. Bagging Predictors [J]. Machine Learning, 1996, 24(2): 123-140.
[18] Schapire R E. The Strength of Weak Learnability [J]. Machine Learning, 1990, 121(2): 256-285.
[19] Hunt E B, Mrin J, Stone P J. Experiments in induction [M].New York: Academic Press, 1966.
[20] Quinlan J R. Induction of decision tree[J]. Machine Learning, 1986, 1(1): 81-106.
[21] Quinlan J R. C4.5 Programs for Machine Learning [M]. SanMateo: Morgan kaufmann Publishers, lnc, 1993.
[22] L Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. Classification and Regression Trees. Wadsworth, Belmont, (1984).
[23] Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization[J]. Machine Learning, 2000, 40(2).
[24] Gislason P O, Benediktsson J A, Sveinsson J R. Random Forests for Land Cover Classification [J]. Pattern Recognition Letters, 2006, 27(4).
[25] Lee S L A, Kouzania A Z, Hu E J. Random Forest Based Lung Nodule Classification Aided by Clustering[J]. Computerized Medical Imaging and Graphics, 2010, 34(7).
[26] Diaz-Uriate R, Andres S A D. Gene Selection and Classification of Microarray Data Using Random Forest [J]. BMC Bioinformatics, 2006, 7(3).
[27] A Bosch, A Zisserman, X Muoz. Image Classification using Random Forests and Ferns. Computer Vision, 2007. ICCV
[28] Lan Guo, Yan Ma, Bojan Cukic, Harshinder Singh. Robust Prediction of Fault-Proneness by Random Forests. Software Reliability Engineering, ISSRE, 2004.
[29] Bo-Suk Yang, Xiao Di and Tian Han，Random forests classifier for machine fault diagnosis, Journal of Mechanical Science and Technology 22 (2008) 1716~1725.
[30] J Shotton, A Fitzgibbon, M Cook. Real-Time Human Pose Recognition in Parts from Single Depth Images. Computer Vision and Pattern Recognition, 2011 ,1297-1304
[31] A Criminisi, J Shotton, E Konukoglu. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised

Learning. Microsoft Research technical report TR-2011-114 ,2011.
[32] Downs J. J. and Vogel E. F. A Plant-Wide Industrial Process Control Problem [J]. Computers & Chemical Engineering, 1993, 17(3): 245-255.
[33] Chen X W, Liu M. Prediction of Protein-protein Interactions Using Random Decision Forest Framework [J]. Bioinformatics, 2006, 21(24).
[34] Pal M. Random Forest Classifier for Remote Sensing Classification [J]. Remote Sens, 2005, 26(1).
[35] Auret L, Aldrich C. Change Point Detection in Time Series Data with Random Forests [J]. Control Engineering Practice, 2010, 18(8).
[36] Patton R. J. Chen. J. A review of parity space: approaches to fault diagnosis. Proceeding of IFAC Fault Detection [J], Supervision and Safety for Technical Processes. Germany, 1991(28): 65-81.
[37] 何小斌, 基于统计学方法的自适应过程监控与故障诊断,上海交通大学博士学位论文.2008.
[38] 刘艳丽,随机森林综述,南开大学硕士学位论文,2009.
[39] 孙烈.随机森林及其在色谱指纹中的应用研究[D].大连理工大学硕士学位论文, 2009.
[40] 雷震.随机森林及其在遥感影像处理中应用研究[D].上海交通大学博士学位论文,2012.
[41] 雍凯.随机森林的特征选择和模型优化算法研究[D].哈尔滨工业大学硕士学位论文,2008.
[42] 邱一卉.随机森林在电信行业客户流失预测中的应用[D].厦门大学硕士学位论文,2008.
[43] 谢芳芳.基于支持向量机的故障诊断方法[D].湖南大学硕士学位论文,2006.
[44] 扈彬.基于随机森林与卡尔曼滤波的人体跟踪方法研究[D].天津师范大学硕士学位论文,2012.
[45] 庄进发.基于模式识别的流程工业生产在线故障诊断若干问题研究[D].厦门大学博士论文,2009.
[46] Amit Y, Geman D. Shape Quantization and Recognition with Randomized Trees [J].Neural Computation, 1997, 9(7).
[47] Wolpert D H, Macready W G. An Efficient Method to Estimate Bagging's Generalization Error [J]. Machine Learning, 1999, 35(1).
[48] Breiman L. Out-of-bag Estimation [EB/OL]. [2010-06-30]. http//stat. berkeley. edu/ pub/ users/ breiman /OOB estimation.ps.
[49] Freund Yoav, Schapire Robert E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting [J]. Journal of computer and system sciences, 1997, 55(1):119-139.
[50] I. Monroy, R. Benitez, G.Escudero, M.Graells, A semi-supervised approach to fault diagnosis for chemical processes [J]. Computers and Chemical

Engineering, 2010, 34:631–642.

[51] Z.B. Zhu, Z.H Song, A. Palazoglu. Process pattern construction and multi-mode monitoring [J]. Journal of Process Control, 2012, 22:(247–262.

[52] L.H. Chiang, F.L. Russell, R.D. Braatz. Fault Detection and Diagnosis in Industrial Systems, Springer, London, 2001.

个人简历、在学期间发表的论文与研究成果

姓名：张晓丹　　性别：女　　民族：汉族　　出生年月：1987.01

教育经历

2010.9-2013.7	中国科学院大学	物流工程	硕士
2006.9-2010.7	郑州大学	生物医学工程	学士

已发表（录用）文章

Yi Peng, **Xiaodan Zhang**, Jianbin Jiao, Dynamic Fault Diagnosis in Chemical Process Based on SVM-HMM , Proc.of IEEE International Conference on Mechatronics and Automation(ICMA), 2013.

在审文章

Xiaodan Zhang, Yi Peng, Xiaogang Chen and Jianbin Jiao, Fault Diagnosis Based on Random Forest in Chemical Process,CJCE 2014 (submitted).

软件著作权

无线上网认证系统　开发日期：2012.03-2012.07 证书获得时间：2013.01

淘车位系统　　开发日期：2011.07-2011.11 证书获得时间：2012.11

致　谢

在中国科学院大学攻读硕士学位的学习生活使我受益匪浅。在毕业论文完成之际，由衷地感谢这几年来曾经给予我无数帮助的老师、同学、朋友和家人。

首先，我要感谢我的导师焦建彬教授给予了我珍贵的求学机会，感谢他在我攻读硕士学位期间对我的悉心指导和鼓励。恩师给了我良好的科研环境和实践机会，让我从科研能力到为人处世方面都受益匪浅。他严谨的治学之风和对事业的孜孜追求深深地影响着我，他诲人不倦的精神和对我的谆谆教导让我感动，他的言传身教将使我受益终生。

其次，感谢叶齐祥、韩振军老师，他们在我的科研学习和理论研究中给予了耐心的指导。他们渊博的专业知识、自强不息的学习精神、扎实的动手能力和对我的谆谆教导，让我受益匪浅。

感谢参加开题及中期评阅的各位老师和专家们，他们丰富的经验和无私的工作对论文方向和研究进度的把握和指点给整个研究工作带来了巨大的帮助。

我还要衷心感谢实验室所有成员，这几年里我们一起学习定时运动，形成了家人一样的氛围。陈孝罡师兄总是在第一时刻给予我学习上的建议，并在论文的撰写和修改中给予了很多帮助，对我的科研学习起到了关键的指导作用；张立国师兄、彭艺师姐、李策师姐无论在科研还是生活中，都给我树立了良好的榜样，总是热心的帮助我们这些师弟师妹们；同届的武利军、梁吉祥、邹佳凌、杨威、高山我们一起学习共同成长，相互帮助互相关心，结下了兄弟姐妹般的友谊；另外感谢室友关婉秋、于新菊，我们一起生活了三年，共同面对成长中的喜悦和困惑，一起凝结了最难忘的回忆。

最后要感谢我的父母，感谢他们多年来一直给我最无私的帮助与鼓励，让我有勇气面对一切。感谢我的父母为我所做的一切，愿他们能够为我而骄傲。

最后，感谢参加论文评审和答辩的各位老师。

张晓丹

2014 年 4 月