（封面、封底用120克白色铜版纸打印，红色字打印时删掉）



**本科生毕业设计[论文]**

（上一行为华文中宋小初号加粗居中）

**基于工业大数据的生产设备故障诊断**

（黑体2号加粗居中）

|  |  |
| --- | --- |
| 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 |

年 月 日（华文中宋3号居中）

**学位论文原创性声明**

（黑体小2号加粗居中）

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

（宋体小4号）

作者签名： 年 月 日

**学位论文版权使用授权书**

（黑体小2号加粗居中）

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密囗，在 年解密后适用本授权书

2、不保密囗 。

（请在以上相应方框内打“√”）

（宋体小4号）

作者签名： 年 月 日

导师签名： 年 月 日

(注：此页内容装订在论文扉页)

**摘 要** （黑体小2号加粗居中）

由于计算机硬件以符合摩尔定律的速度迅猛发展，计算机数据存储、数据传输和分布式计算的成本都大幅度降低。而现代化的工厂中，往往都布置有大量的传感器，而且随着存储成本的降低，读取到的设备信息变得更加丰富，由此便会产生海量的工业数据。

对于这一变化，全世界许多国家都相继提出相应的举措。最初是德国提出了“工业4.0”的概念，之后美国推出“工业互联网”，我国也相继推出“中国制造2025”的概念，其核心都指向智能制造。而工业大数据技术则是这些内容中的核心部分。

在工业设备的运行过程中，自然磨损、设备超载、操作不当等多种原因会导致设备的性能发生下降，甚至于产生故障或者是异常。而通过对设备加装传感器进行监控，获取到设备的实时信息并且加以梳理计算，就可以得到设备各个部位的实时运行状态，从而实现对设备的监控。而如果出现了设备故障现象，则可以通过对历史数据进行数据挖掘、清洗形成故障模型，导入设备的最新运行数据进行故障诊断。

设备的故障诊断方法可以分为三种:基于机理模型的方法，基于数据驱动的方法，基于知识工程的方法[1]。本文将采用基于数据驱动的方法中的基于分类的方法进行故障模型的构建。同时为了对比不同分类方法的性能，本文采用了两种分类方式进行比较。

本文主要研究工作和成果如下：

（1） 建立了以决策树算法为基础的故障诊断模型；

（2） 实现了以支持向量机为核心的数据驱动方法。对支持向量机的核心原理进行了研究，以风力涡轮机齿轮箱的健康数据、故障数据进行了SVM的分类检测，实现了简单的故障诊断；

（3） 比对了两种策略的运行时间，故障诊断正确率以及其他的一些的性能度量。

（宋体小4号，行间距固定1.5倍行距，字符间距为标准）

**关键词：**故障诊断；工业大数据；数据驱动；决策树；支持向量机

（黑体4号加粗） （宋体小4号）

**Abstract**（Time New Roman小2号加粗居中）

（Time New Roman小4号，行间距固定1.5倍行距，字符间距为标准）

As the computer hardware rapidly develops at a rate consistent with Moore's Law, the cost of computer data storage, data transmission, and distributed computing has been greatly reduced. In the modern factories, a large number of sensors are often arranged, and as the cost of storage decreases, the read device information becomes more abundant, and thus a large amount of industrial data will be generated.

For this change, many countries around the world have put forward corresponding measures in succession. Initially, Germany proposed the concept of "Industry 4.0". After the United States introduced the "Industrial Internet," China has also successively introduced the concept of "Made in China 2025." Its core points to intelligent manufacturing. Industrial big data technology is a core part of these contents.

During the operation of industrial equipment, natural wear, equipment overload, improper operation, and other reasons can cause the performance of the equipment to drop, and even result in failure or abnormality. By adding sensors to the equipment for monitoring, obtaining real-time information from the equipment and combing calculations, the real-time running status of various parts of the equipment can be obtained, thereby realizing the monitoring of the equipment. If there is a device failure, you can perform data mining and cleaning on the historical data to form a fault model and import the latest operating data of the device for fault diagnosis.

Currently, there are three types of device diagnostic methods that are popular in the world: mechanistic model-based methods, Data-Driven methods, and knowledge-based methods. This article will use a classification-based approach based on a Data-Driven approach to build a fault model. At the same time, in order to compare the performance of different classification methods, this paper uses two classification methods to compare.

The main research work and achievements of this article are as follows:

(1) Establish a fault diagnosis model based on decision tree algorithm, and study the differences between ID3 and C4.5;

(2) A Data-Driven approach based on Support Vector Machines(SVM) is implemented. The core principle of SVM is studied. The classification and detection of SVM is performed based on the health data and fault data of the wind turbine gearbox, and a simple fault diagnosis is realized.

(3) Compare the running time of two strategies, the accuracy rate of fault diagnosis, and some other performance metrics.

**Key Words：**Fault Diagnosis; Industrial Big Data; Data-Driven Method; Decision Tree; Support Vector Machine

（Time New Roman 4号加粗） （Time New Roman小4号）

**目 录**（黑体小2号加粗居中）

**摘要** Ⅰ

**Abstract** Ⅱ

**1 绪论** 1

1.1 ×××××× 1

1.2 ×××××× 3

1.2.1 ×××××× 7

1.3 ×××××× 10

**……**

**……**

**……**

3**□□□××××××** 20

3.1□□×××××× 20

3.2□□×××××× 23

3.2.1□□×××××× 25

3.3□□×××××× 30

**……**

**……**

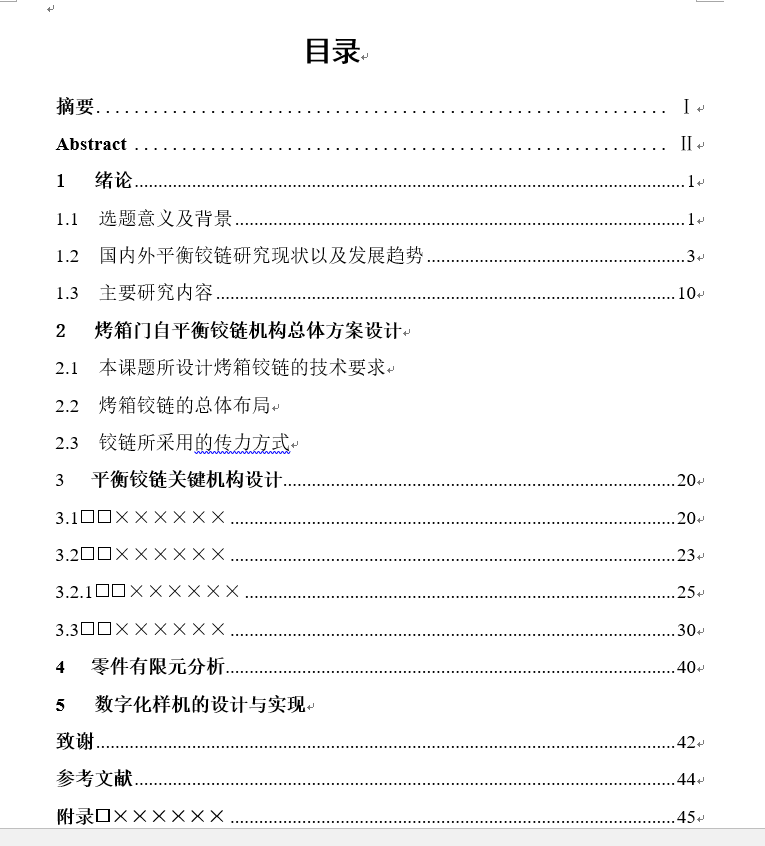
**4□□□结论** 40

**致谢** 42

**参考文献** 44

**附录□××××××** 45

（章为宋体小4号加粗，其余宋体小4号，字母、阿拉伯数字为Time New Roman小4号）

****

1. **绪论**（黑体小2加粗居中）
   1. **选题背景和意义**（黑体4号加粗, 字母、阿拉伯数字为Time New Roman4号加粗）

在计算机行业还未能发展到如今这般规模的时候，人们只能选择抽样的数据、局部的数据和片面的数据，纯粹靠经验、理论、假设和价值观去发现、理解未知领域的规律。而这样做的结果，就是对真实现象的抽象归纳与演绎推理，这就不可避免的包含了各种主观上的因素。同时由于样本的局部性，很多推理归纳出来的结果与实际现象具有极大的偏差。

而如今的所谓大数据，通常都指数据量在“太字节（TB）”即2的40次方以上，一般情况下难以收集、存储、管理以及分析的数据。而且随着科技进步，大数据对于“大”的含义还在不断地刷新。但是大数据不仅仅只关乎于数据量的大小，而且还与其他的因素有着千丝万缕的关系。

大数据其真正的意义在于：我们可以通过各式各样的传感器，实现与真实世界更加紧密、准确的连接。在得到实时数据后进行整合、挖掘、云计算，去逐步的逼近真实世界，挖掘出那些未曾被我们发现的隐藏规律，建立更加符合真实的数学模型，这是大数据的魅力所在。而在大数据的庞大篇幅中，工业大数据占据着重要地位。工业大数据是智能制造的关键技术，主要作用是打通物理世界和信息世界，推动生产型制造向服务型制造转型[2]。

而在工业领域，工业大数据的一大发展方向就是故障诊断，数据的产生和记录贯穿于一台设备从投入生产到损耗的全过程，而通过一定数量的智能传感器，我们可以监控某些生产设备的所有信息，使设备在生产线的实时状态远程监控成为可能，这一方面改善了工作人员的工作环境，另一个更为重要的方面就是提高了设备发生故障或者异常时候的反应速度以及排除故障的效率。

伴随着工业生产水平发展的突飞猛进，工业设备精度越来越高，结构越来越复杂，所以在车间内很多设备的故障都未能得到及时的发现和解决，这一点很容易对工厂造成巨大的损失.由于设备愈加复杂所导致的设备故障信息数据呈现指数型增长，而运行时所产生的海量数据，采用传统基于机理模型的方式已经很难负载如此巨量规模的数据分析，进行故障诊断了。

此外，工业设备结构极其复杂，不同模块之间可能会产生故障的交集，人工分析或者是传统的先验知识故障检测手段已经无法准确、迅速的完成故障的诊断。因此，结合工业大数据对工业设备所产生的海量数据进行数据挖掘分析建立故障诊断模型，对于提高设备维护效率、迅速有效解决故障、降低维修费用有巨大意义。

（宋体小4号，行间距固定1.5倍行距，字符间距为标准）

**1.2 国内外研究现况及发展趋势**

**1.2.1 国内研究现状**（黑体小4号加粗, 字母、阿拉伯数字为Time New Roman小4号加粗）

金风科技股份有限公司基于大量风场的历史故障信息进行大数据分析， 对 SCADA(Supervisory Control and Data Acquisition，监控和数据采集系统)瞬时数据的时间序列模式提取，挖掘桨距角一致性、变桨过程曲线模态、振动模式、变桨电机温差、ng5充电电流差异等断裂征兆模式，通过多模型融合和深度学习，提前90h进行断裂预警，通过预防性维修消除重大故障隐患[3]。张鹏在在线性模型的基础上进行了大量改进，提出了将卡尔曼滤波器和基于非线性模型相结合的方法，并在航空发动机与传感器上进行了验证[4]。徐德民等人以航行器为研究对象，使用连续-离散无迹卡尔曼滤波算法对航行器执行器进行故障诊断[5]。南京航空航天大学鲁峰等人对发动机进行了仿真建模并且获得了影响系数矩阵，成功实现了对发动机中的气路故障进行诊断[6]。

**1.2.2 国外研究现状**（黑体小4号加粗, 字母、阿拉伯数字为Time New Roman小4号加粗）

国际权威专家 Frank 将故障诊断的方法总结为三种:基于机理模型的方法、基于数据驱动的方法、基于知识工程的方法[7] 。葡萄牙科英布拉大学的Marco S. Reis教授和Geert Gins在《Industrial Process Monitoring in the Big Data/Industry 4.0 Era: From Detection, to Diagnosis, to Prognosis》一文中提到，过去的重点都是检测，也就是基于机理模型的一种对比当前数据的检测办法，实现高水平的检测速度和强度是过去IPM研究的主要重点。在处理新流程时，这是一个必要的步骤，但是随着时间推进，越来越多的因素阻碍了更先进的监控手段的发展[8]。因此为了解决这些障碍与挑战，我们需要找到一种方法可以找到故障根源，也就是我们当今时代的主流过程监测手段：故障诊断。未来的发展方向更是令人心驰神往，故障预检测，能够根据当前的运行数据获取未来一段时间内的设备运行状态预测。

瑞典吕勒奥理工大学的Lianwei Zhang开发除了一套专用于大数据检测以及维护的系统以及一种基于自适应核密度的异常检测（Adaptive-KD）方法，在工业场景中具有极大的使用价值。

**1.3 主要研究内容**

1.3.1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；

1.3.2 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；

1.3.3 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；

1.3.4 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。

**2 故障诊断的总体设计方案**

（黑体小2加粗居中, 字母、阿拉伯数字为Time New Roman小2号加粗）

**2.1 故障模型的要求**（黑体4号加粗, 字母、阿拉伯数字为Time New Roman4号加粗）

故障模型应该是基于历史数据构建的，由于我们的模型是基于数据驱动，所以对于机械设备方面的先验知识需求量远小于基于机理模型和基于知识工程的构建方式。另外，我们的模型要能接受新的运行数据，对于数据进行测试，从而进行故障诊断的最终目的。

在精度上，模型的精度应该随着数据的不断完善而改进，构建模型所使用的数据越多，那么我们的模型的泛化程度就越高，对于各种实际情况的解读能力就会进一步提升。

**2.2 决策树建立故障树模型**

**2.2.1 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（也可以理解为样本分类的频率）为：

结合每一个分类的信息量，则此样本总体的信息熵为：

**2.2.2 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：

**2.2.3 ID3算法**

ID3是一种以自顶向下递归的方法构造决策树的贪心算法。其决策树的基本生成策略如下:

（1）树以整体样本作为单个节点开始；

（2）如果当前节点中，所有的样本都属于同一类，则该节点成为叶节点，并标记为当前样本的类；

（3）否则，使用前面提到的信息增益作为判断信息，选择信息增益最大的一个属性作，该各个属性值将成为该节点往下进行分支的依据。在这里，我们假设所有的属性都是分类的，即取离散值，连续值的属性必须离散化[9]；

（4） 对测试属性的每个已知的值创建一个分支，并据此划分样本子集；

（5） 算法使用类似的方法，递归地形成每个划分上的样本决策树。一旦一个属性出现在一个节点上，就不必在该节点的后代上考虑这个属性。

ID3算法虽然简单易用，但是也有很多缺陷：

（1） ID3算法缺乏对于连续值的处理手段，而在现实生活中，很多的应用环境都是采集到的连续值；

（2）计算信息增益的时候对于样本频率P(xi)有极大的依赖性，有时候会对模型造成很大的偏差；

（3）对噪声较为敏感，所谓噪声也就是一些在生成模型的时候就给定的错误数据；

（4）采用递归的方式形成模型，而且整个决策树的生成过程对于数据多次读取存写，所以算法较为低效，而且无法应用于大数据量的场合下。

下面是利用这些概念获得故障树的过程：



图2 ID3算法生成决策树流程图

为了提高决策树的模型精度，去除掉模型创建时一些错误数据的干扰，有两种剪枝方法可以用于提高决策树的正确分类能力：

（1）预剪枝方法(prepruning)，该方法通过提前停止树的向下延伸而对树剪枝。在各个节点向下分支之前，判断通往该分支的样本子集中的判断正确率进行对比，如果在当前节点的正确率高于分支后的子节点，那么就停止生长，这就是预剪枝的主要思想。该方法很多的分支都未曾“展开”，降低了过拟合的风险，而且还显著减少了决策树的训练时间和花销；但是另外一方面，一些分支虽然不能提升整体的泛化性能，但是由其再次展开的分支却有可能导致性能显著提高，而且预剪枝的“贪心”本质给这种方法带来了欠拟合的风险。

（2）后剪枝方法(postpruning)，顾名思义，该方法是预先生成一颗完整的决策树，然后从每一个叶节点往上查看父节点，计算如果该父节点进行剪枝成为叶节点后是否会提高判定精度来决定是否剪枝。这种方式直到无法提高决策树性能为止。对比预剪枝方法，后剪枝方法保留了更多的分支。一般情况下，后剪枝方法的欠拟合风险比较小，泛化性能优于预剪枝方法生成的决策树。但是由于后剪枝方法是在决策树生成后在进行的，所以在训练时间和花销上会比预剪枝方法高得多。

**2.3 支持向量机二分类原理**

**2.3.1 SVM原理**

支持向量机（Support Vector Machine，SVM）是一种经典的二分类模型算法，基本模型定义为特征空间中最大间隔的线性分类器，其学习的优化目标便是间隔最大化，因此支持向量机本身可以转化为一个凸二次规划求解的问题。

对于二分类学习器，假设数据是线性可分的，这时分类学习就是找到一个合适的超平面来进行分类，该超平面能够将不同类别的样本分开，下图中的点是低维数据表示，而这些数据对应的超平面就是中间那根线。对于点来说，线毫无疑问是处于高维的超平面了。随着维数增加，超平面总是比数据的维数多一维。



图3 二维超平面

但是这样的超平面可能存在多个，我们应该寻找的最优超平面该如何获取呢？

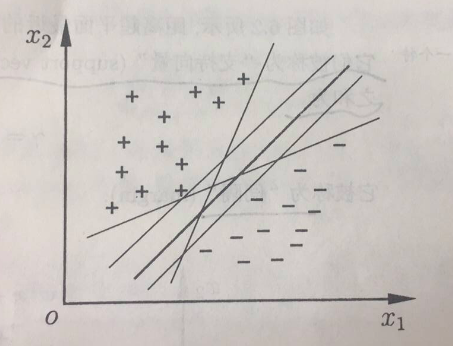


图4 存在多个划分超平面将两类训练样本分开

直观上看的话，我们应该寻找位于两类训练样本“正中间”的超平面作为我们的最优超平面，因为这个超平面对于整个训练样本局部扰动的“容忍”性能最好。换言之，这个超平面所产生的分类结果是最鲁棒的，对于未知的数据的分类能里最强。

在样本空间中，超平面可以通过如下线性方程表示：

其中***w*** = （w1;w2;w3;...;wd）为超平面的法向量，决定了超平面的方向。b为位移，决定了超平面与原点之间的距离。既然有了可以用数学方式表达的超平面，那么样本空间中任意一点到超平面的距离也就可以表示为：

假设超平面能够正确分类，即对于，则有:

那么我们可以得到距离超平面最近的几个训练样本可以使得上式中的等号成立，这几个样本就称为支持向量，两个异类支持向量到超平面的距离之和为：

这个距离就被称作“间隔”（margin）：

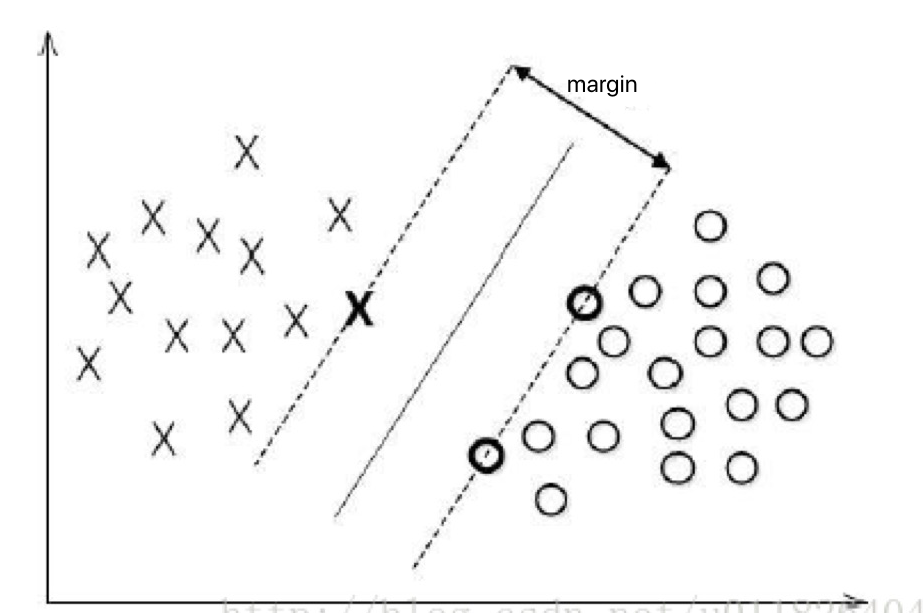


图5 支持向量与间隔

显而易见的，我们的目标就是找到具有最大间隔的超平面作为最优超平面。那么此超平面需要符合下列特点：

由此，为了得到最大间隔下的最优超平面，只需要最小化,于是，上式可以改写为：

这就是“支持向量机”（Support Vector Machine，SVM）这一方法的基本型。

**2.3.2 对偶问题**

由上面的支持向量机方法的基本型，我们可以知道这是一个带约束的凸二次规划问题，解决这个问题的较高效的办法是：对偶问题（dual problem）。

具体的解决方式为给每条约束添加拉格朗日乘子 则上述问题的朗格朗日函数可写为：

其中,由拉格朗日乘数法的思想，令对的偏导为0可以的得到：

利用两个式子代入到中可以消去，就得到了关于基本型的对偶问题：

通过求解出的，我们可以计算得出，从而得到最终具体的模型：

在上述的解答过程中，因为基本型有着不等式的约束条件，所以需要满足KKT（Karush-Kuhn-Tucker）条件：

于是对于任何的训练样本，都要有 = 0 或者 的限制条件。且必须是满足的样本才会出现在最大间隔边界上，说明这一个样本是支持向量。这一点也就代表着，训练完成后大部分的训练样本都不需要进行保留，只需要留下支持向量的那些样本即可。因为非支持向量的都等于0，去除之后不会对模型产生什么影响。

模型建立完毕之后，在对新的数据点进行预测时，实际上就是将这个新的数据点代入到分类函数中，若f(x)得出来的结果大于0，则为正类，否则为负类。

**2.3.3 SVM核函数**

假如训练样本线性不可分，放到我们最开始的例子里面的意思就是：我们无法用一条直线将所有的样本正确分类为两类。如下面的“异或”问题：



图6 异或问题

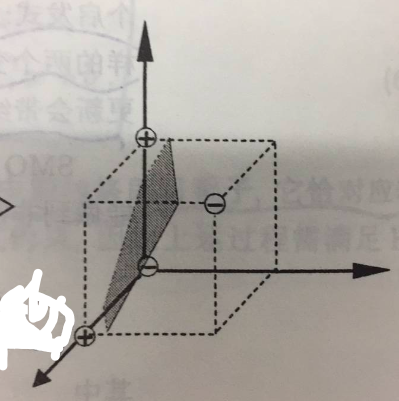
****

图7 非线性映射

对于这种情况，我们可以将原始空间的训练样本映射到一个更高维度的特征空间，使得样本在这个特征空间内线性可分[10]。

令表示将x映射到高维空间中的特征向量，那么对比前面线性可分的模型表示，可以得到相应的模型为：

同理可得：

同样可以得到他的对偶问题，如下：

映射后的特征空间维数可能会很高，会极大地加大我们训练时间和内存开销，直接计算 不一定可行。为了避开这个不稳定的区域，我们可以假设存在这样一个函数：

即在高维特征空间中映射向量的内积等于在原始样本空间中通过核函数计算后的结果。这就是“核技巧”。上述式子可以改写为：

求解之后可以得到：

在线性不可分的情况中，核函数的选择是影响SVM模型的性能至关重要的因素。在明确特征映射的形式之前，我们并不知道具体选择哪一种核函数，而且核函数的选择也隐式的定义了特征空间。如果在模型的构建过程中选择了错误的核函数，那么由此建立的支持向量机模型的性能将会下降许多。

下列几种常用的核函数：

|  |  |  |
| --- | --- | --- |
| 名称 | 表达式 | 参数 |
| 线性核 |  |  |
| 多项式核 |  | d 1 为多项式的次数 |
| 高斯核 |  | > 0 为高斯核的带宽 |
| 拉普拉斯核 | ) | > 0 |
| SIGMOD核 |  | Tanh 为双曲正切函数 |

此外，函数相互组合也可以得到核函数，

1. 如果 和 都是核函数，那么对于任意正数其线性组合：

也是一个核函数;

1. 如果 和 都是核函数，那么核函数的直积：

也是一个核函数;

1. 如果 是一个核函数，那么对于任意函数g(x):

也是一个核函数。

**3 具体方案设计**

×××××××××××××××××××××××××××××××××××

表3-1□××××××××××

|  |  |  |  |
| --- | --- | --- | --- |
| ××××× | ××× | ××× | ××× |
| ××××× | ×× | ×× | ×× |
| ××××× | ×× | ×× | ×× |
| ××××× | ×× | ×× | ×× |
| ××××× | ×× | ×× | ×× |

（表标题：位于表格上方，黑体小4号，字母、阿拉伯数字为Time New Roman小4号，表内容：宋体5号，字母、阿拉伯数字为Time New Roman 5号）

××××××××××××××××××××××××××× (如图3-1所示)



图3-1□××××××××××

（图标题：位于图下方，黑体小4号，字母、阿拉伯数字为Time New Roman小4号）

.......

--------章与章之间插入分页符----------

**参考文献** (黑体小2号加粗居中)

[1] 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.

[2] 卫凤林,董建,张群.《工业大数据白皮书(2017版)》解读[J].信息技术与标准化,2017(04):13-17.

( 宋体小4号)

[3] 王建民.工业大数据技术[J].电信网技术,2016(08):1-5.

[4] 张鹏. 基于卡尔曼滤波的航空发动机故障诊断技术研究[D]. 南京: 南京航空航天大学, 2009.

[5] 徐德民, 刘富樯, 张立川等. 基于改进连续-离散无迹卡尔曼滤波的水下航行器故障诊断 [J]. 西北工业大学学报, 2014, 32(5): 756-760.

[6] 鲁峰, 黄金泉, 孔祥天. 基于变权重最小二乘法的发动机气路故障诊断[J]. 航空动力学 报, 2011, 26(10): 2376-2381.

[7] Frank P M. Fault diagnosis in dynamics systems using ana-lytical and knowledge-based redundancy: a survey and some new results[J]. Automatica, 1990, 26(3): 459-474.

[8] Reis M S, Gins G. Industrial Process Monitoring in the Big Data/Industry 4.0 Era: From Detection, to Diagnosis, to Prognosis[J]. Processes, 2017, 5(3): 35.

[9] 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

[10]  [周志华](http://cs.nju.edu.cn/zhouzh/) 著.机器学习, 北京: 清华大学出版社, 2016年1月. (ISBN 978-7-302-206853-6)

……

……

**附录** (黑体小2号加粗居中)

×××××××××××××××××××××××××××

( 宋体小4号)

……

……

……