中国石油大学（华东）
CHINA UNIVERSITY OF PETROLEUM

项目编号	

大学生创新创业训练计划

创新训练项目申请书

项 目 名 称： 基于数据挖掘技术的核电站主泵状态预测

项目负责人： 刘嘉赓

联 系 电 话： 17854206393

所 在 学 院： 计算机与通信工程学院

学　　号　： 1501030209

专 业 班 级： 软件工程 1503 班

指 导 老 师： 龚安

申 请 日 期： 2017 年 7 月 15 日

教务处制

填 写 说 明

1、本申请书所列各项内容均须实事求是，认真填写，表达严谨，简明扼要。

2、申请人可以是个人，也可以是创新团队，“项目编号”不填。

3、本申请书为A4纸双面打印，左侧装订成册。可网上下载、自行复印或加页，但格式、内容、大小均须与原件一致。

4、负责人所在学院认真审核，经初评和答辩，签署意见后，由学院统一将申请书（一式3份）报送教务处。

承诺书

我保证填报内容的真实性，不存在知识产权争议。如果获得资助，我与本项目组成员将严格遵守学校的有关规定，并按计划认真开展研究工作，在项目研究过程中或结束时，接受学校对本项目的中期检查和结题验收，并按时提交工作总结和结题报告。

负责人（签名）：

成员（签名）：

年　　月　　日

一、基本情况

<table>
<tr><td>项目名称</td><td colspan="3">基于数据挖掘技术的核电站主泵状态预测</td><td>选题来源</td><td colspan="3">□教师拟定　□自拟</td></tr>
<tr><td>所属学科</td><td colspan="3">计算机科学、应用数学</td><td>项目年限</td><td colspan="3">□1 年　□2 年</td></tr>
<tr><td>主要预期成果</td><td colspan="3">研究报告</td><td>研究方向</td><td colspan="3">理工、数学</td></tr>
<tr><td>负责人姓名</td><td>刘嘉赓</td><td>学号</td><td>1501030209</td><td>专业班级</td><td>软件工程 1503 班</td><td>联系电话</td><td>17854206393</td></tr>
<tr><td>指导教师</td><td>龚安</td><td>职称</td><td>副教授</td><td>所在单位</td><td>计算机与通信工程学院</td><td>联系电话</td><td>13375571170</td></tr>
<tr><td>负责人曾经参与科研的情况</td><td colspan="7">参加 2017 年美国大学生数学建模竞赛，对智慧城市和精明增长问题进行了研究，并获得该题的 Meritorious Winner 奖项（一等奖）。</td></tr>
<tr><td>指导教师对本项目提供支持情况</td><td colspan="7">首先，指导老师就技术路线选取方面对团队进行了详细地指导；
其次，指导老师为团队提供了实验室的服务器使用权限，为团队在研究中所进行的大规模计算提供了便利条件。</td></tr>
<tr><td rowspan="5"></td><td>姓名</td><td>学号</td><td>专业班级</td><td colspan="2">联系电话</td><td colspan="2">QQ</td></tr>
<tr><td>刘宇航</td><td>151030106</td><td>地球物理 1501 班</td><td colspan="2">17806234810
13284096199</td><td colspan="2">1770084442</td></tr>
<tr><td>杨文文</td><td>1510020111</td><td>俄语 1501 班</td><td colspan="2">17854285133</td><td colspan="2">741144133</td></tr>
<tr><td></td><td></td><td></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>项目简介</td><td colspan="7">核电站由于其更加绿色的工作方式和更高的能源效率，在 21 世纪为众多国家所青睐。但是由于核电主泵结构的复杂性和大功率、高负荷地连续运转，设备在工作中随时间的推移和内部条件的变化，不可避免地会发生故障。这些故障轻则降低设备性能，重则停机停产，毁坏设备，造成严重的经济损失和人员伤亡。因此基于核电站主泵历史监测数据的状态预测对于核电站合理安排检修工作有着重要的意义。
本研究将把来自核电站主泵若干个测点所测得的历史数据作为高维时间序列，利用数据挖掘技术对其进行分析并对于其在下一阶段的工作状态进行预测。首先，将来自核电站主泵若干个测点所测得的历史数据进行数据清洗，并依据数据各个维度之间的相关性对其进行划分；然后依据时间序列预测的相关理论构建基于数据挖掘技术的状态预测模型，并设计相应评价标准选定预测效果较好的模型；最后，依据所模型的预测结果，对相关部门对主泵的检测与维护提出合理化建议。</td></tr>
</table>

二、立项依据

（一）研究目的

在核电站运行过程中，采用有效方法对运行状态进行监测和诊断，给操作员提供真实、清晰和完整的核电站状态信息，是核电站安全运行的重要保证。关于核电站的工作状态预测技术，国内外很多专家都投入了大量精力进行研究。但是传统的时间序列预测方法主要是基于统计分析的线性模型，例如回归分析法、ARMA 模型，这些模型难以模拟各领域时间序列的普遍存在的非线性特点。近年来人工智能不断发展，时间序列预测方法也随之扩展，基于机器学习的预测方法有支持向量机、人工神经网络等，这些方法及其组合已经成为时间序列研究领域的热点，并在核电站的工作状态预测方面得到较精准的预测效果。

本次研究将来自核电站主泵若干个测点所测得的历年数据作为高维时间序列，主要利用数据挖掘技术中的神经网络及其改进技术对数据集进行分析、处理并对其在下一阶段的工作状态进行预测，以此来判断设备在下一阶段的工作状态是否属于正常工作范围，进而帮助相关部门更加合理地对主泵进行检测与维护，避免不必要的检修，使核电站设备的运行、维护更加合理化、智能化。

（二）研究内容

为了能够成功地对核电站的工作状态进行较为科学准确的预测，达到辅助相关机构对核电站主泵进行合理维护的目的，本次研究将包括以下几个方面的内容。

第一，对原始数据集进行数据清洗，并依据数据各个维度之间的相关性对其进行划分。由于原始数据是设备的直接观测数据，数据集中部分数据存在缺失、重复或无效等问题，所以在对数据集进行分析之前首先要对原始数据集进行数据清洗，处理其中的“脏”数据。此外，核电站主泵中各个监测点之间的状态往往存在着某些关联，通过对其按相关性进行划分有助于对监测点状态进行更为合理的预测，提高预测结果的科学性和准确性。

第二，依据时间序列预测的相关理论构建基于数据挖掘技术的状态预测模型，并设计相应评价标准选定预测效果较好的模型。虽然目前国内外对于时间序列预测技术进行的广泛而深入的研究，但是该技术在核电站主泵状态预测的问题上尚未得到广泛应用，如何将其他领域较为成熟的时间序列预测技术应用到对主泵状态的无监督时间序列预测上并取得较好的预测效果将是本次研究的重点问题。此外，对于不同的预测模型其评价指标也不尽相同，选定一套合理、统一的评价标准将有助于对所构建的各个预测模型进行客观评价，进而选定一个预测效果较好的模型来对主泵状态预测问题进行后续研究。

第三，依据所模型的预测结果，对相关部门对主泵的检测与维护提出合理化建议。利用所选的模型对主泵状态进行预测，结合主泵出现异常或故障时的状态特征，对主泵可能出现的异常或故障进行预报，进而为相关部门安排主泵检修提出合理化建议。

（三）国内外研究现状和发展动态

本次研究所涉及到的技术主要包括时间序列的无监督学习以及数据挖掘技术中的神经网络及其改进的算法，这些技术也是如今智能算法研究的焦点之一，目前国内外关于这些领域的研究现状及发展动态如下。

时间序列无监督算法的研究内容涵盖了时间序列无监督特征提取、时间序列聚类、时间序列异常检测等，由于研究领域的不同，学界对挖掘任务的理解也存在一定的差异。如：Keogh 等人表示子序列聚类是没有意义的，他们试图对所有分割的子序列进行聚类，但是不管输入的数据集是什么样，其聚类结果总是正弦波。为改善子序列聚类总是产生正弦模式的缺陷，Ohsaki.M 等人在滑动窗口和 K-means 的基础上提出一种基于移动平均值的子序列聚类算法，Denton 等人提出径向分布函数实现基于模式的子序列聚类等。

神经网络方法在非线性关系表达中具有良好的拟合能力，在预测领域中受到越来越多的重视。目前已有理论证明，三层神经网络在隐藏层神经元足够多的情况下可以以任意精度模拟任意复杂的非线性或线性函数，然而神经网络预测方法的缺点在于训练效果对网络初始权值的敏感依赖性，网络初始化不够理想往往会使训练陷入局部最优，且单一的模型预测效果有限。为此，国内外的各位研究者针对单一神经网络的缺陷提出不同的模型改进及模型组合。例如：改进的 PSO 优化 BP 神经网络预测模型算法，粒子群加速了因子优化 BP 神经网络预测模型的初始权值和阈值，在模型训练时提高了网络的预测精度；利用小波分解技术对时间序列进行分解，然后运用模糊神经网络进行预测也有较好的性能；此外，寇茜茜等人提出的利用 RBM 改进的神经网络减轻了其对初始权值的依赖，提高了模型的泛化能力，在时间序列预测的问题上也取得了较好的效果。

（四）创新点与项目特色

应用创新：本次研究将何希平等人在 GDP 预测的问题上提出的基于 RBM 的神经网络时间序列预测模型引入到对核电站主泵各监测点数据所组成的时间序列的预测问题上，对该理论的应用领域进行迁移和扩展。

算法优化：本次研究将混沌机制引入粒子群算法的全局搜索过程，并与神经网络中的 BP 算法相结合以提高神经网络的收敛速度和搜索精度，在核心算法层面对神经网络算法进行了改进的尝试。

学科交叉性：本次研究涉及现代数学发展中的模糊数学、混沌理论等多个分支，以及现代计算机学科中智能算法的人工神经网络等多个分支领域，涉及的领域众多、学科交叉性强。

理论分析与实践相结合：本次研究虽然涉及到若干数学分支领域中的众多理论，但是利用这些理论所建立的模型需要通过实际观察结果来不断进行检验，模型的理论分析与实践检验紧密相连。

（五）技术路线、拟解决的问题

目前对时间序列进行数据清洗的技术已经较为成熟，但是对时间序列进行无监督学习以及对其进行预测的方法千差万别，如何构建针对时间序列的预测模型并提高模型的预测效果是本次研究所要解决的主要问题，而本次研究的技术路线也将围绕此问题全面展开。研究团队通过查阅国内外学者对时间序列预测问题研究的相关文献，最终选定通过对以下两种方法的深入研究来解决本次研究的主要问题。

基于 RBM 的神经网络时间序列预测模型：

虽然已有理论证明，三层神经网络在隐藏层神经元足够多的情况下可以以任意精度模拟任意复杂的非线性或线性函数，但是普通 BP 神经网络的预测效果对初始参数存在明显的依赖关系，为了克服这一缺点何希平等人在对 GDP 预测的问题上提出了基于 RBM 的神经网络时间序列预测模型，并取得了较好的效果。其引入 RBM 来初始化输入层与隐藏层之间的权值的做法在本次时间序列预测模型的构建中具有重要的启发意义。

基于 RBM 的神经网络时间序列预测模型的构建方法具体如下：首先，为了优化神经网络的初始权值和阈值，利用引入的受限玻尔兹曼机（RBM）来来学习时间序列的统计特征，即挖掘数据样本服从的概率分布，这一过程为无监督特征学习；其次将 RBM 的编码特征作为预测神经网络模型的输入，将时序值作为网络输出，进一步将网络扩展为回归分析的神经网络；最后利用 BP 算法调整神经网络内部的权值，这一过程为有监督的特征学习。最终达到对由核电站主泵各监测点数据所组成的时间序列进行预测的目的。

基于混沌粒子群改进的神经网络时间序列预测模型：

BP 神经网络的改进目前主要有引入惯性量、动量法、变尺度法和变步长法等几种方法，但是这些改进并没有触及 BP 算法的实质，故改进效果不甚明显。周家林等学者在镗孔加工中的问题提出了基于粒子群算法的神经网络，并提出利用 PSO 算法与 BP 算法相结合这一启发性观点。但是由于 PSO 算法中随机个体的质量往往不能满足算法收敛性的要求，且个体极值和群体极值的更新均带有一定的盲目性，限制了模型的泛化能力。通过查阅文献发现将混沌机制引入 PSO 算法可以显著改善这一问题，故引入了基于混沌理粒子群改进神经网络时间序列预测模型来对核电站主泵各监测点数据所构成的时间序列进行预测。

基于混沌理粒子群改进神经网络时间序列预测模型的构建方法具体如下：首先将表示时间维度的数据依据周期的渐进以及周期内部的递增映射到高维空间，构造一个局部混沌状态，对由数据集中数据构成的粒子群进行混沌初始化，提高算法的收敛速度；其次，在粒子群优化算法（PSO）的全局搜索过程中引入混沌机制，并用改进后的 PSO 算法与神经网络中的 BP 算法相结合以优化其连接的权值和阈值，从而增强神经网络的学习能力，提高神经网络的收敛速度和精度；最后，对模型进行反复的训练、测试、改进，最终利用达到满意度的成熟模型对核电站主泵的工作状态进行预测。最终达到对由核电站主泵各监测点数据所组成的时间序列进行预测的目的。

（六）项目研究进度安排（按照季度编制）

起止时间	研究内容	阶段目标	负责人
2017.9 至 2017.12	对原始数据进行数据清洗	完成数据清洗	刘嘉赓
2018.1 至 2018.3	建立基于 RBM 的神经网络时间序列预测模型	建立模型，利用模型进行预测并调优	刘嘉赓
2018.4 至 2018.6	建立基于混沌理粒子群改进神经网络时间序列预测模型	建立模型，利用模型进行预测并调优	刘嘉赓
2018.7 至 2018.9	建立评价标准，对两个模型的预测效果进行评价	选出最优的模型，并进行预测	刘嘉赓
2018.10 至 2018.12	对相关部门对主泵的检测与维护提出合理化建议	提出合理化建议	刘嘉赓

说明：阶段目标为经过该阶段研究应取得的阶段成果，目标应具体、量化，具有可考核性。

（七）已有基础

1.与本项目有关的研究积累和已取得的成绩

项目组的成员在暑假期间通过对《The Master Algorithm》的阅读对智能算法的整体知识脉络有了较为深入的了解，此外通过参加网络课程、查阅文献等方式目前已经具备了有关神经网络、受限玻尔兹曼机、粒子群算法、混沌理论等本次研究所需的基础知识。

2.已具备的条件，尚缺少的条件及解决方法

由于指导老师可以为本次研究提供实验室的服务器使用权限，目前项目组已经具备了对数据进行大规模计算的条件；此外也已经获得了所需的实验数据，具备了进行本次研究的基础条件。但是目前也存在对于某些理论、算法的细节了解不足的问题，需要通过进一步的学习和实践来对其进行深入的了解。

三、经费预算

开支科目	预算经费（元）	主要用途
预算经费总额	**3000.00**	
办公费	200.00	日常办公用品及其他零星支出等
复印打印费	300.00	复印、打印参考资料等
图书费	600.00	购买项目相关学习所需图书等
出版费	1500.00	出版论文所需要的相关费用等
其他	400.00	购买网络课程等其他学习资料所需的费用等

四、指导教师推荐意见

指导教师签名：　　日期：

五、专家组评审意见

专家组签名：　　日期：

六、学院意见

负责人签名（单位公章）：　　日期：

七、学校意见

主管部门盖章：　　日期