北京志凌海纳科技有限公司

研究开发项目阶段性成果报告

项目名称： 基于AI的超融合运维系统

时 间： 2098年12月31日

**一、项目概况**

SmartX 的超融合软件系统（SMTX OS）旨在为客户提供灵活、高效的基于虚拟化和软件定义技术的 IT 基础架构。客户大多依赖 SMTX OS 提供的虚拟机来运行他们的线上业务系统，或者开发测试系统。无论是支撑业务系统，还是支撑开发测试环境，只要能够准确向客户反映基础架构层的潜在问题，并提供对应的预防处理措施，都具有重大的意义。

举例来说，硬件故障是任何 IT 系统都无法回避的问题。这里考虑磁盘失效这一类型的问题。当一个 IT 系统突发磁盘失效问题时，如果其中的数据没有副本，那么后果就是数据丢失，造成无法挽回的损失。而即使是有副本的情况下，在恢复缺失的副本这个过程中（往往需要数个小时到数天），同样面临读能力下降以及数据丢失风险增大的问题。

观察发现，硬件的部分故障并不是毫无征兆的。通过一些复杂的监测和分析手段，可以在一定程度上预见某些硬件的故障。如果硬件的故障可以被预测，那么运维人员可以提前运用一些运维手段来平滑地替换问题硬件，将突发事件转化为常规事件，降低系统的风险性。

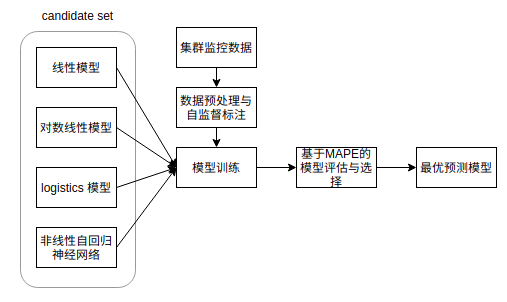
此外，除了预测硬件故障外，对集群资源用量增长的预测也具有一定意义。通常来说，客户业务的增长往往会带动业务系统的扩张，从而需要用更多的 CPU、内存、磁盘等物理资源进行支撑。对这些资源增长趋势的预测能进一步为客户提供准确的集群扩容的建议，避免因基础设施跟不上业务增长而阻碍公司快速发展。

最后，在客户使用 SMTX OS 的实际情况中，不乏一些物理资源分配过多或者过少的虚拟机。前者可能造成一定的资源浪费，而后者可能影响业务系统的良好运行。及时准确地向用户提示这些资源分配不合理的虚拟机可以促进用户提高整个系统的利用率，具有非常积极的作用。

可以看到，一个具有强大分析和预测能力的运维系统对于降低系统运维风险、提高系统效率有重大意义。但是基于人工规则的方法难以实现，而需要使用 AI 技术中流行的机器学习与深度学习方法。

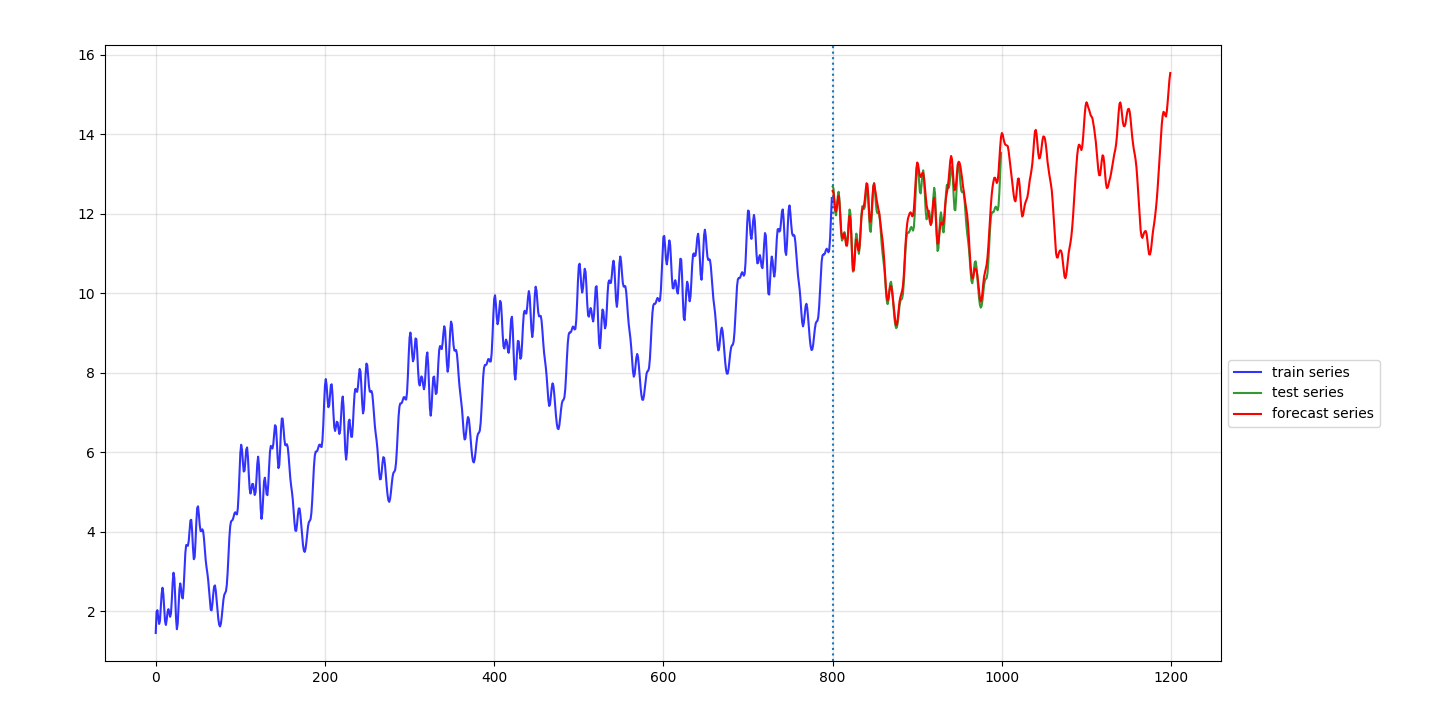
**二、研究成果描述**

2.1 扩容预测

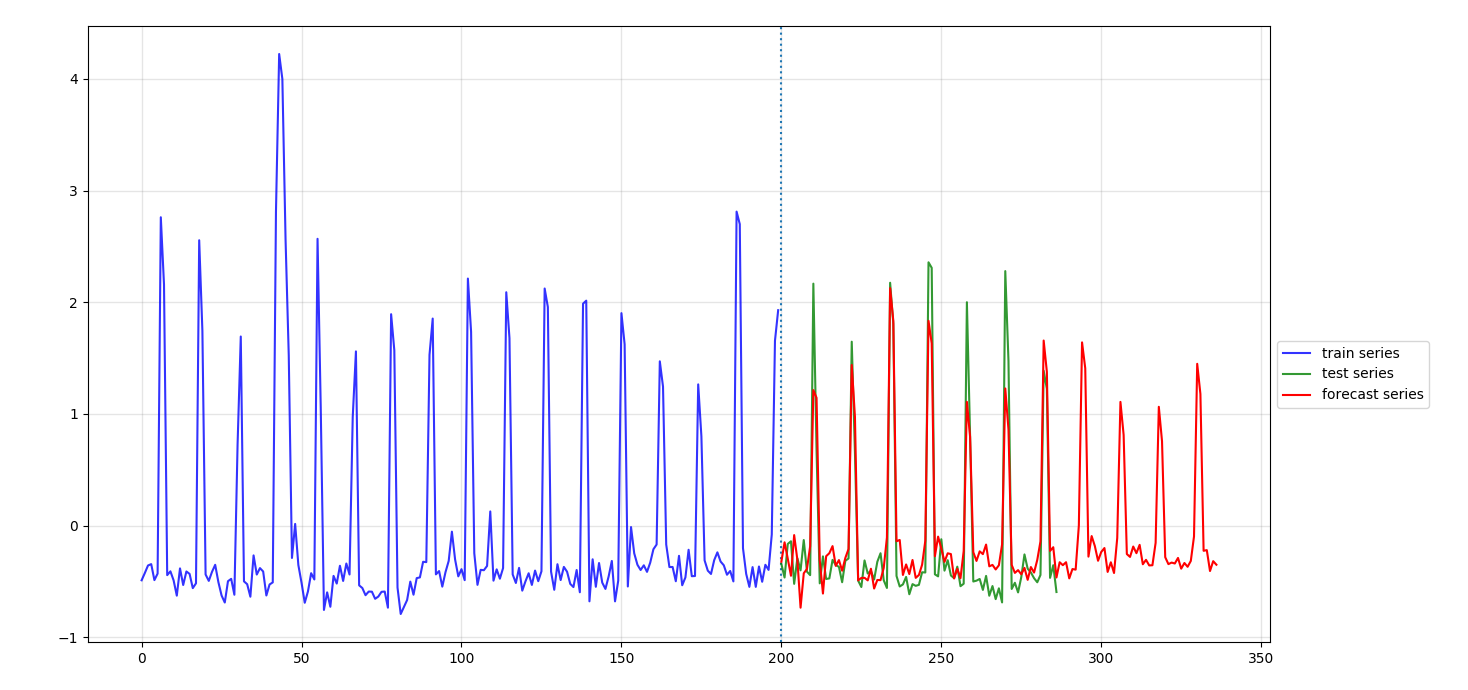


目前为止，candidate set 除原来的线性模型、对数线性模型、logistics模型外，为了预测更复杂的增长模式，还添加基于非线性自回归神经网络。基于这种方法，我们可以构建更多满足复杂场景的 candidate models。从测试看，目前 candidate set 足够满足集群资源预测需求。下面我们提供两个基于非线性自回归神经网络预测效果。

基于神经网络的非线性自回归预测方法的预测效果1:



基于神经网络的非线性自回归预测方法的预测效果2：



上图为基于神经网络的非线性自回归预测方法对超融合集群性能预测效果。可以看到，除了增长模式外，神经网络还可以预测数据中的周期模式以及复杂的非线性模式。由于存在 candidate 机制，不需要担心神经网络过拟合简单的资源增长数据而导致预测无效。因为神经网络一旦过拟合，就会出现较大的验证误差，进而在模型选择阶段被剔除在外。相较于线性类模型，神经网络在训练阶段会消耗一定的时间，但由于训练数据使用滑动窗口对原始时序数据进行自监督标注，通常规模较小，通常能够在 CPU 上秒级别完成训练。而模型推断阶段，能够在毫秒级别内响应。

**虚拟化资源功能的技术路线**

从目前的监控平台看，虚拟资源存在大量监控指标，这些指标表征虚拟资源在一段时间内的使用特点，如资源紧张、资源分配过多、僵尸虚拟机等。我们会使用机器学习中的决策树模型，通过相关的监控指标学习虚拟机资源特点并做出资源状态判断。这种方法无需用户定义决策规则。决策树模型有很好的解释性，且可以把决策规则导出，研发、售后或用户可以根据自身需求修正决策规则。

**虚拟化资源功能的特点**

* 可解释性：用户可以根据决策树的决策规则理解虚拟机的资源使用特点。
* 可定制性：用户可以根据自身的需求或运维经验修改决策树学习到的规则从而定制更个性化的判定规则。

**集群扩容预测功能的技术路线**

集群资源使用情况从长周期来看存在三种增长特点：指数增长、线性增长、带饱、趋势增长。这三类增长特点都可以通过微分方法推导出来，具有很好的场景解释性。

目前从监控平台获得资源使用的时序数据后，我们会使用 candidate models方法，使用机器学习方法把获得的时序数据拟合成多个模型，并在同样的评估指标下选出最好的预测模型，并使用这个预测模型来进行集群资源使用情况的长周期预测。由于所关注的是集群长周期的资源使用情况，如几小时内的等短期内的波动情况不予考虑，但是，仍会评估预测模型的误差，给出误差区间。

**集群扩容预测功能的特点**

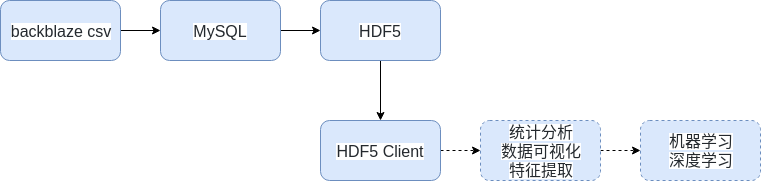
扩容预测中模型选择使用 candidate models 方法，此方法扩展新模型十分便捷。当有新的增长特点，只需要把相应的模型加入到candidate set即可，预测模型会根据数据的特点自动选择最好的增长模型。目前 candidate set 包括线性模型、对数线性模型、带饱和增长的线性模型以及基于自回归预测方法的神经网络。

**磁盘故障预测的技术路线**

磁盘故障分为两部分，第一部分为磁盘（临近）故障分类；第二部分为磁盘剩余寿命预测。这两部分均采用机器学习与深度学习理论与技术实现。

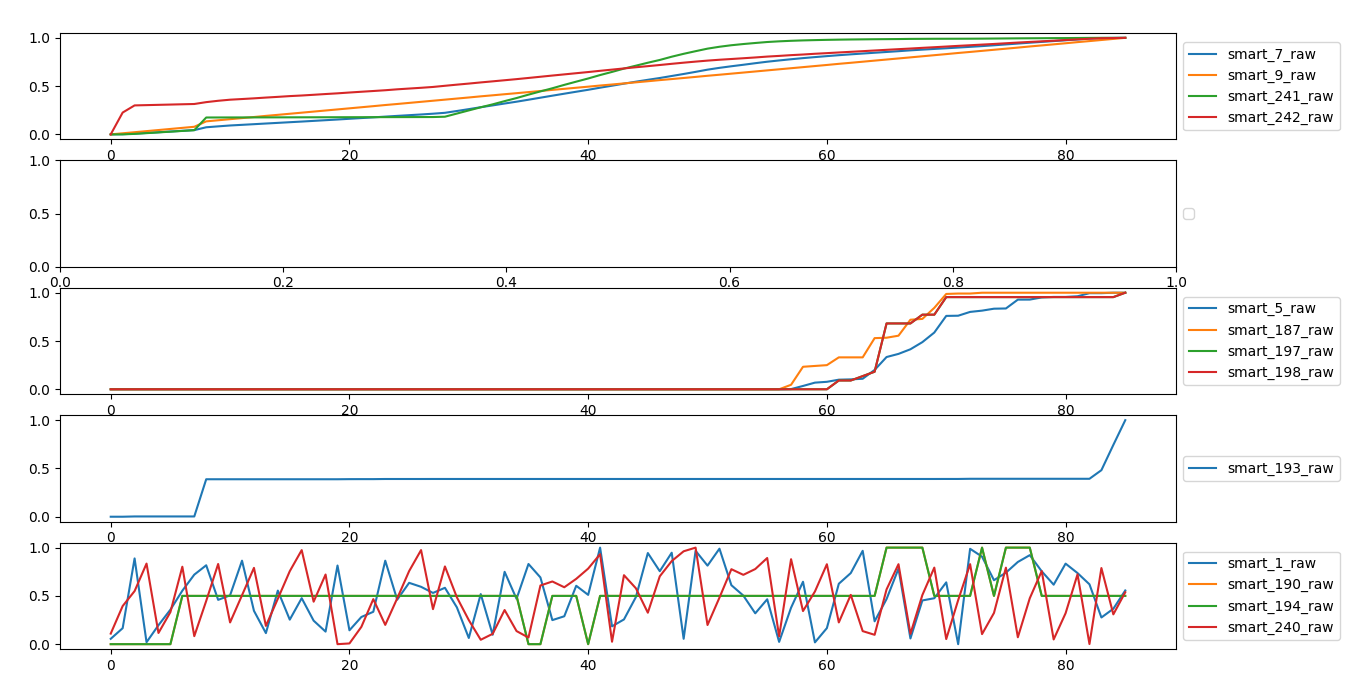
首先叙述磁盘故障分类的技术路线。故障分类的目标为根据磁盘的监控数据判断磁盘是否健康或（临近）故障，并为此定义评估故障分类好坏的性能指标，用于反映机器学习模型的预测效果。从技术和产品的角度看，目前方案使用两个指标：（1）真阳性率（2）假阳性率，分别用于反映磁盘故障预测的检测率和误报率。

定义好评估模型的性能指标后，我们会使用自研的监控平台监控采集磁盘的各种指标，具体分为两大类：一类是磁盘 S.M.A.R.T （Self-Monitoring, Analysis and Reporting Technology）；另外一类是性能指标，包括 IOPS、IO Latency、AML 等等。第一类指标反映磁盘的内部静态特性，第二类指标反映磁盘的外部所处环境的动态特性。

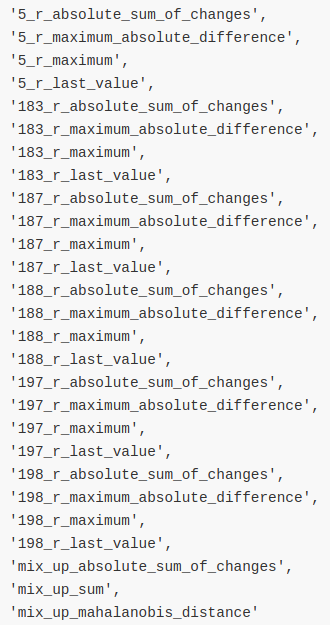


在获得海量数据后，会对数据进行预处理，包括使用插值技巧补充缺失值、数据标准化、数据归一化等等。经过这一系列的处理后，会对数据以磁盘 model 和 serial number 为基准进行基于 HDF5 格式的层次化存储，以方便下一步数据探索、特征工程，也有利于为大规模的机器学习训练构建 pipeline。

在完成数据的预处理后，会获得标准化的数据。为有效进行机器学习任务，我们需要从中提取领域相关的特征。特征工程这一步往往需要大量的领域知识以及对数据特点的探索。下图为我们对某块磁盘的 SMART 数据聚类后的可视化：



可以知道，SMART 多维时序数据可以大致分为三类：持续增长、突变、波动。从现有的数据我们分析到，磁盘故障前部分指标存在突变现象，我们会使用 time series segmentation 和 compact time series representation 方法把包括突变现象的多维时序数据压缩为特征向量，并作为机器学习模型的输入。基于以上分析和对 SMART 数据的理解以及我们已有数据的探索、认知，我们设计了 57 个特征，由于由于篇幅有限，我们列举部分的特征：

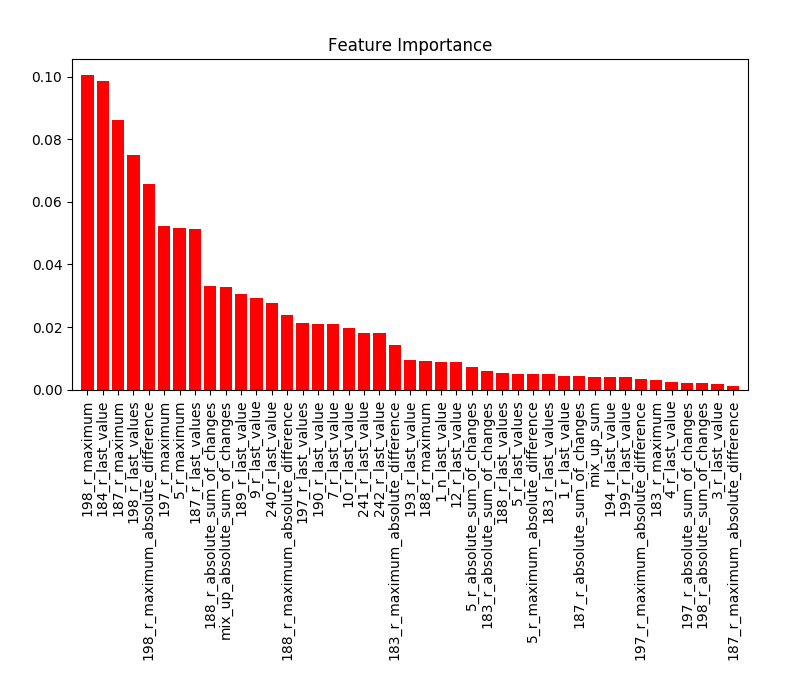


为丰富故障类型的信息和解决磁盘正负样本不平衡问题，一般来说，我们会把磁盘寿命终结前的一段内标注为故障，并以此作为模型的输出。这样，一块故障盘可以构造若干个正类样本。为避免数据泄露，我们严格限制一块磁盘所对应的多个样本要么在训练集中，要么在测试集中。

利用以上方法，我们把原始的监控数据转换到特征空间并获得数据的标签，在此基础上划分训练集、验证集、测试集，这些数据可以直接作为机器学习模型的输入与输出。

在机器学习模型的选择上，我们使用基于树类的集成学习方法。前者会综合上千个预测器的预测结果作为最终的故障预测结果。从偏差-方差分解角度看，这种方法的好处是可以有效降低预测结果的方差。

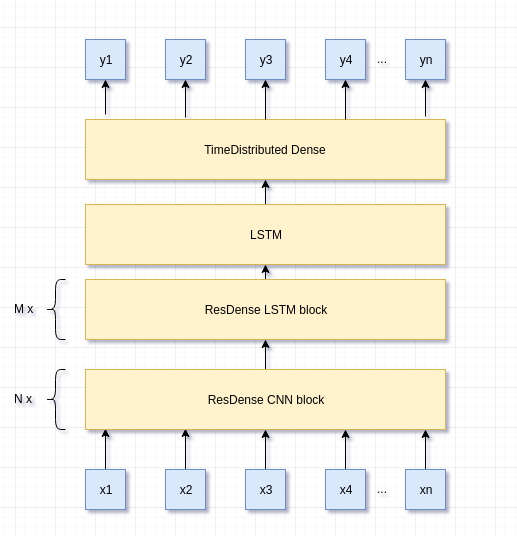
基于机器学习的磁盘故障预测方法，我们做了大量的特征工程，基于训练好的模型评估这些特征的重要性（百分比），其排序如下：



基于以上特征，模型在数据集 ST4000dm000 上的真阳性率为 80%。

在我们训练好预测模型后，会对模型进行评估，当达到我们要求的性能指标后就会集成到我们的产品中，否则进一步优化预测模型。此外，由于磁盘类型繁多，为复用训练好的模型，我们使用 transfer learning 计算。

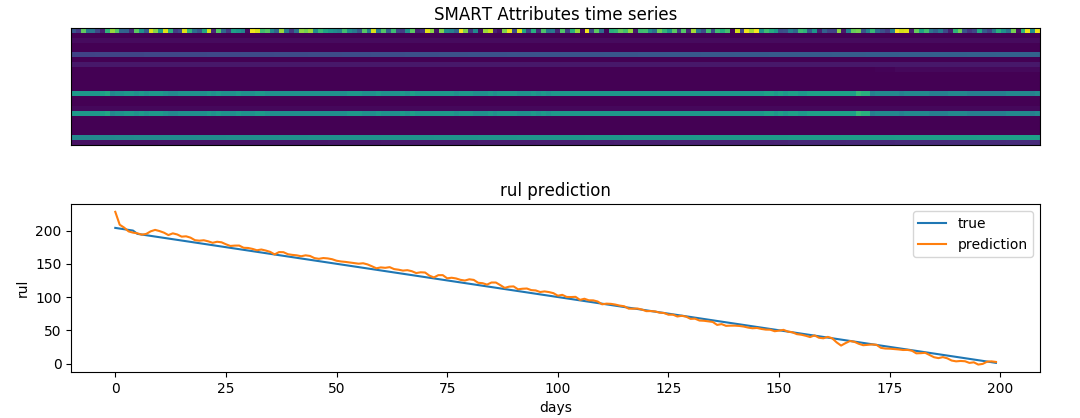
磁盘剩余寿命预测，我们使用深度学习方法，它能为我们提供 end-to-end 的便捷性，一般来说，只要定义好模型的输入与输出，神经网络的网络架构，设计好控制过拟合的方法即可自动进行学习，无需关注特征或专家领域知识。最后，我们会在我们的 GPU 上训练。通常来说，模型训练都设计大量的超参数，这些参数无法让机器学习自动选择最优，为此我们使用 grid search 方法，在 GPU 上搜索最优的超参数。我们构建并训练大量神经网络模型，目前在 ST4000dm000 数据集上效果最好的模型为：



该模型使用 CNN 捕捉局部语义特征，使用 LSTM 捕捉全局特征，为促进信息的流通以及获得稳定的梯度，层与层之间添加 skip connection。为了让模型收敛更快，每层在激活输出前都使用 LayerNormalization 进行归一化。

在模型的设计和训练上，需要考虑如下四点。第一点，数据标注使用线性递减或非线性规则方法对每个 time steps 进行标注，并作为一个回归问题，使用 MAPE 作为 loss 训练神经网络模型。第二点，我们不需要使用大量的领域知识设计各种特征，只需要设计一个足够复杂度的能够捕捉局部和全局特征的神经网络模型。第三点，为避免数据泄露，即模型使用未来才能获得的数据，需要设置掩码机制。第四点，为提高训练效率，我们在数据 pipeline 引入 pre-fetched 技术，当一个 batch 被 CPU 预处理后送到 GPU 上训练时，马上 pre fetch 下一个 batch 数据并预处理放到 cache 中。这样，GPU 就会有源源不断的训练数据，提高 GPU 的利用率，进而缩短训练时间。考虑到 pre-fetched 的瓶颈在存储上，后期我们会尝试把 pipeline 和我们的分布式存储整合在一起。

模型训练好后的预测效果，以磁盘类型 ST4000dm000 为例，预测剩余寿命预测效果如下：



图（上）为磁盘生命周期中的 SMART 数据的可视化。图（下），蓝色线为基于线性递减的磁盘剩余寿命的标注，橙色线为模型的预测效果。可以看到，刚开始（10天内）的预测误差比较大，因为还没有采集到足够的数据，随着数据的采集，预测越来越好。目前预测的平均误差为 7 天。

**磁盘故障预测的特点**

磁盘（临近）故障预测可以在磁盘实际故障前一段时间把磁盘健康状态预测出来，这可给用户足够的时间进行数据迁移与安排换盘计划。机器学习训练出来的模型可以以极小的资源消耗部署到超融合环境下。在实际场景下，磁盘类型（model）非常多，我们先使用一个大型的数据集 pre-training 一个通用的模型，然后在目标磁盘类型的预测上进行 fine-tune。只要目标磁盘类型的数据量足够，我们的模型就可以表现较好的预测效果。如果某个数据中心需要支持某类磁盘的预测，则前期需要采集一定量的数据以便在 pre-training 模型上进行 fine-tune。

磁盘剩余寿命预测可以预测磁盘在当前环境下（集群负载）还能用多久（以天为单位），对于大型数据中心来说，有足够的时间制定数据迁移、换盘与磁盘采购计划。

**三、项目实施情况**

3.1 执行期

|  |  |  |  |
| --- | --- | --- | --- |
| **阶段** | **主要工作内容** | **时间** | **备注** |
| 阶段一 | 项目启动 | 2018.1.1 | 已完成 |
| 阶段二 | 需求分析说明 | 2018.1.1 –3.5 | 已完成 |
| 阶段三 | 概要设计 | 2018.3.6 –4.10 | 已完成 |
| 阶段四 | 详细设计 | 2018.4.11 –6.30 | 已完成 |
| 阶段五 | 软件开发与单元测试   * 虚拟化资源功能开发与测试 * 集群扩容预测功能开发与测试 * 磁盘故障预测功能开发与测试 | 2018.7.1 – 2019.11.1 | 进行中 |

3.2 关键技术及技术指标

* 机器学习与深度学习。模型的训练与模型的选择均使用机器学习、深度学习相关方法。
* 使用 MAPE 或 MSE作为多个模型的统一评估指标。
* 基于统计学的误差估计。

**四、成果的创新性与先进性**

* 使用 candidate models 方法，不再局限在单一模型上，这种方法非常容易扩展，只需要把新模型添加到 candidate set 中即可。
* 模型训练时消耗的CPU、内存非常小。模型推断响应快。

**五、人才培养**

通过本项目的实施，公司融合了一支精诚合作、团结进取的优秀研发团队。通过创新人才培养方法，充分发挥了管理人才和技术人才的作用，提高了研发能力和水平，造就了一支创新型研发人才队伍。建立了研发人员绩效考核办法，以公司制定的新产品开发计划为基础，以公司新产品开发目标和经营目标为导向，年终结合目标责任制要求进行绩效考核，根据绩效考核情况最终确定其报酬，并设立薪资奖励、职务晋升和特殊业绩奖励等多种奖励机制，稳定了技术人员队伍，为公司发展打下了坚实的基础。

**六、技术条件**

公司的员工通常拥有本科及以上学历，以及在国内外知名企业的工作经验，其对相关产品认识深刻，具有相关应用行业的丰富开发经验。公司建立了完善的培训机制，包括外部培训和内部培训。公司结合已完成的项目，建立了关键软件模块化库以作为资源共享库，为技术工程师的学习和交流提供了平台。

公司技术研发部成立了项目小组，按照项目管理的要求，依据管理规定、管理办法、作业标准三级管理层次，落实并执行项目立项、项目计划、设计编程调试、测试验收等管理办法，各节点均有检查和验收。

**七、资金使用情况**

项目投入经费 万元，详见《研发项目加计扣除研究开发费用情况归集表》。