机器学习服务框架－－MLaaS

王汉杰M201676090

December 6, 2016

摘要

本文以ICML会议在2015年发表的论文《MLaaS: Machine Learning as a Service》为主要学习内容，初步了解深度学习作为服务框架的设计和实现过程。这篇论文首先介绍和分析了深度学习、SCA 架构等相关技术；其次设计了基于SCA （Service Component Architecture）规范的深度学习服务框架；最后以电力需求量预测这一现实问题作为实例，对文中提到的内容做了实现和对比测试，基本实现了设计预期的功能。

本文大致按照原论文的思路进行展开，查找有关论文中涉及到却没有展开的内容的相关资料，对文章的内容进行梳理和总结，将主要分析原论文中应用到的已经学过的理论和技术，以便进行知识的迁移和转化，便于认识和理解；此外在最后对论文和相关的体会认识进行总结整理。

关键字：深度学习服务框架,监督学习,回归预测,SCA规范

1 绪论

随着移动互联网的普及和物联网技术的蓬勃发展，数据获取的途径越来越多，数据量也愈来愈大。对于各种设备产生的大量数据做出有效而正确的处理成为当代信息技术发展的重要方向之一。多种深度学习技术已经用于从海量数据中提取有用信息，并且取得很好的成果。但是，深度学习需要大量资源作为支撑，这对于大公司而言还可以承受得起，但是中小公司在运用这一技术时却面临着诸多困境。

为此，西安大略大学电子与计算机工程系的三位研究人员Mauro Ribeiro, Katarina Grolinger, Miriam A.M.Capretz于2015年在ICML 会议上发表了题为《MLaaS: Machine Learning as a Service》的论文（以下简称“论文”）。论文中分析了深度学习技术服务化的需求，对深度学习、SCA等技术做了一定的梳理分析，在此基础上设计了MLaaS机器学习服务框架，并对框架进行了模块分析和流程简介，最后以电力需求预测这一问题作为实例进行分析、实现和测试。最后得出MLaaS基本实现了预想的功能需求。而这篇文章则是作为原论文的阅读笔记在对论文的内容进行充分解析的基础上，对文中内容进行分析和梳理。

本文的内容结构包括以下几个部分：第2节简要解析论文中的相关背景知识；第3节概述分析MLaaS框架；第4节结合实例进行测试分析；第5节总结全文。

2 相关背景

本节重点内容在于对论文进行解析和总结。主要内容将会按照如下的两小节展开。

2.1 深度学习

深度学习是近年来计算机科学内发展非常迅猛的领域之一，它主要是运用数学建模和统计模型根据数据样本集做出推断，因为其需要自适应与变化的环境，所以也作为人工智能的一个分支。 机器学习如图1所示的可以分为三类：有监督的学习，无监督的学习和增强学习，而有监督的学习指的是训练集是有标签的（意即包括模型试图估计的标签），而标签按照离散和连续又可以进一步分为分类和回归两种[1]。其中本文在MLaaS框架的框架实现和实例分析中主要针对的是回归预测分析。

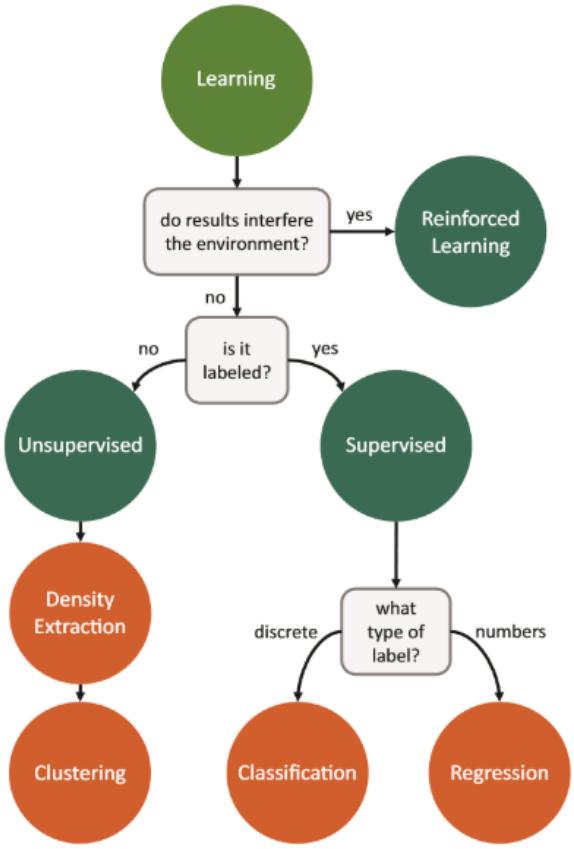

图 1: 机器学习方法分类

无论机器学习技术有多少种分类，其大致运行过程总是相似的[1]：模型从训练集上获取信息，对新的集合做出推断。这种共性为我们创建可以运行多种不同算法的统一框架提供了可能。本文主要的关注点在于为回归预测模型设计支持多算法的统一框架。预测模型有两个关于准确性的要点需要把握：选择正确的算法和输入参数；对可能的预测错误做出估计。常用的用来评估准确率的技术是K阶交叉检验[2]。K阶交叉检验就是把原始的数据随机分成K个部分。在这K个部分中，选择一个作为测试数据，剩下的K-1个作为训练数据。交叉检验的过程实际上是把实验重复做K次，每次实验都从K个部分中选择一个不同的部分作为测试数据（保证K 个部分的数据都分别做过测试数据），剩下的K-1个当作训练数据进行实验，最后把得到的K 个实验结果平均。根据K阶交叉检验，我们可以对预测的结果进行对比分析。理论分析和实践表明，不同的算法对于不同的问题而言其准确性是不同的，因此为用户提供统一的、可适配多种算法的统一化平台是十分有意义的。

单就机器学习的思路而言，其与数理统计的相关理论知识存在着某种相似之处：都是从一定的样本数据中分析某种规律，创建一种相应的运算模型，之后再将其运用

到其他的数据集上，得出相应的统计推断并对误差做出估计。只是数理统计中规律的发现、总结和模型的构造都是由人力完成的，而机器学习则是要将“归纳总结”这样的过程交给计算机来完成。

2.2 SCA规范

根据IBM DeveloperWorks中对于SCA的解释[3]，SCA是一组用于构建和实现基于SOA（Service Oriented Architecture）应用系统的编程模型。因而SCA架构也具有SOA架构的优点。在论文中将SCA的实现简要划分为：组件、组合和服务。其中组件用以实现业务逻辑；组合则是组件的联合，已构成业务解决方案；而服务则是这些业务逻辑的远程调用接口。

深度学习发展到现在的阶段，自然也形成了多种体系框架，比如PredictionIO框架、Baldominos框架、OpenCPU框架等[1]，这些框架都因为其构建在一种具体的分析工具下而受到限制，这对于添加关于数据存储、服务部署的新算法而言非常不灵活。因此论文中构建的MLaaS框架就是基于更为一般通用的SCA规范，该框架强化了对新算法的适用性，具有更好的通用价值。

3 MLaaS框架

这部分就论文中的架构设计部分进行简要的介绍，MLaaS框架目的在于支持从多数据集中，应用多种算法，构建对应的机器学习模型。文章以回归预测这一机器学习的分支为主，但相关的理论和思想并不局限于此，可适用于其他方面的应用中。

3.1 模块分析

图2中给出了MLaaS框架的结构图，从图中可以看出其模块比较多，因此我们此处仅以主要流程涉及的模块进行简要介绍。

Data Gatherer组件用以接收和预处理数据，同时将数据转发至Modeler 组件。其三种功能是流水线式的并行运行的。Modeler组件用以构建预测模型，一个预测模型是由一种特定算法生成的Modeler-μ的示例。MLaaS 中可以同时运行多个Modeler-μ对象，每个对象可以拥有不同的设置。框架中存在的多个Modeler-μ组件对应多个算法。其更为详细的耦合关系在下一小节中详细描述。

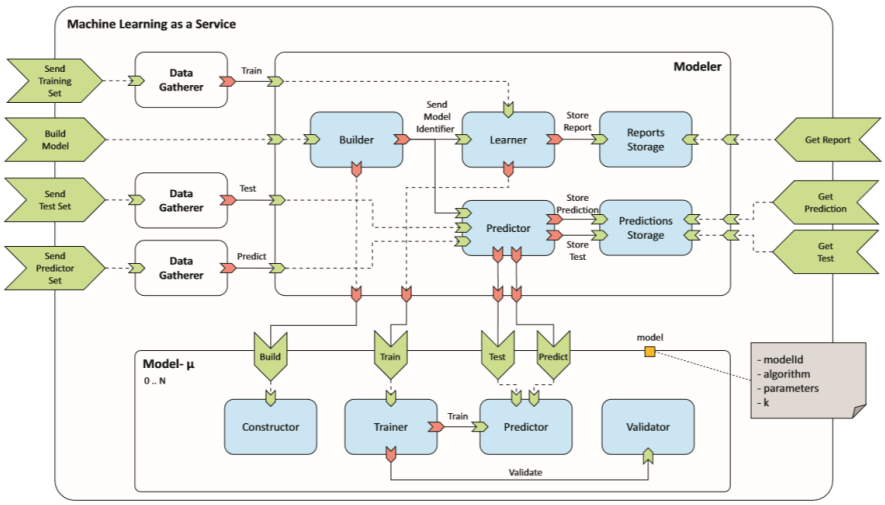

图 2: MLaaS架构模型

框架的大致功能流程可以概述为三个阶段：建模过程、训练过程和预测过程。其中建模过程指的是Modeler组件中的Builder模块将会根据用户命令创建和配置一个Modeler- μ 对象。训练过程则是由Send Training Service为框架获取原始数据。由Data Gatherer组件进行对其进行预处理，其后Modeler组件根据处理好的数据训练Modeler- μ 实例对象，并对模型的误差做出估计。预测过程则是由用户通过Send Predictor Service输入待预测的数据集并指定某种特定的模型，对应的Modeler- μ 对象则根据待预测的数据集做出推断预测，并通过Get Prediction Service将结果传递到外部模块。

3.2 框架总结

从上文中MLaaS的架构探讨中我们可以将其较为突出的特点总结如下几个方面：①分离思想。显而易见，框架内包括大量较为独立的模块，模块间的交互多以消息传递的形式进行，这本身就使得模块间处于非常松散的状态。而每个模块的功能也非常单一，这非常相似于面向对象思想中的单一职能原则，这种设计使得以该框架为基础的系统拥有更好的可维护性。②抽象接口思想。框架中提供了统一的算法接入接口，这样可以灵活动态的加载多种不同的算法（只要实现系统中提供的统一算法接入接口即可），这种面向接口的编程就各种算法而言拥有更高的灵活性，而灵活切换算法正是这次开发实现的核心目的。③分布式思想。框架是基于SOA的架构思想构建的，本身就考虑到了分布式应用得需求，而框架本身的设计又有着较大的并行性，可以很好的支持多个算法并发计算运行。这在当前基础数据集越来越庞大的今天是十分必要

的，可以更好的满足性能需求。

4 实例测试

接下来论文中就电力能源需求进行实例分析，应用不同的算法以判断一种最佳算法，具体的数据和实现情况就不再赘述，此处仅就实验中使用的算法和结果进行简要介绍：①MLP（Multi-Layer Perception）：常用的机器学习算法之一，尤其在电力消耗问题方面有广泛应用。②SVR（Support Vector Regression）：也是电力消耗问题领域内广泛应用的算法。③KNN （K-Nearest Neighbors）：一种易实现易理解的机器学习算法[1]。实现过程中只有最后一种算法是论文作者自行编码实现的，前两者均是直接获取的第三方代码库。正如上文所述，框架提供的统一接口使得无论是第三方的程序包还是自行编码实现的算法都可以很方便的“挂载”到当前实现的系统上。同时在进行测试之后，看到各个算法之间可以实现并发运行。

根据实验测试的结果，我们可以分2个指标来探讨：（1）准确性：SVR拥有更优的准确性，而另两种算法都存在相对较大的误差。（2）执行时间：在确认阶段KNN执行最快，而测试阶段MLP执行最快。当然就该篇论文而言，到底那种算法更优并不是讨论的重点，而是框架实现了预期的功能并为用户提供了这些常用的性能指标（统计图表输出），可以供用户进行参考和决策。

5 结论

至此，从论文中我们可以有以下三点结论：（1）文中提供了一种新颖的基于SCA规范的预测建模机器学习服务框架。该框架支持配置多种算法从多种数据集中独立并行的获取预测分析结论。（2）文中根据该框架实现了一个电力需求预测系统。主体框架和一种算法实现后，其他算法的实现和添加就十分简便了，而框架本身提供了一些指标对算法进行衡量。（3）文中最后指出这种框架还可以进一步拓展到其他机器学习领域内。

通过对论文的阅读和这篇总结的写作，我有以下几点体会：首先，更多的了解了有关机器学习的内容，虽然只是很粗浅的涉猎，对其间理论和实现技术还有很多不了解，但是还是深深感受到了深度学习的优点，通过对机器进行训练可以更好、更自动的对数据进行处理得出有用的信息或结论，这在当今数据爆炸的时代毫无疑问是非

常有意义的。其次，论文的主题除了深度学习之外，还包括架构设计的相关问题，从文中可以看出一个优良的软件架构有着非常的优越性，组件思想的应用降低了彼此间的耦合关系，提供了更好的拓展性，这在软件开发过程中是十分值得思考和认真对待的。最后，英文文献中的逻辑性和严密性十分值得我们进行学习，不仅在文档写作方面，还在日常的学习工作中都应该保持好的逻辑性。

参考文献

[1] K. G. Mauro Ribeiro and M. Capretz, "Mlaas: Machine learning as a service," ICMP, 2015.

[2] "K阶交叉检验." http://blog.sina.com.cn/s/blog_688077cf0100zqpj.html.

[3] "Sca专题." http://www.ibm.com/developerworks/cn/webservices/lp/sca/.