白皮书

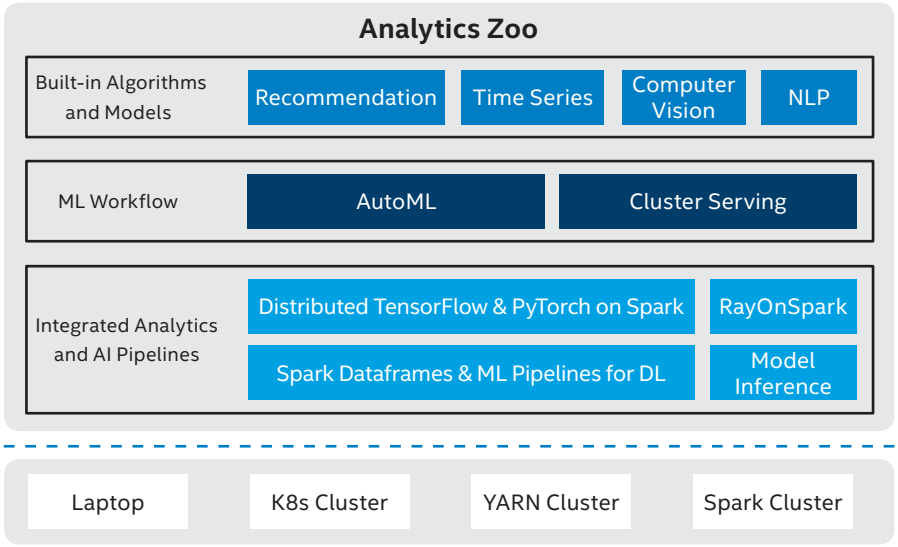
Analytics Zoo 助力腾讯云提升智能钛机器学习平台性能

Tencent 腾讯

目录

前言

英特尔® Analytics Zoo 是一个统一的大数据分析和人工智能平台，它可以将 Tensorflow、Keras、PyTorch、Spark、Flink 和 Ray 程序等集成在一个统一的流水线中，并且可以从笔记本环境透明地扩展到大规模集群，对生产数据进行处理。图一所示是 Analytics Zoo 的总体架构。



图一 Analytics Zoo 架构图

通过使用 Analytics Zoo，用户可以轻松地创建端到端的人工智能应用，并加以部署，例如在 Spark 程序中书写 TensorFlow 或者 PyTorch 代码，并进行分布式的训练和推理；或者通过 RayOnSpark，在大数据集群中直接运行 Ray 的应用程序。通过使用 Analytics Zoo，用户可以使用高级机器学习流水线，来实现大规模机器学习应用程序开发过程的自动化，例如用户可使用自动分布式的 Cluster Serving 来进行 TensorFlow、PyTorch、Caffe、BigDL 和 OpenVINO™ 模型的推理；或者通过可扩展的 AutoML，来进行时序数据的预测。此外，Analytics Zoo 还提供了用于构建推荐、时序数据、计算机视觉和自然语言处理程序等不同应用场景的各种算法和模型。

在典型的机器学习应用中，用户一般要先通过恰当的数据预处理、特征工程、特征提取和选择等，使数据集能够有效地被机器学习应用所使用；数据预处理之后，用户还必须

使用恰当的模型算法，以及通过超参数优化，来使得机器学习模型和算法的预测性能最大化。很显然，这些步骤都极具挑战性，使得一般人难以利用机器学习的技术。

自动机器学习 (AutoML) 是一种把机器学习应用到解决真实世界问题的自动化手段，它涵盖了从原始数据处理到可部署的机器学习模型的整个流程。由于机器学习的应用需求不断增长，AutoML 被认为是应对这一问题的一种人工智能解决方案。高度自动化的 AutoML 使得非专业人士也可以利用机器学习模型和技术，而不必先成为这一领域的专家。通过应用端到端的自动机器学习技术，用户能够获得类似人工智能的解决方案，更快地构建这些解决方案，并且这些方案大多在性能上还可超越人工调试的模型。

腾讯与英特尔合作推进 AutoML 项目

腾讯云智能钛机器学习 (TI Machine Learning) 是基于腾讯云强大计算能力的一站式机器学习生态服务平台。它能够对各种数据源、组件、算法、模型和评估模块进行组合，使得算法工程师和数据科学家能够在其之上方便地进行模型训练、评估和预测。腾讯云智能钛机器学习平台 (TI ONE) 支持多种计算框架，例如 PySpark、PyTorch、TensorFlow 等。

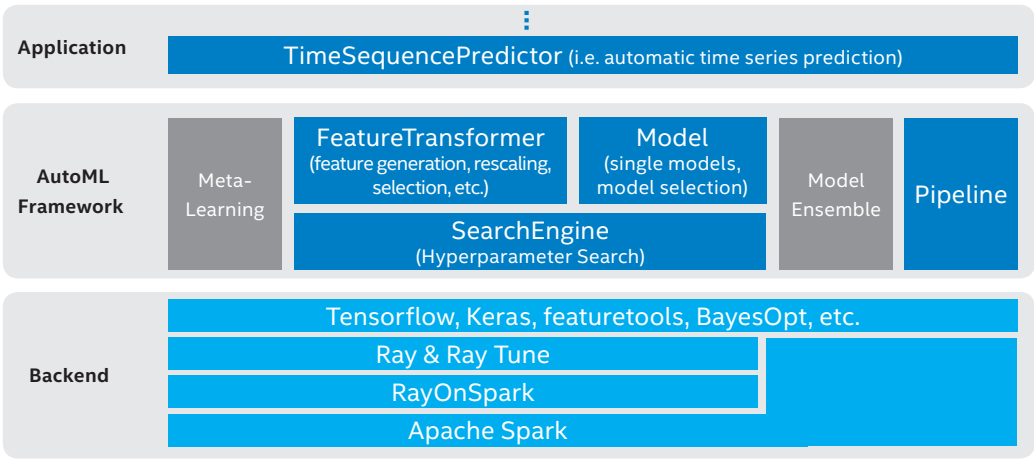
英特尔与腾讯的机器学习团队通过深度技术合作，将 Analytics Zoo 集成到腾讯云智能钛机器学习平台，使该平台获得了更强大的 AutoML 特性，让 AI 初学者也能轻松使用。使用 Analytics Zoo 的 AutoML，可以很方便地进行时间序列分析，如时序预测，异常检测等。

时间序列数据，顾名思义，是按照时间顺序收集的一系列数据。时序数据预测是指用过去的时间序列数据作为输入，来预测未来时间的数据值。在很多真实的应用场景，例如网络运营商的网络质量分析、数据中心运维的日志分析、高价值设备的预防性维护等，都可以利用时序数据预测的技术。时序数据预测还可以作为异常检测的起点，使其在真实值和预测值偏差较大时触发警报。

经典的时序数据预测通常使用描述性模型或者统计方法。这些方法经常需要对数据的分布进行假设，并且要做时序分解 (把时序数据分解为周期、趋势、噪声等成分)。相比较而言，基于机器学习的时序数据预测方法 (例如基于神经网络的模型) 对数据几乎不做假设，比经典的线性预测模型 (ARIMA, ES等) 对复杂模式的识别表现更好。实际上，神经网络模型在时序预测领域已经有了不少成功的案例。为时间序列数据预测构建机器学习应用是一个需要大量专业知识的费时费力的过程，Analytics Zoo AutoML 框架实现了自动化的特征生成和选择，模型选择，和超参调优的功能，可以使得训练时序分析模型的过程更加容易。

Analytics Zoo 中基于 AutoML 的时间序列数据预测工具构建于 Ray 和 Ray Tune 之上。Ray 是一个由加州大学伯克利分校 RISE 实验室开源的分布式计算框架，用于开发新型的人工智能应用；而 Ray Tune 则是一个在 Ray 之上运行的可扩展超参数优化库，用户可以在一个大规模集群上高效地进行很多实验。Analytics Zoo 支持 RayOnSpark，允许用户在已有的大数据集群上直接运行基于 Ray 的各种新兴人工智能应用，并且可以无缝整合到大数据处理和分析流水线中。

下面我们将描述如何使用 Ray Tune 和 RayOnSpark，来实现 AutoML 框架和自动的时间序列数据预测。图二所示是 Analytics Zoo 中的 AutoML 框架。



图二 Analytics Zoo 中的 AutoML 框架

AutoML 框架利用 Ray Tune 在 RayOnSpark 上进行超参数搜索，已实现的超参数搜索涵盖了特征工程和建模。在特征工程中，搜索引擎从各种特征生成工具 (比如 featuretools) 自动生成的特征集合中选择最佳的子集；在建模中，搜索引擎搜索超参数，例如每层的节点数、学习率等。在本项目中，使用流行的深度学习框架，例如 TensorFlow 和 Keras，来构建和训练模型，并使用 Apache Spark 和 Ray 来进行分布式的运行。

AutoML 框架目前包括四个基本组件，即 FeatureTransformer、Model、SearchEngine 和 Pipeline。Feature Transformer 定义了特征工程流程，其通常包括一系列操作，如特征生成、特征缩放和特征选择；Model 定义了模型 (如神经网络) 和使用的优化算法 (如 SGD、Adam 等)。此外，Model 还可能包括模型/算法选择；SearchEngine 负责搜索 FeatureTransformer 和 Model 的最佳超参数组合，控制实际的模型训练过程；Pipeline 则是一个集成了 FeatureTransformer 和 Model 的端到端的数据分析流水线，Pipeline 可轻松保存到文件中，方便后续加载重新使用。

利用 AutoML 框架训练模型的一般流程包括以下步骤：

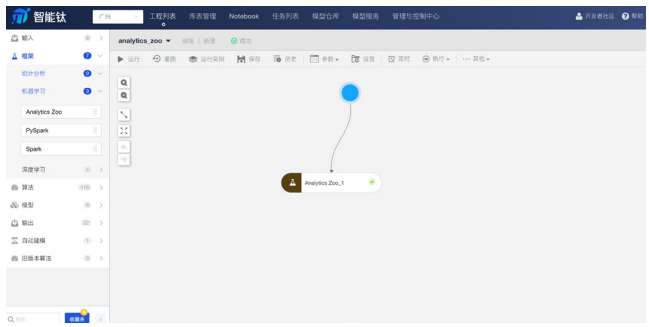
1. 首先实例化 FeatureTransformer 和 Model，随后对 Search Engine 进行实例化，并由 FeatureTransformer、Model 及一些搜索预设(指定超参数搜索空间、奖励指标等)进行配置。

2. SearchEngine 运行搜索程序。每次运行将生成多个试验，并使用 Ray Tune 在集群中分布式运行这些试验。每个试验使用不同的超参数组合完成特征工程和模型训练流程，并返回目标指标。

3. 在所有试验完成后，可根据目标指标检索出一组最佳的超参数，并得到训练好的模型，用于创建最终的 FeatureTransformer 和 Model，以及构成 Pipeline。Pipeline 可被保存至文件中，以便通过后续加载用以推理和/或增量训练。

Analytics Zoo 提供了一个方便的接口 TimeSequencePredictor，将上述 AutoML 框架的一般步骤加以封装，集成了大量时序相关的特征处理和模型，专门用于时序预测模型的训练。用户可以直接调用这个接口进行自动化的时序预测模型训练，输出 Pipeline 和保存，方便后续进行预测、部署以及增量训练更新。

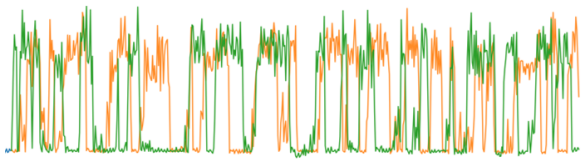
腾讯云智能钛机器学习平台 TI ONE 已经整合了 Analytics Zoo 组件，如图三所示。有兴趣的用户可以使用智能钛机器学习平台的 Analytics Zoo 组件，进行时间序列数据的分析以及机器学习建模。



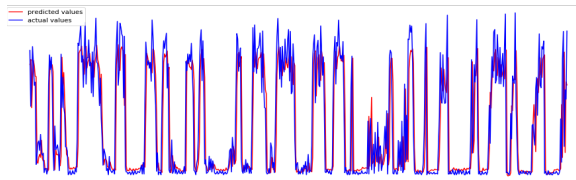
图三 整合了 Analytics Zoo 的腾讯云智能钛机器学习平台 TI-ONE

Analytics Zoo AutoML 的优势

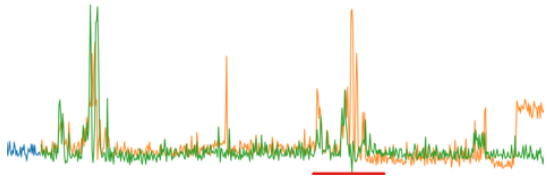
基于 Analytics Zoo 的 AutoML 不仅可以实现特征生成、模型选择和超参数调优的过程自动化，而且由其训练生成的模型准确率通常会超越传统方法或者手工调优措施。如下图第一组对比，当时序数据中的周期长度不是特别规则的时候，使用传统方法进行时间序列数据预测会产生较大偏差，而基于 Analytics Zoo AutoML 的预测值和实际值吻合度较高。如下图第二组对比，传统方法对时间序列数据中峰值不是特别规则的情况预测偏差较大，而 Analytics Zoo AutoML 的模型的吻合度相对较高。



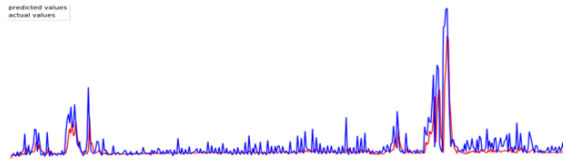
图四 传统方法的时间序列数据预测



图五 Analytics Zoo AutoML 的时间序列数据预测



图六 传统方法的时间序列数据预测



图七 Analytics Zoo AutoML 的时间序列数据预测

在腾讯云智能钛机器学习平台上使用 AutoML 检测数据异常

下面将通过一个案例介绍 Analytics Zoo AutoML 在腾讯云智能钛机器学习平台上的使用方法。

腾讯云将先进的第二代英特尔® 至强® 可扩展处理器应用于腾讯云智能钛机器学习平台。第二代英特尔® 至强® 可扩展处理器支持英特尔® 深度学习加速 (Intel® Deep Learning Boost) 技术，极大提升了人工智能负载，特别是深度学习负载性能。Analytics Zoo 通过利用英特尔® MKL-DNN 的优化和加速，高度释放了英特尔® 至强® 可扩展处理器的模型训练和推理性能。

在腾讯云上创建基于英特尔® 至强® 可扩展处理器的实例步骤如下:

新建实例 - 自定义配置 - 选择地域与机型 - 2 核 (Cascade Lake) 4GB 或更优配置

图八 在腾讯云上创建基于英特尔® 至强® 可扩展处理器的实例

Analytics Zoo 提供了一个 notebook 例子，将 AutoML 用于时间序列数据的异常检测。它将历史数值作为模型的输入来训练模型，然后使用训练好的模型预测下一个数据点。当实际值与模型预测值相距较大时，定义为异常。

这个案例使用了 Numenta Anomaly Benchmark 的一个数据集 (NYC taxi passengers) 来示例。该数据集包含 10,320 条样本，每条样本表示特定时间纽约市的出租车乘客总数。数据格式如下所示:

```
timestamp,value
2014-07-01 00:00:00,10844
2014-07-01 00:30:00,8127
2014-07-01 01:00:00,6210
2014-07-01 01:30:00,4656
2014-07-01 02:00:00,3820
2014-07-01 02:30:00,2873
2014-07-01 03:00:00,2369
2014-07-01 03:30:00,2064
2014-07-01 04:00:00,2221
```

在运行案例之前，需要先下载数据，解压压缩包，并将数据文件 nyc_taxi.csv 上传到 cos 上去。

以下是使用 AutoML 训练时序模型的关键步骤:

首先用必要的参数初始化一个 TimeSequencePredictor 对象，然后调用 TimeSequencePredictor.fit，以分布式的方式对历史数据自动地进行机器学习训练，在训练结束后得到一个 TimeSequencePipeline 对象。

```
from zoo.automl.regression.time_sequence_predictor import
TimeSequencePredictor
tsp = TimeSequencePredictor( dt_col="datetime",
                             target_col="value",
                              extra_features_col=None,
                             future_seq_len=1)
pipeline = tsp.fit(train_df,
                   metric="mean_squared_error",
                    recipe=RandomRecipe(num_samples=100),
                   distributed=True)
```

TimeSequencePredictor 的输入数据 (train_df) 是包含一系列记录的 (Pandas) Dataframe，每条记录包含一个时间戳 (dt_col) 及与时间戳关联的数据点数值 (target_col)，每条记录还可包含额外的输入特征列表 (extra_feature_col); TimeSequencePredictor 训练完成之后得到 TimeSequencePipeline，用于预测未来时步的相应 target_col。

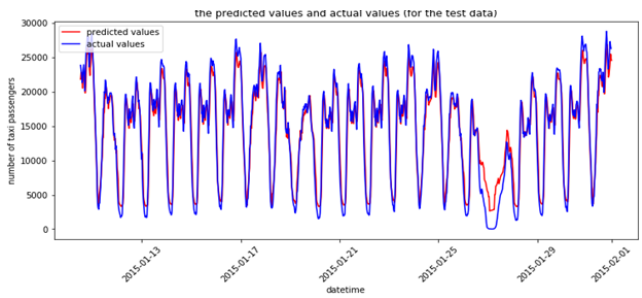
recipe 参数包含 TimeSequencePredictor 所需的参数，用于在训练时指定搜索空间、停止条件和样本数量 (即搜索空间中生成的样本数量)。目前可用的 recipe 包括SmokeRecipe、RandomRecipe、GridRandomRecipe 和 BayesRecipe。

可以将训练结束时获得的 TimeSequencePipeline (已包含最佳超参数配置和 AutoML 框架返回的训练好的模型) 保存至文件中，并在后续对其进行加载，用于评估、预测或增量训练，具体细节如下所示。

```
pipeline.save("/tmp/saved_pipeline/my.ppl")  #save

from zoo.automl.pipeline.time_sequence import load_ts_pipeline
pipeline = load_ts_pipeline("/tmp/saved_pipeline/my.ppl")  #load
rs = pipeline.evaluate(test_df, metric="r_square"])  #  evaluation
result_df = pipeline.predict(test_df) # inference
pipeline.fit(newtrain_df, epoch_num=5)  #  incremental training
```

下图使用 AutoML 展示了下一个时步的预计出租车乘客量。



图九 使用 AutoML 展示下一个时步的预计出租车乘客量示例

总结

英特尔与腾讯的机器学习团队通过深度技术合作，将 Analytics Zoo 集成到腾讯云智能钛机器学习平台，使该平台获得了 AutoML 高级特性，让 AI 初学者也能轻松使用。利用已整合 Analytics Zoo 的智能钛机器学习平台，可对各种数据源、组件、算法、模型和评估模块进行组合，使得算法工程师和数据科学家在其之上能够方便地进行模型训练、评估和预测。

目前智能钛系列产品支持公有云访问、私有化部署以及专属云部署。

参考材料

https://github.com/intel-analytics/analytics-zoo

https://cloud.tencent.com/product/ti