分布式计算框架的大数据机器学习探析

□ 邱莉萍 鞠海军 龚晓敏 邓拓 刘林玉

摘要：为了实现大数据挖掘多样性以及实时性的要求，本文利用分布式计算框架设计了机器学习系统，实现大数据的迭代计算。根据模型的建立过程，对数据迭代计算分为粗调以及细调阶段，通过对不同迭代过程中对不同样本对模型参数向量的影响程度不同，形成了优化算法。从而形成迭代过程的简化，使计算过程效率进一步提高。按照本文设计的优化算法，进行了分布式计算框架机器学习模型的建立及验证，结果表明本文设计的分布式计算框架的大数据机器学习方式能够提高训练精度，提升了大数据挖掘的使用性。

关键词：分布式计算框架；大数据；机器学习

在现代智能研究领域，机器学习是其中一门分支学科，主要是为了实现计算机能够自主学习，从而增强处理问题的能力。机器学习需要足够的运算内存提供支持，从而形成大数据与机器学习彼此的共同完善，本文通过对机器学习模型的探讨，形成高效大数据处理方式。

一、Spark 方法

针对 Spark 方法，在进行计算过程主要的特点可以归纳为如下几点，第一，该方法计算速度快。在进行循环数据流的运算过程中运用无环图方式，将运算形成的中间数据存储到系统内存，从而形成高效运算；第二，使用方便，可以兼容多数语言；第三，该方法具有较高的通用性，具有多种功能组件。Spark 方法能够提供十分全面的组件库，比如流失计算、数据查询、机器学习等组件。

通过将模型各项参数存储到 driver 节点，形成与 Workews 之间的信息交流，从而实现参数进行迭代更新。在进行大量数据部署的过程中，模型形成的参数不能在 driver 进行储存，因此需要在进行迭代过程中进行 RDD 运算对参数更新之后的数据进行保存[1-2]。

二、基于分布式的机器学习算法

（一）机器学习算法的迭代计算

需要利用多次迭代实现，在每一次迭代计算之后都会形成新的参数，通过迭代过程的不断进行模型形成的参数向量会逐渐变少。因此，需要进行迭代次数的判断，利用参数 a 作为阈值，当迭代过程中形成的模型参数大于 a 的时候，就需要对迭代次数进行粗调，每次都会形成模型参数改变。当模型参数小于 a 的时候，通过对迭代次数的调整，最终形成最优的模型参数。

（二）算法的优化

根据上述分析的学习样本的差异性对于模型参数的影响，本文进行了算法的优化：

输入：α、β、$flag \leftarrow false$

输出：模型参数向量 ω

方法：对结点 r 第 t 次迭代计算过程进行计算。

（1）$if\ flag == false$

（2）按照当前模型进行梯度计算

（3）else

（4）for each sample i

（5）$if \| g_i^{(t-1)} \| \leq \beta$

（6）$g_i^{(t)} \leftarrow g_i^{(t-1)}$

（7）else

（8）$g_i^{(t)} = \triangle f(w^{(t)}, x_i y_i)$

（9）end if

（10）end for

（11）end if

（12）进行模型总体度计算

（13）计算得到的总体度 g 发送到服务器节点

（14）获得新模型参数数据集

（15）$if \| w_i^{(t)} - w_i^{(t-1)} \| \leq \alpha$

（16）$flag \leftarrow true$

（17）else

（18）$flag \leftarrow false$

（19）end if

通过上述优化算法可以看出，在样本对模型参数梯度变化贡献较大的时候，需要在每一次的迭代计算进行梯度计算，否则在进行微调计算时不需要进行计算，使用上次的梯度数据，虽然计算存在一定的误差，但是由于对整体梯度影响较小，可以忽略不计。

在梯度数据集计算中出现误差较大的时候，迭代计算就会进行粗调计算，此时需要对所有样本进行精确的梯度计算，从而保证计算精准性。

三、实验结果和分析

本文通过利用多层神经网络数据对单层神经网络进行训练，验证所设计的优化算法的合理性，通过对均方误差以及迭代过程进行对比分析，结果表明，利用本文设计的优化算法可以实现误差减小、收敛速度提高的效果，可以实现对数据的精准预测。

参考文献

[1] 须成杰，肖喜荣，张敬谊等．基于 Spark 的大数据分析平台的设计和应用 [J]．中国卫生信息管理杂志，2019(5)：633.

[2] 王万良，张兆娟，高楠等．基于人工智能技术的大数据分析方法研究进展 [J]．计算机集成制造系统，2019，25(3)：5-23.

（作者单位：南华大学计算机学院）