# **Data Management in Machine Learning: Challenges, Techniques, and Systems**

## 文章摘要

使用统计机器学习（ML）的大规模数据分析为许多现代数据驱动的应用程序奠定了基础。 数据管理社区已经为解决ML工作负载中出现的与数据管理相关的挑战而工作了十多年，并且已经建立了多个用于高级分析的系统。该文章介绍了相关方向的系统情况，并 分析了数据管理中的挑战和技术。 我们专注于以下三项工作：（1）将ML算法和语言与现有的数据系统（如RDBMS）集成在一起；（2）将受数据管理启发的技术（如查询优化，分区和压缩）适应新系统 （3）结合数据管理和ML思想来构建可改善与ML生命周期相关的任务的系统。 最后，我们确定了重要的开放数据管理挑战，以供将来在这一重要领域中进行研究。

文章从ML工作负载模式、数据系统与ML、基于DB的ML系统、结合ML生命周期的系统以及一些开放问题入手进行分析。

## ML中的数据系统

文章介绍了将ML算法与RDBMS或更新的数据流系统集成在一起的系统，以使ML计算更接近数据所在的位置（例如，在RDBMS或HDFS中）。 因此，它们避免或减少了数据的移动成本。

在ML系统设计时会整合SQL、ML算法等综合设计系统。

## 使用DB的思想优化ML系统

异步执行、压缩以及分布式下的扫描共享

许多机器学习算法是迭代的，并执行重复的矩阵向量乘法。 由于矩阵矢量乘法类似于传统的表扫描，即使是内存I / O绑定，现有工作也试图通过压缩来减小数据大小或减少扫描次数。

## 通过整个ML的声明周期来看待系统

ML系统还包括机器学习生命周期中的其他重要任务，包括在训练之前发生的预操作。诸如特征提取，模型选择和模型管理之类的任务。 这些系统中的许多应用了以DB为中心的思想，例如declarativity, interactivity, 与 optimization，通常将这些思想与ML技术相结合。

文中提到特征提取、模型选择与管理。