Abstract

提升树是一种高效，广泛应用的机器学习方法。在本文中，我们描述了一个称为XGBoost的可扩展的端到端树升级系统，由数据科学家广泛使用，以获得许多机器学习挑战的最先进的结果。我们为稀疏数据和加权分位数草图提出了一种新颖的稀疏感知算法，用于近似树学习。更重要的是，我们提供了对高速缓存访​​问模式，数据压缩和分片的洞察，以构建可扩展的树升级系统。通过结合这些见解，XGBoost通过比现有系统少得多的资源，超越了数十亿个示例。

Introduction

机器学习和数据驱动方法在许多领域变得非常重要。 智能垃圾邮件分类器通过从大量垃圾邮件数据和用户反馈中学习来保护我们的电子邮件; 广告系统学会将正确的广告与正确的语境相匹配; 欺诈检测系统保护银行免受恶意攻击; 异常事件检测系统帮助实验物理学家找到导致新物理学的事件。 驱动这些成功的应用程序有两个重要因素：使用有效的（统计）模型来捕获复杂的数据依赖性和可扩展的学习系统，从而了解大型数据集的兴趣模型。