可细分部分

大数据背景下的机器学习算法

在大数据背景下，大规模数据集要求我们开发出的机器学习算法更加高效，并行度更大。在机器学习研究领域，学者们已经针对大规模数据集做了不少卓有成效的工作。

早期的工作，有比方说PSVM（Parallel Support Vector Machines）[4]的开发。PSVM利用了近似矩阵分解的方法。这种分解方法是基于行抽取的，它可以在算法运行时大大降低内存使用。该算法同时从而可以通过增加参与并行计算节点的数量来减少算法运行时间。该算法针对机群的可扩展性在上百台这样一个数量级上。

此后，研究者又提出了PLDA (Parallel Latent Dirichlet Allocation) [26]算法。该算法可以大幅提高LDA算法的运行效率。它采用随机抽取的方法。PLDA算法所发布的Hadoop版本具有很强的鲁棒性，它具有承受节点失败的能力，这其中既利用了算法的随机性，也利用了Hadoop的容错优势。

最近，Dean等学者的工作[18]表明并行化的优势可以再深度学习（deep learning）领域起到重大作用。他们的工作可以把机器学习算法的并行度推进到上亿个计算节点。这是迄今为止，机器学习研究领域所达到的最大并行度，同时也促使深度学习算法在图像识别、语音识别和很多其他应用上取得了目前最好的学习效果。

除了Hadoop并行系统所代表的框架，机器学习学术界也有很多对其他并行框架的尝试。GraphLab[17]是卡耐基梅隆大学开发的，专门针对机器学习中大规模图算法的一个计算工具。它可以大幅提升此类型算法的运算效率和在机群上的可扩展性。

并行算法框架

本文工作建立在三个并行系统上，这三个系统在运行机器学习算法时的算法框架都有其各自特点。

Apache Hadoop[27]作为一个软件代码库，实现了对大规模数据集的分布式处理。它被广泛应用与各类机器集群上，并且其使用了简洁的编程模型。它是一个开源并行框架，支持对高密度分布式数据的处理。而Hadoop所支持的机器集群可以是大规模的并且支持商品级硬件。Hadoop是从Google的MapReduce和GFS（Google File System）中演化而来的。Hadoop使用Java语言。它的开源特性保证了全球大量开发者的贡献都体现在了这个Apache的顶层项目中。Hadoop平台在一般意义上，包含Hadoop kernel，MapReduce和Hadoop Distributed File System (HDFS)，以及一些其他相关的项目，包括Apache Hive和Apache HBase。Hadoop的设计目标就是要大幅提高机器集群的可扩展性，从几台服务器上升到成千上万个运算节点。在Hadoop中每个节点都能提供本地的计算和存储。Hadoop框架同时对运行在其上的应用同时透明的提供数据可靠性保证以及数据通信保证。

Hadoop所执行的运算模型为MapReduce[7]。每个应用被分割为大量小片段，每个片段可以在机器集群中的任意一个节点上执行和重复执行。在MapReduce运算模型能够应用的问题中，输入会被解析为键-值对的集合。一个map函数会把这些键值对经过用户所定义的计算转为中间结果的键-值对。一个reduce函数会接着把这些中间结果按照键值索引聚集（其聚集方法是用户所定义的）。事实上，MapReduce所能应用的问题很广泛，不少问题都可以转化为一个或者一系列MapReduce任务。这种运算模型也使得它很易于底层的并行实现。所有的处理都是独立的，所以可以分配给不同的节点分别执行。除此之外，Hadoop还提供了分布式的文件系统HDFS。它可以给整个机器集群提供很高的通信带宽。这用map/reduce以及HDFS的设计保证了当节点失败时可以自动被系统处理而不影响整个任务执行。而这种容错性保证，并不依赖于硬件，而是Hadoop代码库本身的设计在应用层考虑了节点失败的检测和处理。所以尽管每个节点可能是易于出现问题的，但整个机器集群作为一个运算资源来说是可靠地。

Apache Mahout[20]是一个专门针对大规模机器学习算法所开发的代码库。其核心设计目标就是要提高在大规模数据上，提高机器学习算法的可扩展性。它的核心算法包括聚类算法、分类算法和协同过滤。这些算法的很大一部分是建立在Apache Hadoop的基础之上运用MapReduce框架的。同时，Mahout的核心代码库也含有一部分非并行计算程序（针对罗吉斯回归模型的算法就属于这一部分）。它们在代码优化上做了大量的工作，从而达到了很多好的运算效果。Mahout甚至广泛应用与商业领域，它良好的可扩展性，以及对很多流行机器学习算法的完整支持都促进了它的推广。

目前，Mahout主要支持以下四个应用场景：

1. 推荐系统：把用户行为作为输入，并根据用户行为进行个性化推荐。
2. 聚类：例如把文档按照主题进行分类。
3. 分类：例如从已经分好类的文档中学习模型，并将一个新的未被分类的文档分入正确的分类中。
4. 频繁模式：将一系列事项集合作为输入（例如查询词、销售图表等等），从中识别出那些事项模式出现的较为频繁。

Mahout当前支持很多机器学习算法，包括协同过滤算法、基于用户和基于产品的推荐算法、K-Means算法、模糊K-means聚类算法、平均偏移聚类算法、狄利克雷过程聚类算法、LDA算法、奇异值分解算法、并行频繁模式算法、互补朴素贝叶斯分类器和随机决策树分类器。

Spark[29]系统是由加州大学伯克利分校的AMPLab开发的。它是一个开源的分布式计算系统。Spark的设计目标是使得数据分析更加快速：一方面是运行时间的加速，另一方面是读写的加速。为了使程序运行得更快，Spark提供了基于内存和缓存的集群计算基本操作：一个任务可以将数据最充分地载入内存中，从而使得反复不断地访问操作要比硬盘读写来的更快。而硬盘读写多这一点，正是Hadoop系统的最大劣势。为了使得在Spark系统上的编程开发更为快速，Spark支持Scala和Python语言，并提供了简洁的API。用户甚至可以以交互式的方式，通过Scala和Python语言来快速处理大数据集。Spark系统在设计之初，主要是面向两类应用问题：迭代程序（这是在机器学习算法中广泛使用的）和交互式数据挖掘。如果能够把需要重复利用的数据始终保存在内存中，这两类应用问题的算法执行效率都必然能够获得极大提升。而随后的测试也表明，相比于Hadoop系统，Spark确实在这两个应用中取得了100倍以上的加速效果。

Spark系统较为年轻，但也已经在工业界获得了不少应用。Spark系统是Shark系统的后端引擎，而Shark系统是一个与Apache Hive系统相兼容的数据仓储系统。同样的，Shark系统也比Hive系统有100倍以上的加速。尽管Spark系统是新开发的一个并行框架，它支持访问HDFS上的数据，从这个意义上讲，它也与Hadoop兼容。这个特性使得很多基于Hadoop的程序不用再重新开发，从而给Spark也带来了更广的应用前景。

Spark是建立在Mesos上的系统。Mesos是一种运行于集群上的操作系统。Mesos可以使得多个并行程序良好地共享一个集群。它也同时提供了API来支持其上的应用在集群上部署并行任务。Spark通过利用Mesos系统可以在集群上与其他系统良好共存，比如说同样建立在Mesos系统上的Hadoop和MPI。另外，利用Mesos系统，使得Spark的开发过程节约了原本所需的大量编程工作。

Spark系统得以成功的关键在于运用了弹性分布式数据集（Resilient Distributed Dataset，简称RDD）这个概念。每个弹性分布式数据集代表一块分布在集群中一部分节点上的只读对象。它具有容错特性，既当一个弹性分布式数据集的分布方式信息丢失时，系统可以自动重构这个弹性分布式数据集，从而保证数据的可靠性。用户可以清晰地在应用程序代码中创建RDD，并根据需要把它载入整个集群的缓存中，并在像如前所述的MapReduce框架下的map和reduce函数中不断重复访问这些数据。弹性分布式数据集通过仅支持线性操作（lineage）来保证容错性。如果一个弹性分布式数据集丢失，系统可以通过它所对应的线性操作方式（也可以是一串线性操作）来对重新生成这个弹性分布式数据集。

次线性方法

近些年来，次线性方法逐渐为人所熟知。Clarkson等学者充分运用随机算法提出了这一新的方法[5]。他们在机器学习领域首先提出了这种次线性时间的近似优化算法。他们把这个方法应用在线性分类器和最小闭包球这两个基础问题上。次线性方法结合了一种新型抽样方法和一种新的可乘式更新算法。他们同事证明了算法下界。该算法下界表明算法理论运行时间对于RAM模型来说，已接近最优。

Hazan等学者[16]