决策树算法和MAPREDUCE技术：综述

Bisma Bashir[1]，Farheen Siddiqui[2]
1.控制科学与工程，SEST，佳米雅综合大学，新德里，110062，南德里，印度
2.控制科学与工程，SEST，佳米雅综合大学，新德里，110062，南德里，印度

摘　要：在当今时代，大数据和数据挖掘起着非常重要的作用。大数据可以是结构化的，也可以是非结构化的。这种巨大的数据增长引发了处理这些数据的新技术和工具的开发。不同的算法被用来对大数据进行分类，但是基于决策树的算法类别被认为是所有其他算法中最好的分类器。决策树算法实现起来简单快速。本文讨论了基于MapReduce框架的基于决策树的算法。随着数据的增长，决策树算法有一些需要解决的局限性。MapReduce技术是采用决策树算法进行大数据处理的最佳解决方案。

关键词：大数据；决策树算法；分类技术；ID3；C4.5；MapReduce

Decision Tree Algorithm and MapReduce Technology: A Review

Bisma Bashir[1], Farheen Siddiqui[2]
1. Department Of CSE, SEST, Jamia Hamdard, New Delhi, 110062, South Delhi, India
2. Department Of CSE, SEST, Jamia Hamdard, New Delhi, 110062, South Delhi, India

Abstract: In today's era bigdata and data mining plays a very vital role. Big Data can be structured as well as unstructured. This enormous data growth has triggered development of new techniques and tools to process this data. Different algorithms are employed to classify bigdata, but the category of algorithms that are decision tree based are considered the best classifiers among all the other algorithms. Decision tree algorithms are simple and fast to implement. This paper discusses decision tree based algorithms that are based on map reduce framework. As the data grows decision tree algorithms have some limitations which need to be solved. MapReduce technology is the best solution for the bigdata processing with the decision tree algorithm.

Keywords: Big data, Decision tree algorithm; Classification techniques; ID3; C4.5; MapReduce

一、引言

在当今时代，数据增长非常快。数据可以是结构化的，也可以是非结构化的。由于数据的爆炸式增长，迫切需要新技术和工具来处理这些海量的数据并将其加工成有用的信息。分类和预测技术等数据调查技术可用于从大数据中提取重要数据。大多数分类技术都受内存限制，因此可以在小型数据集上正常工作[1]。

对于海量数据的分类，使用了各种技术，例如决策树、k-最近邻、贝叶斯和基于神经网络的分类器。这些技术中最好的技术是决策树分类。与其他分类方法相比，决策树算法具有较高的效率和准确性[2]。决策树算法实现起来很便宜。在决策树中，有一个没有传入边的根节点，并且有许多内部节点和子节点。从根节点到子节点的每条路径形成一个决策规则。在海量数据上应用决策树算法会导致树构建延迟[3]。随着数据的不断增加，构建决策树变得非常耗时。MapReduce技术用于解决决策树构建中大数据增加的问题[4]。

MapReduce 是一种编程模型，它由一个执行过滤和排序的 map 过程和一个执行汇总操作的 reduce 过程组成[4]。对于大规模数据的处理和分析，MapReduce 是最流行的框架。大数据集通过 MapReduce 模型分割成更小的数据集，每个数据集在不同的计算机上单独处理，然后将每个子过程的结果聚合起来，生成最终结果。MapReduce 架构允许比传统数据更快地处理大规模数据[5]。

本文的结构如下；第 2 节描述了本文的文献调查。第 3 节描述了 MapReduce 过程。第 4 节描述了决策树算

法。第 5 节给出了 ID3 和 C4.5 决策树算法的比较，并介绍了这两种算法的缺点。第 6 节描述了基于 MapReduce 的决策树，第 7 节总结了本文。

二、文献综述

A. Tiwari 等人 2012[6] 在他们的论文中收集了实验数据，以找出最适合不同数据集的决策树算法。通过对实验数据使用所实现的方法，计算了算法的性能，最后对结果进行了仿真。Lakshmi 等人 2013[7] 在他们的论文中发现了决策树算法的性能。他们根据学生的定性数据对ID3算法、C4.5算法和CART算法进行了比较。他们提出了他们研究工作的实验设计，并得出结论，与 ID3 算法和 C4.5 算法相比，CART 具有更好的性能和最佳的分类精度。H. Chauhan 等人 2014[8] 在他们的论文中提出了决策树算法的实验性能评估。他们得出结论，C4.5 算法在分类准确率方面优于ID3算法。他们已经对可用的数据集进行了实验。具有许多实例的数据集更适合 C4.5 算法。W. Dai 等人 2014[3] 在那篇论文中提出了在分布式环境中使用 MapReduce 计算模型的 C4.5 算法。他们使用了 Hadoop 集群。研究的目标是加快决策树的构建，保证分类的准确性。本文通过实验证明了所提算法的效率和可扩展性。B.R. Patel 等人 2014[1] 在他们的论文中表明，算法的性能取决于熵、信息增益和数据集的性质。本文介绍了各种决策树算法的基本思想，称为PDTSSE：基于 MapReduce 的可扩展并行决策树算法。PDTSSE 算法旨在改善以前算法存在内存限制和运行时间延迟的缺点。PDTSSE 用于 MapReduce 框架中的大规模决策树学习。D. Wei 等人 2014[9] 在他们的论文中提出了 C4.5 算法的 MapReduce 实现。随着大数据的增多，传统的决策树算法已经不适用。随着日期的增长，树的构建非常耗时，并且由于大数据无法放入内存，因此操作的输出成本也变得非常高。所提出的算法提高了效率和可扩展性。F. Yuan 等人 2015 [10] 在他们的论文中提出了一项基于遗传算法（GA）优化方法使用 MapReduce 框架的决策树并行实现的研究。本文提出的MR-GAOT算法证明了该算法的可行性，同时也表明GA优化比传统的决策树算法具有更高的分类精度和更短的运行时间。S. B. Evangeline 2016[11] 在他们的论文中介绍了决策树算法的分析及其在各个领域的应用。为了对现有信息的结果进行分类、决策、识别和跟踪，决策树算法在其中起着至关重要的作用。他们得出结论，决策树算法是一种可以在应用程序中使用的重要技术。

三、MapReduce

MapReduce是一种分布式计算的编程模型。MapReduce 编程模型中包含两个主要任务，即 map 任务和 reduce 任务。在Map任务中，数据集被分解为元组（键/值对）。在reduce阶段，map阶段的输出作为该阶段的输入，map任务完成后，reduce任务开始[12]。MapReduce 的主要优点是可以轻松地在多个节点上处理数据。MapReduce 模型下的数据处理原语称为映射器和缩减器[12]。

如图 1 所示的MapReduce编程模型的基本阶段是[13]：

·Map阶段：将输入数据分成M个Map函数，称为Mapper。Mappers 并行运行，MapReduce 的输出是中间键值对。

·Shuffle and Sort 阶段：映射器的输出通过散列输出键进行分区。这里分区的数量等于reducer的数量。在 shuffle 阶段，所有的 key 和 value 对共享同一个 key，属于同一个部门。在对MapReduce进行分区之后，每个分区都被一个键缩短以合并该键的所有值。

·Reduce 阶段：第二阶段的输出被划分为R的Reduce函数，称为 Reducer。Reducers 处理不同的中间键并且并行运行。

MapReduce 的基本阶段描述如下[14]：

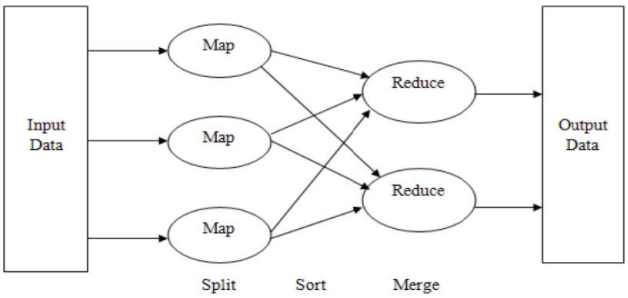

图1　MapReduce 的基本阶段

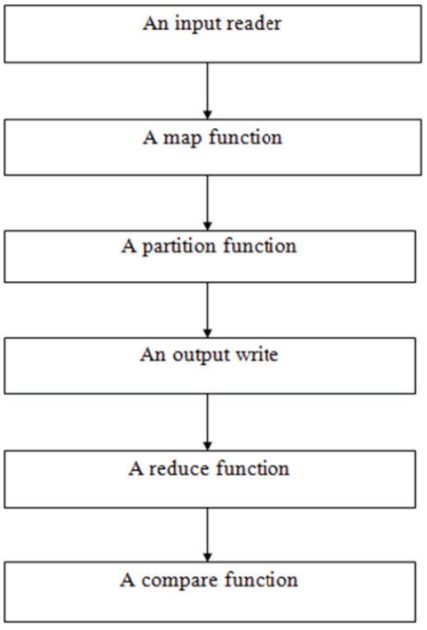

图2　MapReduce 编程模型的数据流

图 2 显示了 MapReduce 编程模型的数据流，显示数据流包括六个步骤[15]。

四、决策树算法

决策树算法是一种用于分类问题的监督学习算法[16]。决策树算法有很多，但这里我们只考虑下面讨论的 ID3 和 C4.5 算法。

ID3 算法：ID3 决策树算法由 Ross Quinlan 开发。在数据集上采用自上而下的贪心搜索，以测试每个节点的每个属性，然后选择最有价值的属性进行分类[6]。为了确定决策树中每个节点的合适属性，使用了信息增益。选择信息增益最高的属性集作为当前节点的测试属性[1]。该算法用于机器学习。

C4.5 算法：C4.5 算法是由 Ross Quinlan 开发的，作为 ID3 算法的扩展。C4.5算法被称为统计分类器，因为C4.5算法生成的树可以用于分类。C4.5算法以信息增益作为分割标准。C4.5 算法接受分类值和数值。缺失值由 C4.5 算法处理，因为它们不用于增益计算 [1]。C4.5 算法的优点是同时处理连续和离散属性以及处理缺少属性值的训练数据[11]。

图 1 显示了决策树算法的流程图[17]。

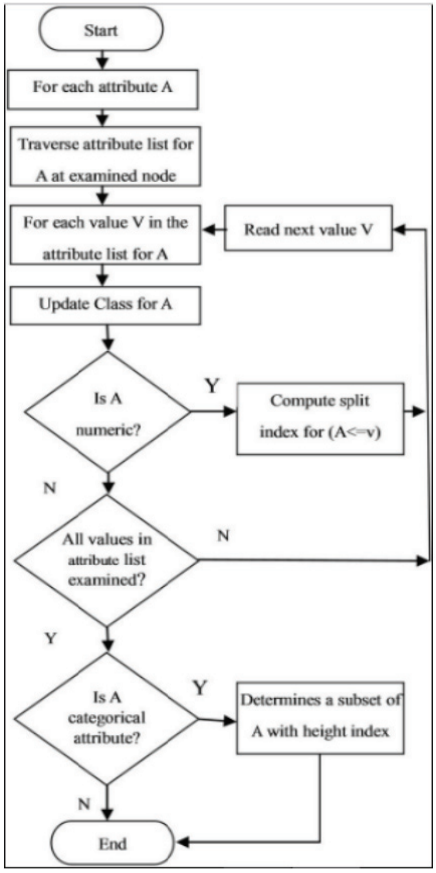

图3　决策树算法流程图

图 3 中的流程图描述了决策树检查树构造的属性并检查属性列表的完整逐步过程。如果属性为数值，则将这些值与给定的索引进行比较。逐步检查列表中的所有值。如果属性是分类属性，则计算具有最高索引的属性并继续进行树构造。具有最高信息增益的属性是根节点。具有零信息增益的属性没有子节点。

五、ID3 与 C4.5 算法对比

ID3 和 C4.5 算法的比较见表 1[18]。

表 1　ID 3 和 C4.5 的比较

Algorithm	Splitting Criteria	Attribute Type	Missing Values	Pruning Strategy
ID3	Information Gain.	Handles only Categorical values.	Do not handle missing values.	No Pruning is done.
C4.5	Gain ratio.	Handles both categorical and numerical values.	Handles missing values.	Error based pruning is used.

5.1 ID3 算法的缺点

ID3 算法的缺点如下[18]：

- 当测试一个小样本时，数据可能会被过度固定
- 为了做出决定，一次测试一个属性
- 缺失值和数值属性不被算法处理

5.2 C4.5 算法的缺点

C4.5 算法的缺点如下[18]：

- 一次只能拆分一个变量
- 可能会形成不平衡的树

六、基于 MapReduce 的决策树

随着大数据的增长，构建决策树变得越来越困难，因为树的构建需要大量时间。随着大数据的增加，对内存的要求非常高。因此，MapReduce 是解决决策树中大数据挑战的最佳解决方案。MapReduce 实现提高了决策树算法的时间效率和可扩展性[19]。

随着 MapReduce 在决策树算法中的实现，随着节点和处理器数量的增加，性能得到了提高。由于决策树的不规则性，其性能取决于各种因素。在 MapReduce 环境中，各种因素都会影响决策树的 MapReduce 结果。例如，硬件布局会影响决策树的性能。MapReduce 的实现依赖于不同的领域，在不同的情况下，MapReduce 的实现有利于战利品，但另一方面，它的实现也会导致各种缺点[20]。

七、结论

决策树算法是最好的分类器，用于各种分类技术。我们对 ID3 和 C4.5 算法进行了分析，我们的分析证明 C4.5 算法的精度高于 ID3 算法。由于 C4.5 算法可以处理数值数据和分类数据，但 ID3 算法不能处理数值数据。由于这种 C4.5 算法比 ID3 算法具有更高的准确性。

随着数据的增长，构建决策树变得越来越困难。在大数据上构建决策树的最佳方法是借助 MapReduce 技术。MapReduce 用于大型数据集的并行计算。在 MapReduce 的帮助下，决策树的构建变得容易并且树的

构建是准确的。MapReduce还有助于形成没有重复子节点的平衡树。

参考文献：

[1] Patel, B., & Rana, K. (2014). A Survey on Decision Tree Algorithm For Classification. IJEDR, 2(1).

[2] Cui, Y., Yang, Y., & Liao, S. (2014). PDTSSE: A Scalable Parallel Decision Tree Algorithm Based on MapReduce.

[3] Dai, W., & Ji, W. (2014). A Scalable Parallel Decision Tree Algorithm Based on MapReduce. A Scalable Parallel Decision Tree Algorithm Based On Mapreduce, 1, 49–60.

[4] Tiwari, A., & Athavale, V. (2012). A Survey on Frequently Used Decision Tree Algorithm and There Performance Analysis. International Journal Of Engineering And Innovative Technology (IJEIT), 1(6).

[5] Lakshmi, T., Martin, A., Begum, R., & Venkatesan, V. (2013). An Analysis on Performance of Decision Tree Algorithms using Student’s Qualitative Data. I. J. Modern Education And Computer Science, 5, 18–27.

[6] Chauhan, H., & Chauhan, A. (2014). Evaluating Performance of Decision Tree Algorithms. International Journal Of Scientific And Research Publications, 4(4).

[7] Yuan, F., Lian, F., Xu, X., & Ji, Z. (2015). Decision Tree Algorithm Optimization Research Based on MapReduce.

[8] Evangeline, S., & Sudhasini, P. (2016). An Introduction to Decision Tree algorithm on Various Field of Applications. International Journal Of Digital Communication And Networks (IJDCN), 3(3).

[9] Python), A., Python), A., & Team, A. (2017). A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python). Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one

[10] (2017). Retrieved from https://www.quora.com/What-are-the-differences-between-ID3-C4-5-and-CART

[11] Decision Tree. (2017). Saedsayad.com. Retrieved from http://www.saedsayad.com/decision_tree.htm

[12] Karim, M., & Rahman, R. (2017). Decision Tree and Nave Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing.

[13] "Mapreduce". En.wikipedia.org. N.p., 2017. Web. 4 Mar. 2017.

[14] What MapReduce can't do. (2017). Analyticbridge.com. Retrieved from http://www.analyticbridge.com/profiles/blogs/whatmapreduce-can-t-do

[15] Hadoop MapReduce. (2017). www.tutorialspoint.com. Retrieved from https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[16] Bisma B. An Approach of MapReduce Programming Model For Cloud Computing. International Journal of Advanced Research in Computer Science, 8 (2), March 2017, 43–45

[17] 2017. Retrieved from https://www.google.co.in/serch?q=fundamental+phases+of+mapreduce

[18] Sushant, S, N. Ashishkumar . International Journal of Science and Research (IJSR) (2013),

[19] Wei, D. Wei, J. (2014). A MapReduce implementation of C4.5 Decision Tree Algorithm. International Journal of Database Theory and Application, Vol.7, No.1.

[20] Tianyi, Y. and Anne, H, H, N. Implementation of Decision Tree Using Hadoop Map Reduce. (2016), International Journal of Biomedical Data Mining. Volume 6 Issue 1