## 《大数据导论》研读讨论作业

## Main Ideas and Major Contributions of Readings Discussion 4 (in Chinese)

备注：小组讨论翻译并理解给出的研读文献，无需全文直译，但需要撰写文献的核心思想、主要过程和主要贡献，依据撰写质量评价给分

Name: \_\_\_\_\_黄昊\_\_\_\_\_ Student ID: \_\_20204205\_\_

On the use of MapReduce for imbalanced big data using Random Forest. Information Sciences 285 (2014) 112–137

这篇文章首先对文章的内容做了简介，然后依次对“大数据和不平衡数据集的分类方法”，“随机森林算法及其一些能够分别处理大数据和不平衡数据集的版本”，“使用随机森林算法为大数据采用的不平衡方法”，“所进行的实验研究”这四部分内容分别进行了详细的说明，最后对总结和归纳了作者所进行的工作。

研究背景：

随着数据生成速度的不断提高，需要处理的数据量变得庞大且复杂，并且这些数据难以用传统技术来处理或分析，所以出现了大数据的术语。大数据被定义为需要新形式的数据处理的具有高容量、高速度、多样化、真实性和有效性等特征的信息。

而MapReduce编程框架则被设计为以极其平行的方式处理大数据的平台，它将计算过程抽象为Map和Reduce两个阶段。Map阶段将输入的数据集分割为独立的子问题交给工作节点处理，工作节点以平行的方式处理较小的问题后将答案回传给主节点；在Reduce阶段主节点将所有的子问题答案合并后形成最终结果。

在数据处理时又出现了不平衡数据集分类问题，不平衡数据集指不同类别样本的数量有明显的差异。而在某些领域实例数较少的类通常是最重要的，若不能正确识别这些类别，将会造成很高的成本。数据集的不平衡率以及数据内在特征都会对不平衡数据集的分类产生负面影响，于是出现了许多技术用于处理不平衡数据集。这些方法主要被分为数据层面的方法（分为过量取样方法、欠量取样方法和混合取样方法），算法层面的方法，对成本敏感的学习方法和集合方法。

数据集分类质量的衡量标准是由混淆矩阵建立的，如下表所示：

|  |  |  |
| --- | --- | --- |
|  | Positive prediction | Positive prediction |
| Positive class | True Positive (TP) | True Positive (TP) |
| Negative class | Negative class | Negative class |

这篇文章使用了Geometric Mean（GM）（）和b-f-measure（b-f-m）（）两种指标对数据集进行分类。

主要内容

首先介绍了数据的分类方法，随机森林及它的两个变体。

随机森林（RF）是用于分类的决策树组合，RF从原始样本集得到新的样本集用于建立几个不同的随机树，然后通过多数投票汇总决策树群的预测值来计算样本的预测值。

RF的第一个变体是Mahout Partial implementation (RF-BigData)算法，该算法是一种为数据的不同部分建立多个树的算法。分为两个不同的阶段：第一阶段创建模型，第二阶段利用先前学到的模型估计与数据集相关的类别。

在第一阶段，随机森林按照MapReduce程序从原始训练集建立的。由三个步骤组成：Initial，Map和Final。Initial步骤将训练数据集分割成独立的数据块，然后复制这些数据快并在不同的处理节点之间传输。Map步骤，每个map任务用其分区的数据块建立一个森林的子集，并生成一个包含所建树的文件。Final步骤对所有mapper生成的输出文件进行解析，提取树木。所有树的集合构成了森林。

第二个阶段，分类阶段，估计与数据样本集相关的类别。由三个步骤组成：Initial，Map和Final。Initial步骤将可用的数据样本集分割成独立的数据块；将它们复制并转移到其他机器上，最后由每个map任务并行处理。Map步骤中每个mapper使用前一阶段建立的RF模型中的树所预测的类别的多数票来估计其中可用的例子的类别。Final步骤中每个mapper产生的预测结果被连接起来，形成最终的预测文件。

RF的第二个变体是加权随机森林，一种基于成本敏感学习的RF版本，目的是解决不平衡的数据集。它在RF建立树的过程中和分类步骤中，修改了所使用的投票方案。在构建树的过程中，成本被纳入到所有的内部计算中。并且使用加权投票方案。计算出每一片叶子的权重，这个权重被计算为与该叶子中的实例相关的成本比例。在预测时使用叶子节点的权重的聚合加权投票决定最终的分类。

然后这篇文章的团队建立了一个新的RF版本（RF-BigDataCS），用来对不平衡的大数据进行分类。他们在基本的RF实现中加入了对成本敏感的操作。在数据集被分配给mapper之前，估计了每个类别的成本。修改了用于选择最佳分割的标准，以建立与错误分类成本有关的实例加权的树，并且还考虑了计算哪个类别与叶子相关的成本。

接下来作者介绍了几个适用于大数据场景的不平衡类的数据预处理方案。

一是大数据的随机超抽样。ROS算法按照MapReduce的设计来处理大数据，其中每个Map进程负责通过随机复制少数类别的实例，调整mapper分区中的类别分布，Reduce进程负责收集每个映射器产生的输出，形成平衡的数据集。

这个过程由四个步骤组成：Initial，Map，Reduce和Final。在Initial步骤中，算法将输入数据集分割成独立的数据块，并将其复制和转移到其他机器上。在Map步骤中，每个map任务通过随机复制少数类别的实例来平衡类别分布。Reduce步骤收集每个mapper产生的输出，并随机化平衡数据集中的实例。在Final步骤，将Reduce步骤产生的平衡数据集作为RF-BigData算法的入口数据。

二是大数据的随机欠抽样。适应于处理大数据的RUS版本采用了MapReduce设计，每个Map进程负责将其数据分区中的所有实例按类别分组，Reduce进程负责收集每个mapper的输出，并通过随机消除多数类别的实例来平衡类别分布，形成平衡的数据集。这个过程由四个步骤组成：Initial，Map，Reduce和Final。在Initial步骤中，将输入数据集分割成独立的数据块后被自动复制并在不同的集群节点之间传递。在Map步骤中，每个map任务对其数据部分的所有实例进行处理并按类分组。Reduce步骤收集每个mapper产生的输出，并平衡类别分布，随机地消除多数类别的实例。在Final步骤，将Reduce步骤产生的平衡数据集作为RF-BigData算法的入口数据。

三是大数据的SMOTE。SMOTE算法每个Map进程对少数类进行过度采样，Reduce进程随机化每个mapper产生的输出，以形成平衡的数据集。由四个步骤组成：Initial，Map，Reduce和Final.。在Initial步骤中，算法将输入的数据集分割成独立的数据块后复制并传输给其他mapper。在Map步骤，每个map进程使用基本的SMOTE算法在可用数据上平衡mapper分区中的类别分布。Reduce步骤收集每个mapper产生的输出并随机化最终的数据。在Final步骤，将Reduce步骤产生的平衡数据集作为RF-BigData算法的输入数据。

然后作者介绍了他们团队为比较不同方法的性能而进行的实验研究，这些方法使用随机森林算法处理不平衡的大数据。首先作者介绍了实验中的分类问题、使用的算法和参数的选择。然后研究了顺序版本在处理规模增大的数据集时遇到的限制。接着对不平衡大数据的各种分类方法的有效性分析。然后评估了所有用于比较的不平衡大数据技术所花费的运行时间。最后得出结论：

顺序版本不是处理不平衡大数据的适当方法，有必要解决这些问题以便在可用数据规模增加时提供适当的解决方案，如MapReduce方法。当mapper的数量增加时，执行时间通常会减少，但由于样本量小的问题，mapper的数量太多可能会对性能产生负面影响。

在使用RF算法时，并没有一个处理不平衡大数据的最佳方法，因为它取决于问题的类型以及密度不足对特定方法的影响。当SMOTE算法适用于大数据时，它在所考虑的不平衡大数据问题中的分类性能很差，因为它被认为是不平衡分类中比较有竞争力的方法之一。当增加mapper的数量时，RUS版本的分类性能比其竞争对手下降得更多，因为它在更大程度上受到小样本量问题的影响。

从所获得的结果可以得出结论，需要更深入地分析该问题，我们要应对以下挑战：

1. 需要确定少数类别的阈值与mapper的数量，以便找到最佳配置，提供更好的性能和更少的响应时间。本篇文章的结果表明，在考虑的极度不平衡的大数据问题中，平均结果要低得多，并且遇到了小样本量。
2. 考虑到SMOTE算法的结果，有必要设计新的技术，以便在使用MapReduce框架时，能够以最佳方式生成代表少数类实例的协同数据。
3. 有必要分析一些与此问题相互作用并加剧问题的数据内在特征，以提高不平衡大数据的性能。需要特别注意噪声和边界实例的存在，当mapper分区中包含的少数类实例数量太少时，这些特征会使不平衡问题变得更加困难；在选定的数据集上诱发小样本量问题。当原始数据的规模缩小时，数据中存在小的不连贯性；类之间的重叠可以通过将原始数据划分为子集来扩展，并且当增加拆分次数时数据集移位问题会更加严重。

主要贡献

本文建立了一个新的RF版本（RF-BigDataCS）用于对不平衡大数据进行分类，通过使用RF算法及其不同版本的变体和不同的不平衡分类方法对不平衡大数据进行处理分别实验，得出随机森林算法的顺序版本不是处理不平衡大数据的适当方法，在使用RF算法时并没有最佳的不平衡分类方法来处理不平衡的大数据问题。继而给出了在使用RF算法处理不平衡大数据问题需要应对的三点挑战以及建议对当前方法改进以减少使用更多mapper时观察到的性能损失。