BDCC 19/20
Worksheet #4
May 26th, 2020

Questions about paper:

- A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning

General questions

1. What is this paper about? Could you summarise its contribution in a paragraph?

This paper presents the distributed implementation of an algorithm that learns the structure of Bayesian networks.

2. How does this work differ from others mentioned in the paper?

There are quite a few implementations of Bayesian network learning in the literature. I believe the authors fail to explore the state-of-the-art in this field.

3. Do the authors present experiments? What is the methodology used? Does it sound correct? Why?

Yes, experiments are performed to answer several questions. The methodology used relies on the use of the Kepler workflow to parallelize the construction of the network structure. Data is partitioned and each process tries to build a network using the MMHC algorithm that works in two steps: finding for each variable a set of candidate parents, and refining the resulting network by adding, removing or reversing an edge. I am not quite sure this algorithm is correct by working with partitions of the data.

4. What are the main results/findings/conclusions? Are the results useful/relevant? Why?

Results show that it is more efficient to learn the networks using the parallel method proposed.

Technical questions

1. What is Distributed Data Parallelism (DDP)?

The name says it all...Parallelism is explored at the data level, not at the control level.

2. In the sentence: “the SBNL workflow partitions the data set into data partitions of reasonable size” (Section IV), what woud be a “reasonable size”? Explore SBNL and find out what is the criterium to choose the data partition size.

According to the authors: SBNL has a score based algorithm to dynamically determine the best partition size to balance both learning complexity and accuracy. Then, data partitions are sent evenly to each local learner. The local learner will first use the value ofS_{Arc} to examine the data partition’s quality. If the quality is good, SBNL then enters local ensemble learning (LEL) step, each local learner will run MMHC algorithm separately on each local data partition to learn an individual BN. Then, local learner applies our proposed ensemble method on individual BNs to generate a final local BN. During local learning, the best local data partition is obtained in each local learner.

3. Discuss the consequences of building a Bayesian network by partitioning the data and building local Bayes nets that later will be combined. Is this affecting the results when compared with the sequential execution? How do you compare the results produced by the parallel implementation with the sequential one? What would be the best model?

These issues are not discussed in the paper. To make it a better paper, these issues need to be discussed. There is no theoretical discussion on the impact of the quality of the models when using partitioned data versus using all data.

4. From a theoretical point of vew, does the probabilistic model generated in parallel approximate the optimal probabilistic function?

This is not discussed in the paper and it is difficult to evaluate from those experiments alone.