6375 ML Project Proposal

Ninad Khargonkar (nxk180069@utdallas.edu)

Problem

We base our research project on extending cutset networks [4] which are a part of tractable proabilistic models. Cuset networks are OR trees with Chow-Liu trees at their leaves. Learning a Chow-Liu tree is quadratic in the number of feature variables and this cost may be prohibitive in scaling the learning process of cutset networks. We plan to develop approximate methods to reduce the cost of learning the Chow-Liu trees at the leaves and hence arrive at a faster approximate algorithm with linear or pseudo-linear costs.

Technical approach

Computing the Chow-Liu of d data points with n feature variables involves finding the maximum spanning tree of the graph with nodes as the feature variables and the pairwise mutual information for all variables as the edge weights. The computation of the mutual information for all pairs takes $O(n^2 d)$ time and further the tree computation takes $O(n^2)$ time.

The above cost is the best we can do in finding the exact tree structure. In order to reduce it further we need to forgo computing some mutual information pairs which will have the double benefit of less edge weight computations and less edges in the graph (making it less dense). We will work with the assumption of binary variables for now and try to derive an approximation from the formula of mutual information between two binary variables as described in the method described in the paper by Meil et al [3].

This method uses the count number for each variable and tries to do develop a heuristic approach which tries to "guess" for pairs having a high mutual information value. We can do this either by sorting by the count value and doing a sort of greedy search for pairs. We plan to develop the approximation by constructing a spanning tree one node at a time by leveraging individual counts in the computation for a high mutual information pair.

In addition to such a heuristic we can use a sampling based approach on the data if the number of data points (d) is also very large since the overall time complexity also consists of a d term which might overpower n^2 too! Each of the approximation scheme can be employed for learning both the internal nodes of the OR tree and the CLtree at the leaf since learning of the internal nodes also depends on entropy calculations requiring a pass over the complete data.

Evaluation

We will use the test set log likelihood as a performance measure in order to compare against other algorithms in [4] and also measure the running time for the time tradeoff. We plan to use some if not all of the benchmark datasets used in [4]. The expectation is that the test set log likelihood will not show a significant drop while the runtime benefits from the approximation scheme.

Further experiments (under ideal circumstances)

If time permits we plan to apply this approximate method and do a time vs accuracy tradeoff analysis for different settings of cutset networks in multi-label classification [1, 6], ensembles of cutset networks [2] and compiled models [5].

References

1. Mauro et al, “Multi Label Classification with Cutset Networks”, PGM, 2016.
2. Rahman et al, “Learning Ensembles of Cutset Networks”, AAAI, 2016.
3. Meil et al, “An accelerated Chow and Liu algorithm: fitting tree distributions to high-dimensional sparse data”, ICML, 1999.
4. Rahman et al, “Cutset Networks: A Simple, Tractable, and Scalable Approach for Improving the Accuracy of Chow-Liu Trees”, ECMLPKDD , 2014.
5. Rahman et al, “Look Ma, No Latent Variables: Accurate Cutset Networks via Compilation”, ICML, 2019.
6. Rahman et al, “Cutset Bayesian Networks: A New Representation for Learning Rao-Blackwellised Graphical Models”, IJCAI, 2019.