A Comparative Study of Different R Frameworks for Large Graph-Based Semi-Supervised Learning

Prithish Banerjee[1,1,⋆], **Mark Culp**[2,1]

1. Department of Statistics, West Virginia University, Morgantown, WV 26506, USA
⋆Contact author: pbanerj1@mix.wvu.edu

Abstract

There has been substantial interest from both computer science and statistics in developing methods for graph-based semi-supervised learning. The attraction to the area involves several challenging applications brought forth from academia and industry where little data are available with training responses while lots of data are available overall. Ample evidence has demonstrated the value of several of these methods on real data applications. The general framework for graph-based semi-supervised learning is to optimize a smooth function over the nodes of the proximity graph constructed from the feature data. The smoothness associated with the function is captured in a penalty matrix constructed from the graph and regularization with respect to this matrix controls the fit versus smooth tradeoff [1].

As the literature has progressed the interest has shifted to developing faster and more efficient graph-based techniques on larger data. Naturally, the issues associated with the graph construction phase are a central component to this research direction. In [2] the authors present a linear time complexity and highly parallelized *Local Anchor Embedding* (LAE) algorithm where the data points are represented through a weighted average of *Anchor Points* which were chosen as the *k-means* cluster centers. This allows one to construct the graph penalty matrix without actually constructing the graph. Most of these approaches are not implemented efficiently or available for general use. Moreover the reference BLAS and LAPACK libraries that come with vanilla *R* are not optimized for large matrix operations required for this work. In this talk we compare some of the accelerated frameworks for large LAE graph construction and label prediction. Some of the open source solutions used in this work are OpenBLAS, a GotoBLAS fork; Apple's accelerated framework vecLib BLAS library which is a combination of C-BLAS, C-LAPACK and some of their own implementations and Intel MKL libraries compiled in Revolution R Open. The computation time and performance accuracy of these frameworks are compared via three datasets.

Keywords: Semi-Supervised Learning, Graph based Learning, Basic Linear Algebra Subprograms (BLAS), vecLib, Revolution R Open (RRO)

References

[1] M. Belkin, P. Niyogi, V. Sindhwani *Manifold Regularization: A Geometric Framework of Learning from Labeled and Unlabeled Examples*, Journal of Machine Learning Research 7, 2006, 2399-2434

[2] W. Liu, J. He and S.F. Chang *Large Graph Construction for Scalable Semi-Supervised Learning*, 27th International Conference on Machine Learning, 2010