Graph Clustering Performance

Pierre Miasnikof *[1], Alexander Y. Shestopaloff[2], Yuri Lawryshyn[1], and Anthony J. Bonner[3]

[1]University of Toronto, Dept. of Chemical Engineering and Applied Chemistry, Toronto, ON, Canada
[2]The Alan Turing Institute, London, United Kingdom
[3]University of Toronto, Dept. of Computer Science, Toronto, ON, Canada

Abstract

Graph clustering and network community detection is a topic that has gained much attention recently. Indeed, many graph clustering/community detection algorithms have appeared in the recent literature.

However, performance evaluation of these clustering algorithms remains an open problem. Clustering on graphs can be broadly categorized as an unsupervised learning task, for which we do not benefit from the benchmarks provided by pre-labeled or pre-clustered data sets. To address this lack of performance measurements, many authors test their algorithms on "ground truth" data sets. These data sets, typically drawn from social networks, are instances where individuals, modeled as graph vertices, have identified their community affiliations (clusters).

While this reliance on "ground truth" data sets does indeed provide objective reproducible performance measurements, it does not guarantee the algorithm will perform similarly well on an unlabeled data set. Arguably, the quality of a clustering returned by a specific algorithm on an unlabeled data set is only assumed to be accurate, because the algorithm performed well on another data set.

We introduce statistical measurements of clustering performance, which can be applied to any unlabeled graph/network data set, with overlapping or non-overlapping clusters. Our suggested measurements allow for the objective comparison of algorithm performance on single data sets and across different data sets. In both cases, they help determine if the clusterings returned by an algorithm are significantly different from a random partitioning of vertices.

Estimating the number of clusters (communities) on a graph is another open problem. In fact, many clustering algorithms require the number of

*Offers thanks to Prof Derek Corneil of the University of Toronto Dept of Computer Science and Amit Bermanis of the University of Toronto Dept of Mathematics

clusters as an input parameter. In such cases, the current practice is to begin with an "educated guess" and iteratively re-apply the clustering algorithm with different inputs, until reasonable results are obtained. This iterative process is very time-consuming and may be infeasible when dealing with very large data sets. Here, a suitably estimated starting point estimate may be useful. Also, for algorithms that do not require the number of clusters as input parameter, having an estimate of the number of clusters provides an additional benchmark for clustering accuracy.

The eigengap heuristic has been suggested as a possible estimate for the number of clusters. Unfortunately, this heuristic approach relies on the spectral decomposition of Laplacian matrices, a very costly operation.

We review one spectral approximation technique from the literature and also propose our own. The former, drawn from the literature, makes use of the Gershgorin theorem and only estimates bounds on eigenvalues, without explicitly computing them. Our own approximation technique is based on random sampling of the adjacency matrix. Initial results suggest our sampling approach provides a better approximation of the spectra.

While our study has just recently begun, we are collaborating with a bank, which has provided us with a very large real-world network data set. We anticipate applicable results shortly.

Research supporteded by a MITACS-CIBC Accelerate grant.