Measuring General Relational Structure Using the Block Modularity Clustering Objective *

Adam Anthony and **Marie desJardins** and **Michael Lombardi**
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250
(410) 455-8894
{aanthon2,mariedj,lombard2}@umbc.edu

Abstract

The performance of all relational learning techniques has an implicit dependence on the underlying connectivity structure of the relations that are used as input. In this paper, we show how clustering can be used to develop an efficient optimization strategy can be used to effectively measure the structure of a graph in the absence of labeled instances.

Introduction

Relational learning refers to machine learning techniques that take as part of their input a set of relations between learning instances. These relations can be used in several ways. In supervised learning, for example, the class label of a particular instance can be modeled using the class labels of neighboring instances (Neville & Jensen 2003). In unsupervised learning, relations are typically analyzed more directly for significant connectivity patterns that uniquely identify a partitioning of objects (Nowicki & Snijders 2001). If the edges in a relation are randomly distributed, the patterns these algorithms take advantage of do not exist. Therefore, it would be useful to have a computationally inexpensive randomness test that would help to avoid running a costly relational learning algorithm on a data set that includes a random relation. This test could be used to identify and filter out relations that do not have meaningful structure.

In this paper we propose a graph-based metric, *block modularity*, that evaluates the structure in a relation, given a set of labels. After partitioning a relation graph with respect to the labels, the block modularity score is used to measure whether the distribution of relations within and between class groups is significant. Our preliminary results show that the average block modularity, computed using an efficient optimization technique, can be used as an *a priori* indicator of relational structure.

Related Work

To our knowledge there is no prior work that focuses on determining a relation's usefulness for a learning task. Recent approaches have combined propositionalization (*i.e.* flattening structured data) with feature selection to determine the best propositionalized attribute vector for a relational data set (Popescul & Ungar 2007). Such techniques could be considered to be a form of relation selection, in that the algorithms must choose which relation types to "follow" to find relevant features.

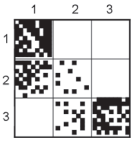

Figure 1: An example block model, in which the majority of links fall into blocks b_{11} ,b_{12} and b_{33}.

The block modularity measure, which we introduce below, is an extension of the well-known modularity objective for directed graphs (Leicht & Newman 2007). The two objective functions are identical in their choice of graph models and underlying theory. The difference between our objective and Leicht & Newman's is that ours is extended to measure the amount of structure in any kind of graph (*e.g.* bipartite, hierarchical, ring) while the original modularity is only able to measure the quality of structure in graphs that have *communities*, or groups that are densely connected with few links leaving the group.

Approach

Figure 1 shows a representation of a relation called a *block model* (Faust & Wasserman 1992). A block model is merely an adjacency matrix with the rows and columns rearranged such that objects in the same cluster are adjacent to one another in the matrix. The lines drawn show the separation between clusters, creating k^2 blocks, which are labeled b_{ij}. It is clear that if a clustering can be found for which most of the edges fall into a small number of blocks, leaving the remainder very sparse or empty, then the graph has a high level of structure. If the graph were random, then for all possible clusterings, one would expect to see the same number of edges in each block. The block modularity objective is based on this observation.

Given a partitioning of the vertices in a relation graph into k groups, the block modularity objective (defined over a relation S and a clustering C that divides the data into k clus-

*This material is based upon work supported by the National Science Foundation under Grant No. #0545726.

ters) is computed as:
(1)

$$BM(C,S) = \frac{1}{k^2|S|}\sum_{l=1}^{k}\sum_{m=1}^{k}\left|\sum_{i,j\in b_{lm}}\left[I[e_{ij}] - \frac{d_{i_{out}}d_{j_{in}}}{|S|}\right]\right|,$$

where $d_{i_{out}}$ and $d_{j_{in}}$ are the in- and out-degrees of vertices i and j, respectively, and $I[\cdot]$ is equal to 1 if the edge exists, and 0 otherwise. The term $(d_{i_{out}}d_{j_{in}})/m$ represents the degree-conditioned probability of an edge existing between vertices i and j in a null model random graph. By summing over all pairs of objects in each block b_{ij}, the function is computing the difference between the number of observed edges in each block and the expected number in a random graph. The absolute value ensures that blocks which have either significantly more, or significantly fewer, edges than are expected will receive a high score.

We we are currently in the process of developing an efficient clustering algorithm that uses block modularity as an objective. Our first version of this algorithm uses a greedy K-means style of optimization:

1. Randomly initialize the vertices into k clusters.
2. Until convergence, do:
   1. For each vertex, in random order (without repeats):
      a. Assign the vertex to the cluster that increases block modularity the most
      b. Update the block modularity score, given the greedy selection

Because there could be many clusterings for which the block model of a relation is significantly different from a random relation's block model, this algorithm locates several peaks for which block modularity is exceptionally high. In practice, because the algorithm can be efficiently implemented, we run the algorithm multiple times and take the highest-scoring clustering as the best solution. In this paper, we have found a unique use for the remaining high-value, but non-maximal, clusterings.

We hypothesize that a structured graph will have many more peaks with high block modularity scores than a random graph does. Therefore, by averaging the locally optimal value found by our algorithm over several runs, we can effectively measure the inherent structure in a relation without making any *a priori* assumptions about the expected structure (*i.e.* there is no need to search for the best community, k-partite, or hierarchical structures individually). One could argue that the relation selection step is unnecessary if the process itself requires clustering, but we argue that we are only applying a uni-relational algorithm, whereas multi-relational clustering algorithms (*e.g.* Kemp *et al.* (2006)) that are executed over a set of N relations have a much higher cost than running the above algorithm N times. We now show some preliminary results under this hypothesis.

Preliminary Results

To test our hypothesis, we started with a structured graph, which we generated using a generative model developed by Nowicki & Snijders (2001), which is a probabilistic interpretation of a block model. We then randomized the graph by removing a percentage of the total edges in the relation and reinserting the same number of edges at random. We specifically chose a random graph model that is different from block modularity's null model to show that this method is applicable to graphs with different distributions.

% Random	0	10	20	30
Average BM	0.116	0.111	0.105	0.092
% Random	40	50	100	
Average BM	0.085	0.089	0.087	

Table 1: Average block modularity scores for an increasingly randomized graph.

Table 1 shows the average block modularity score (over 100 runs) for an increasingly randomized graph. Statistical significance between each successive average (*e.g.* between 30% and 40%) was confirmed using a T-test, with each test yielding p-values less than 0.001. There is a downward trend from 0%-30% that shows that we can measure the degree of randomness between different graphs. Between 40%, 50% and 100% random, the values are very close. We believe this is because once a high number of edges are rewired, the result is virtually indistinguishable from a 100% random relation.

Conclusion

In this paper, we presented the block modularity objective function, which, given a clustering, measures the amount of significant relational structure in a relation, compared to a random graph model. We presented an algorithm that optimizes block modularity and how the results of multiple runs can be used as a statistic for measuring the inherent structure in a relation, without requiring any structural assumptions. In future work, we intend to continue investigating this structural statistic and its applications to relation selection. We also intend to further develop the block modularity objective and optimization technique to allow more control over the clustering output.

References

Faust, K., and Wasserman, S. 1992. Blockmodels: Interpretation and evaluation. *Social Networks* 14(1-2):5–61.

Kemp, C.; Tenenbaum, J. B.; Griffiths, T. L.; Yamada, T.; and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. In *AAAI 2006*.

Leicht, E. A., and Newman, M. E. J. 2007. Community structure in directed networks.

Neville, J., and Jensen, D. 2003. Collective classification with relational dependency networks. In *Proceedings of the Second International Workshop on Multi-Relational Data Mining*. ACM SIG-KDD.

Nowicki, K., and Snijders, T. A. B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96(455).

Popescul, A., and Ungar, L. 2007. *Feature Generation and Selection in Multi-Relational Statistical Learning*. Cambridge, MA: MIT Press, 1 edition. chapter 16, 453–476.