Agglomerative Clustering of Bagged Data Using Joint Distributions

David Arbour[*] DARBOUR@CS.UMASS.EDU
James Atwood[*] JATWOOD@CS.UMASS.EDU
Ahmed El-Kishky[†] AHMED-EL-KISHKY@UTULSA.EDU
David Jensen[*] JENSEN@CS.UMASS.EDU
[*]School of Computer Science, University of Massachusetts Amherst, 140 Governors Dr., Amherst, MA
[†]Tandy School of Computer Science, University of Tulsa, Rayzor Hall, East 5th Place, Tulsa, OK

Abstract

Current methods for hierarchical clustering of data either operate on *features* of the data or make limiting model assumptions. We present the hierarchy discovery algorithm (HDA), a model-based hierarchical clustering method based on explicit comparison of joint distributions via Bayesian network learning for predefined *groups* of data. HDA works on both continuous and discrete data and offers a model-based approach to agglomerative clustering that does not require prespecification of the model dependency structure.

Introduction

Researchers interested in understanding the relationships among groupings of data frequently use hierarchical clustering to assess the similarity of these groups. For example, researchers who are interested in the citation patterns of the scientific community may be interested in the influences within particular venues. A natural question is whether the predictors of citation are the same across venues. An obvious approach would be to apply a hierarchical clustering algorithm. This, however, groups data using the similarity between *features* of venues when the intent is to compare the *underlying generative distributions* of venues. While methods for model-based hierarchical clustering do exist, they require that the form of the dependency structure be specified *a priori.*

We present the hierarchy discovery algorithm (HDA), a method for hierarchical model-based clustering that clusters groups of data using the similarity of their joint distributions without making assumptions about the dependency structure of the data. HDA learnins a Bayesian network for each of a set of predetermined groupings, or partitions, of a dataset. This analysis allows comparison of the joint distributions that generate data instances rather than the specific features of data instances produced by those distributions. This enables researchers to enjoy the benefit of model-based clustering without the need to specify the dependency structure *a priori*, something that is often difficult in practice. Further, there are well-established measures for assessing the similarity between Bayesian networks, such as structural Hamming distance or relative entropy.

A distinguishing feature of HDA is that it operates over *partitionings of data* rather than *individual data instances.* Many large data sets have one or more such pre-existing partitionings that represent geographic regions, temporal periods, organizations, or other natural boundaries that divide the data into subsets. While this heterogeneity may appear to be an inconvenience, it can be leveraged to perform efficient clustering through comparison of the learned joint distributions of the bags.

In this paper, we describe HDA, introduce some suitable candidate measures for comparing models and examine their relative strengths and weaknesses. We then analyze the consistency of HDA in the construction of the hierarchy with experiments on synthetic data. Finally, we show the results of applying the algorithm to the PubMed Open Access Collection of scientific venues and their associated papers.

Presented at the International Conference on Machine Learning (ICML) workshop on Structured Learning: Inferring Graphs from Structured and Unstructured Inputs ,

Hierarchy Learning with Bayesian Networks

Algorithm 1 Hierarchy Discovery Algorithm

```
**Input:** dataset $D$, initial partitioning $P$ of $D$ such that $\bigcup_{i=1}^{|P|} p \in P = D$
$\mathbf{S} = learn\_model(p)\ \forall p \in P$
while $|S| > 1$ do
    find $s_i, s_j$ such that
    $dist(s_i, s_j) = min(dist(s_x, s_y)) \forall x, y \in S$
    $p_{new} = p_i \cup p_j$
    $s_{new} = learn\_model(p_{new})$
    $S = S \backslash \{s_i, s_j\} \bigcup s_{new}$
end while
return $S$
```

Learning the Initial Models

Algorithm 1 specifies the hierarchical discovery algorithm. It takes as input the full dataset, D, and an initial partitioning of the data, P. The initial partitioning may be made in any manner, so long as each partition within the initial partitioning is homogenous, i.e., all data associated with the partition are consistent with a single model. For each partition, p, within the initial partitioning, we learn a Bayesian network representing the underlying joint distribution of p. For this work we used PC, a constraint-based method for learning Bayesian networks (Spirtes et al., 2000). While any algorithm for learning complete Bayesian networks could be used in principle, constraint-based methods such as PC provide an advantage over search and score methods by returning the same model every time the algorithm is applied to data. Search and score-based methods are unable to provide this guarantee due to the inherent stochasticity of the search process. These initial learned models now constitute the set S.

Distance Measures

The algorithm then measures the distance between each pair of models in S. Any symmetric measure of distance could be used in principle. For this paper, we examine the efficacy of three distance measures between Bayesian networks: structural Hamming distance, symmetric relative entropy and proportional loss in Bayesian information criterion. Structural Hamming distance is defined as the L_1 distance between the adjacency matrices of two graphs (Tsamardinos et al., 2006). By using structural Hamming distance between two graphs, the algorithm explicitly compares the *conditional independence structure* of two distributions. This stands in contrast to traditional measures used in hierarchical clustering which seek to minimize some aspect of the grouped feature data but do not explicitly consider the *dependency structure* of the data being compared. While structural Hamming distance provides a principled way for comparing the structure of joint distributions, it does not take into account the strength of dependence between variables. This is problematic in cases where models share dependency structure but differ in parameterization. Take, for example, the models shown in Figure 1. While all three models encode the same dependency structure, it is clear that graph A is much more similar to graph C than to B, given the strength of effects. To account for the strength of effect, we consider both the the loss in the Bayesian information criterion between models and their constituent data and the *symmetric relative entropy* between two Bayesian networks as alternative distance measures.

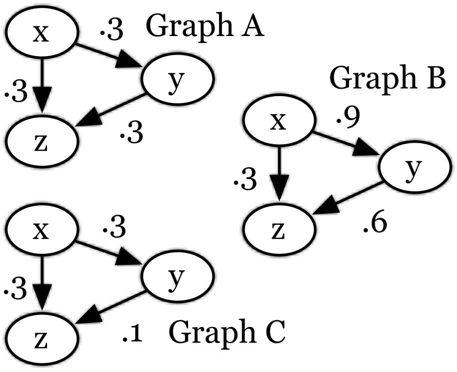

Figure 1. A situation where three models contain the same dependencies but vary in strength of effect.

The *relative entropy* (Koller & Friedman, 2009) between two Bayesian networks $\mathcal{P}$ and $\mathcal{Q}$ where $\mathcal{P}$ is consistent with graph $\mathcal{G}$ is defined as

$$\mathbf{D}(\mathcal{P} \parallel \mathcal{Q}) =$$

$$-H_Q(\mathcal{X}) - \sum_i \sum_{pa_i^{\mathcal{G}}} \mathcal{Q}(pa_i^{\mathcal{G}}) \mathbf{E}_{\mathcal{Q}(X_i|pa_i^{\mathcal{G}})}[ln(P(X_i|pa_i^{\mathcal{G}})]$$

Where $H_{\mathcal{Q}}(\mathcal{X})$ is the entropy of model $\mathcal{Q}$ defined as

$$H_{\mathcal{Q}}(\mathcal{X}) = \sum_i \mathbf{E}_{\mathcal{Q}}[-ln(P(X_i|Pa_i^{\mathcal{Q}})]$$

Relative entropy is known to be asymmetric. We define a symmetric measure by summing the relative en-

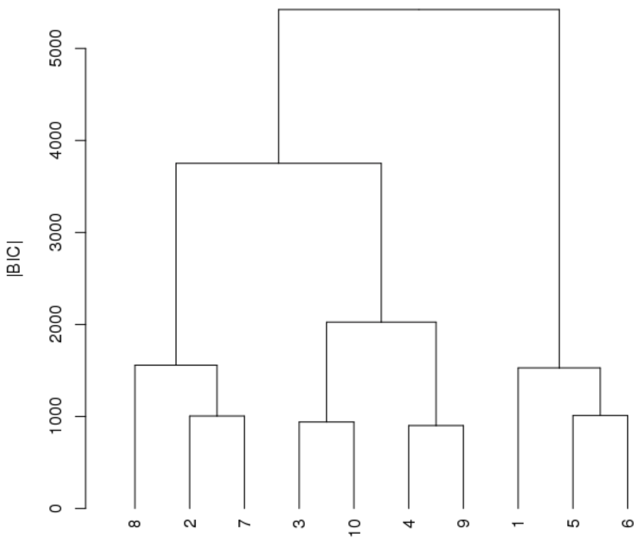

Figure 2. An example hierarchy. The height of each node is determined by the absolute value of the BIC for its associated Bayesian network.

tropy in both directions. Thus our relative entropy-based distance function is defined as

$$dist(\mathcal{Q},\mathcal{P}) = \mathbf{D}(\mathcal{P} \parallel \mathcal{Q}) + \mathbf{D}(\mathcal{Q} \parallel \mathcal{P})$$

The use of relative entropy to compare models requires an assumption that both models are generated from some member of the exponential family. In cases where this is not appropriate, structural Hamming distance and proportional loss in BIC provide more generally applicable measures.

In addition to structural Hamming distance and relative entropy, we define a distance measure that is directly related to the goodness of fit of models to data. Given two joint distributions, P and Q, and their associated datasets, D_P and D_Q, the proportional difference in BIC is defined as:

$$\frac{BIC(Q,D_P)}{BIC(Q,D_Q)} + \frac{BIC(P,D_Q)}{BIC(P,D_P)}$$

Where $BIC(P,D_P)$ is the BIC of the joint distribution of P with regard to dataset D_P This measure is the relative loss in goodness-of-fit of each dataset when applied to the other model. Thus, the metric represents a relative closeness in terms of how exchangeable one model is for the other.

Constructing the Tree

Once the distance has been measured between all models, the pair of models s_i and s_j with the minimum distance are chosen. The data associated with s_i and s_j are then pooled and a new Bayesian network, s_{new}, is learned from the pooled data. We set s_{new} as the *parent* of s_i and s_j and define the distance between each child and its parent to be the difference in Bayesian information criterion between the models. The pair s_i and s_j are then removed from S and s_{new} is added. This process is repeated until there is only one element remaining in S, which is the root node of the hierarchy. Figure 2 is an example of a learned hierarchy.

There is a clear interpretation of each level of the hierarchy. Specifically, if we assume that we have correctly learned the Bayesian networks, for any two levels i and j such that $height(j) < height(i)$:

$$\sum_{p \in P_i} ln\mathcal{L}(\theta_p|x_p) \leq \sum_{q \in P_j} ln\mathcal{L}(\theta_q|x_q)$$

This follows from our initial assumption of homogenous distributions in the initial partitions. This gives rise to an intuitive intrepretation of the tree. As the tree is traversed from the leaves upward, the cost associated with increased generality is reflected with the decreasing likelihood of the models learned. Additionally, if a group of models are very similar, then there will be a small distance between members.

Experiments

Synthetic Data

We compared HDA to feature-based clustering and model-based clustering using a simple marginal model of the data. For the feature-based model, we define the distance function to be:

$$dist(\mathcal{Q},\mathcal{P}) = \frac{1}{|\mathcal{P}||\mathcal{Q}|} \sum_{p \in \mathcal{P}} \sum_{q \in \mathcal{Q}} |q \cap p|$$

which represents the mean correspondence between feature vectors in two groupings. The marginal model naively assumes independence between features and measures distance using symmetric relative entropy. For this experiment, we first generate k Bayesian networks at random with m nodes each. Then, for each model, 500 observations are generated and randomly split into two groups. Each of the split groups is then treated as an initial partition. Hierarchies are then learned on these partitions using structural Hamming distance, symmetric relative entropy, marginal symmetric relative entropy, average vector correspondence, and proportional loss in BIC as distance measures. We also evaluate the efficacy of "clustering" by randomly agglomerating node pairs to establish worse-case performance. We evaluate

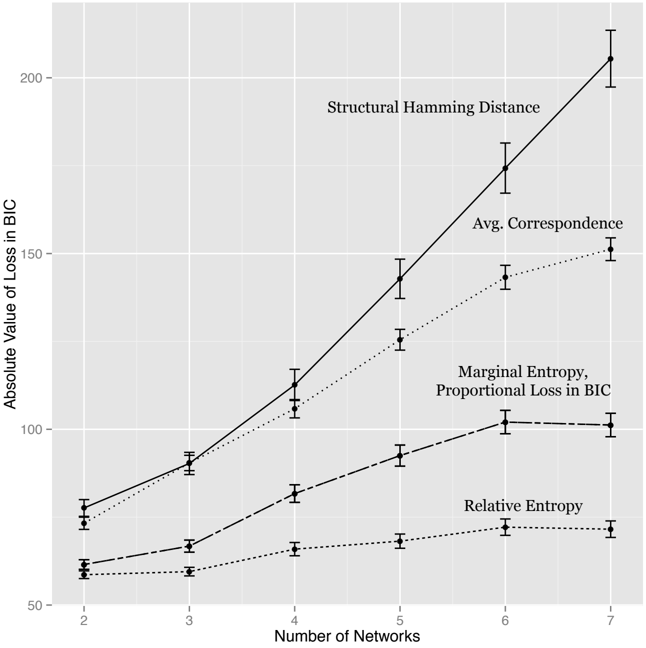

Figure 3. Results from the synthetic split-partition experiment on models with three variables. Random agglomeration, not shown, returns an absolute average BIC difference that is roughly an order of magnitude greater than the models shown.

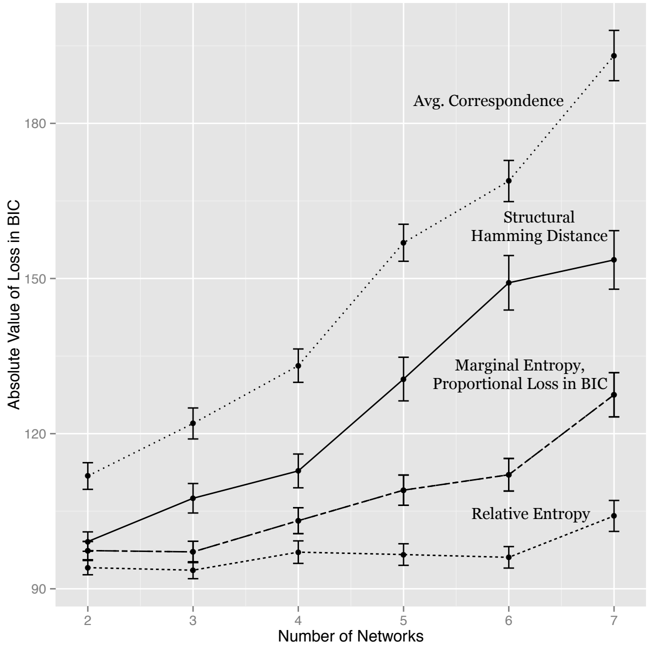

Figure 4. Results from the synthetic split-partition experiment on models with five variables. Random agglomeration, not shown, returns an absolute average BIC difference that is roughly an order of magnitude greater than the models shown.

the performance of the clusterers by determining the distance between each pairing of nodes that were drawn from the same model. That distance is defined to be the average difference in BIC between the lowest common ancestor of the nodes in the learned hierarchy and the nodes themselves. More formally, if two groups m_1 and m_2 are both drawn from a model M and we find their lowest common ancestor in the hierarchy to be a, we define the distance as

$$\frac{1}{2}(BIC(a) - BIC(m_{one}) + BIC(a) - BIC(m_{two}))$$

The best performing model is that which minimizes the sum of tree distances.

The results can be seen in Figures 3 and 4 for models with three and five variables, respectively. All distance measures significantly outperform random agglomeration. Relative entropy provides the best performance, followed by marginal entropy and proportional loss in BIC. Surprisingly, marginal entropy and proportional loss in BIC perform nearly identically in each case. We intend to study this behavior and the relationship between the two measures in the future. In the three-variable case, structural Hamming distance has the worst performance, while average correspondence performs worst in the five-variable case.

The success of relative entropy as a distance measure can be attributed to its ability to leverage both the full conditional structure and parameterization of models. While both marginal entropy and proportional BIC make use of the model parameterization, they do not take full advantage of model structure. Conversely, structural Hamming distance fully considers model structure, but completely ignores the parameterization.

We hypothesize that the relative increase in the perfomance of structural Hamming distance as the number of variables increases can be explained by a boost in discriminative power. The number of possible network structures for models with five variables is much greater than the number possible with three variables. Accordingly, there is a much higher chance of structural overlap between models in the three-variable case. Given that ssuch overlap reduces the discriminative power of structural Hamming distance, we would expect it to perform worse in such an environment.

Real World Data

We also evaluated HDA in a more realistic setting by applying it to the PubMed Open Access Collection, which indexes medical publications that have been re-

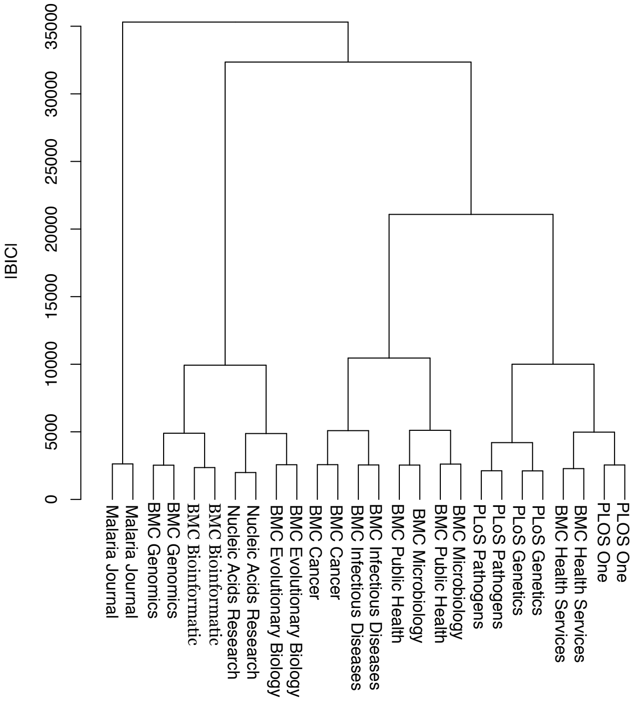

Figure 5. The learned hierarchy for PubMed Open Access Collection with split tags using proportional loss in BIC as a distance measure. Each leaf represents a random 50% of the venues data.

leased to the public under the Creative Commons license.

Each data instance represents a single paper and consists of the following features: number of co-authors, word count of the title, word count of the body, maximimum h-index of all co-authors, average h-index of co-authors, impact factor of the venue of publication, and impact factor of the institution of the primary author. The h-indices and impact factors were derived from the corpus using standard techniques. The initial partitions were defined by the venue of publication. We evaluated this dataset with the same method used for synthetic data. For each venue, we randomly split the data into two equally sized groupings. The hierarchy was then learned over the split venues. Accuracy was defined in terms of the average loss of BIC between the two split venue tags and their lowest common ancestor. Figure 6 summarizes the results. Interestingly, both marginal entropy and proportional BIC outperform relative entropy. We hypothesize that this is because relative entropy requires a correct specification of both model structure and parameters. For a domain with a large amount of uncertainty in model dependencies, as is the case with the PubMed data, the performance of relative entropy suffers.

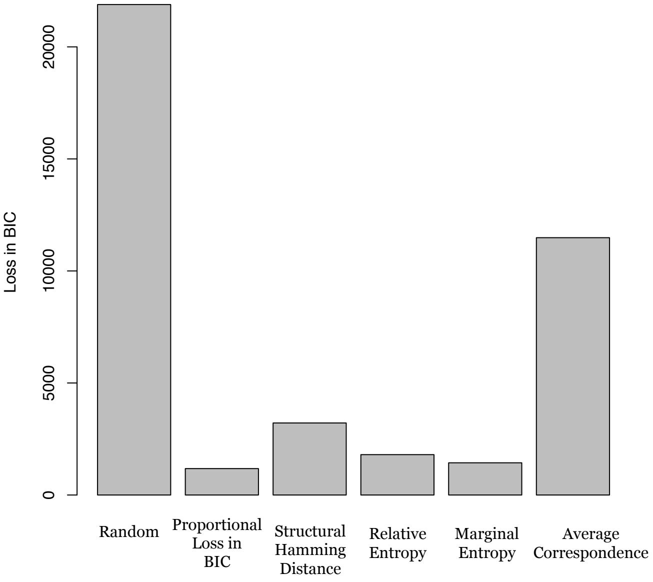

Figure 6. Average Loss in BIC for each hierarchy built with different distance functions.

Related Work

HDA is conceptually very similar to traditional hierarchical clustering techniques (Ward, 1963). However, hierarchical clustering algorithms perform the clustering task by considering the distance between the data features rather than the data model. While this can be an effective technique, it requires that the user supply a distance measure. Constructing such a measure is challenging, particularly when the data consist of a variety of variable types and distributions. Furthermore, the underlying meaning and utility of a candidate measure is not always clear. For instance, it is not immediately obvious how one would measure the distance between two colors. The wavelength of each could be used, but whether this is a useful notion of distance is highly dependent on the task at hand.

Model-based clustering techniques address this by clustering over the likelihood of a set of data given a model. Unlike unconstrained measures of distance, likelihood is a consistent measure with uniform interpretation. While there are existing methods for hierarchical model-based clustering (Vaithyanathan & Dom, 2000), they require that the stucture of a model be specified *a priori*. We relax this requirement by using graphical models, which can represent a much broader class of distributions. While there are other techniques which perform model-based clustering using Bayesian networks (Thiesson et al., 1998), the clustering is not hierarchical and thus does not consider the relationship *between* the models. There has also been some work on clustering based on Bregman divergence (Banerjee et al., 2005), but the authors do not consider graphical models. To our knowledge, there is no existing work on clustering based on the divergence between graphical models learned from data.

Future Work

In principle, this technique can be applied to any task where graphical models can be learned from partitioned heterogeneous data. For instance, one could cluster a set of timeseries by learning a dynamic Bayesian network from each series and constructing a hierarchy from their mutual divergence. Images could be clustered by conditioning a grid-structured conditional random field on each and considering the divergence between the resulting Markov fields. Furthermore, the technique is not limited to propositional models; relational data sets could be clustered based on some notion of distance between learned relational models.

Conclusion

We presented HDA, a method for hierarchical clustering based on the divergence between joint models of partitioned data. This algorithm is particularly useful in large, heterogeneous data domains that contain predefined partitions. We leverage these partitions to learn joint models which provide a notion of distance beyond simple comparison of features. Finally, we demonstrate its effectiveness through synthetic and real-world comparisons with feature-based techniques.

Acknowledgements

The authors would like to thank Cindy Loiselle for her editing expertise. We would also like to acknowledge the members of the Knowledge Discovery Lab for their feedback and commentary, and the Information Extraction and Synthesis Laboratory. This effort is supported by the Intelligence Advanced Research Project Agency (IARPA) via Department of Interior National Business Center Contract number D11PC20152, Science Applications International Corporation (SAIC) and DARPA under contract number P010089628, and the Research Experience for Undergraduates Program of the National Science Foundation under NSF award number CNS-1062749. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of IARPA, DoI/NBC, SAIC, DARPA, NSF, or the U.S. Government.

References

Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit S., and Ghosh, Joydeep. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search.* MIT press, 2nd edition, 2000.

Thiesson, Bo, Meek, Christopher, Chickering, David Maxwell, and Heckerman, David. Learning mixtures of dag models. In *UAI*, pp. 504–513, 1998.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 2006.

Vaithyanathan, Shivakumar and Dom, Byron. Model-based hierarchical clustering. In *In Proc. 16th Conf. Uncertainty in Artificial Intelligence*, pp. 599–608. UAI, 2000.

Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.