Supervised Learning of a Generative Model for Edge-Weighted Graphs

Andrea Torsello
Dipartimento di Informatica
Università Ca' Foscari
Venezia, Italy

David L. Dowe
Clayton School of Information Technology
Monash University
Clayton, Vic. 3800, Australia

Abstract

This paper addresses the problem of learning archetypal structural models from examples. To this end we define a generative model for graphs where the distribution of observed nodes and edges is governed by a set of independent Bernoulli trials with parameters to be estimated from data in a situation where the correspondences between the nodes in the data graphs and the nodes in the model are not not known ab initio *and must be estimated from local structure. This results in an EM-like approach where we alternate the estimation of the node correspondences with the estimation of the model parameters. Parameter estimation and model order selection is addressed within a Minimum Message Length (MML) framework.*

1 Introduction

Graph-based representations have been used with considerable success in computer vision in the abstraction and recognition of object shape and scene structure, as they can concisely capture the relational arrangement of object primitives, in a manner which can be invariant to changes in object viewpoint. Despite their many advantages and attractive features, the methodology available for learning structural representations from sets of training examples is relatively limited, and the process of capturing the modes of structural variation for sets of graphs has proved to be elusive.

Recently, there has been considerable interest in learning structural representations from samples of training data, in particular in the context of Bayesian networks, or general relational models [2]. However, these approaches rely on the availability of correspondence information for the nodes of the different structures used in learning. In many cases the identity of the nodes and their correspondences across samples of training data are not known, rather, the correspondences must be recovered from structure.

In the last few years, there has been some effort aimed at learning structural archetypes and clustering data abstracted in terms of graphs. Torsello and Hancock [6] define a superstructure called tree-union that captures the relations and observation probabilities of all nodes of all the trees in the training set. However, the model structure and model parameter are tightly coupled, which forces the learning process to be approximated through a series of merges. Further, all the observed nodes must be explicitly represented in the model, which then must specify in the same way real structural variations and random noise, limiting the generalization capabilities of the model. In [5] was proposed a generalization for graphs which allowed to decouple structure and model parameters and used a stochastic process to marginalize the set of correspondences, however the approach does not deal with attributes and all the observed nodes still need be explicitly represented in the model. Further, the issue of model order selection was not addressed.

The aim in this paper is to develop an information-theoretic framework for the learning of generative models of graph-structures from sets of examples. The major characteristics of the model are the fact that the model structure and parameters are decoupled, and that we have two components to the model: one which describes the *core* part, or the proper set of structural variations, and one which defines an isotropic random structural noise.

2 Generative Graph Model

Consider the set of undirected graphs $S = (g_1, \ldots, g_l)$. Our goal is to learn a generative graph model $\mathcal{G}$ that can be used to describe the distribution of structure in S. To develop this probabilistic model, we make an important simplifying assumption: We assume that the observation of each node and each edge is independent of the others. Hence, the proposed structural model is a complete graph $\mathcal{G} = (V, E, \Theta)$, where $V = \{1, \ldots, n\}$ is the set of nodes, $E \subseteq V \times V$ is the set of edges and Θ is a set of observation probabilities such that node $i \in V$ is present with probability θ_i. Futher, edge (i, j) is present with probability τ_{ij}, conditioned on the fact that both nodes i and j are present.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

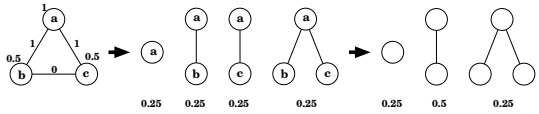

Figure 1. The graph observation process.

Weight distributions can be added to the node and edge models to deal with weighted graphs. In this paper we assume that graphs are attributed with edge lengths, and we model this observed quantity as a normal distribution of mean μ_{ij} and variance σ_{ij}^2 for each edge (i,j). These parameters will be learned together with the other model parameters.

We work under the assumption that the correspondence information between the observation's nodes and the model's nodes that generated them is not known. We can model this by saying that an unknown random permutation is applied to the nodes of the sample. For this reason, the observation probability of a sample graph depends on the unknown correspondences between sample and model nodes. Figure 1 shows a graph model and the graphs that can be generated from it with the corresponding probabilities. Here the numbers next to the nodes and edges of the model represent the values of θ_i and τ_{ij}. Note that, when the correspondence information (letters in the Figure) is dropped, we cannot distinguish between the second and third graphs anymore, yielding the final distribution.

With this definition, since every node in the generated graphs must originate from a node in the model, the only operation we can do to the structural model to generate a new graph is the removal of nodes and edges. This implies that the model must describe every possible structural variation encountered in the data, be it central to the distribution, or simply structural noise that is encountered with very low probability. To avoid this we allow for nodes to be added to the model by saying that, with a certain probability. There *external* nodes have identical probability $\bar{\tau}$ of being connected to any other node and the same length distribution of parameters $\bar{\mu}$ and $\bar{\sigma}^2$, where we force the observation probability to be equal to the average density of the core part of the structural model and the length model parameters to be equal to the average over the parameters of the core model. Hence, external nodes model isotropic (or spherical) noise with the same average edge distribution as the core model. In general, a generative model will generate a graph with k external nodes according to a geometric distribution $P_k = (1-\bar{\theta})\bar{\theta}^k$, where $\bar{\theta} \in [0,1]$ is a model parameter that quantifies the tendency of the model to generate external nodes.

Let us assume that we have a model $\mathcal{G}$ with n nodes and that we want to compute the probability that graph g with m nodes was sampled from it. Let g be a graph and $\sigma : (1,\dots,n) \to (1,\dots,m+1)$ be a set of correspondences from the model nodes to the nodes in g where $\sigma(i) = m+1$ if model node i has no corresponding node in g, that is, if model node i is not observed in graph g. Further, let $\pi : (1,\dots,m) \to (1,\dots,n+1)$ be the inverse set of correspondences, where $\pi(h) = n+1$ if h is an external node, otherwise $\sigma(\pi(h)) = h$, and $\pi(\sigma(i)) = i$ if $\sigma(i) \neq m+1$. With this notation, the probability that a graph g was sampled from a model $\mathcal{G}$ given the correspondences σ and π is

$$P(g|\mathcal{G},\sigma) = (1-\bar{\theta}) \prod_{i=1}^{n}\prod_{j=i}^{n} \Theta_{ij}^{\sigma(i)\sigma(j)} \prod_{h=1}^{m}\prod_{k=h}^{m} \overline{\Theta}_{\pi(h)\pi(k)}^{hk} ,$$

where Θ_{ij}^{hk} is the probability that model edge (i,j) generated graph edge (h,k), $\overline{\Theta}_{ij}^{hk}$ with $i = n+1$ or $j = n+1$ is the probability that edge (h,k) is external to the model, and pairs with the same index represent a node instead of and edge. Letting $G = (g_{hk})$ be the adjacency matrix of graph g, and letting $\Pr(x|\mu,\sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$, we define Θ_{ij}^{hk} and $\overline{\Theta}_{ij}^{hk}$ as follows:

$$\Theta_{ij}^{hk} = \begin{cases} 0 & \text{if } i=j \wedge h \neq k \text{ or } i \neq j \wedge h = k \\ \theta_i & \text{if } i=j \wedge h=k \wedge G_{hh}=1 \\ 1-\theta_i & \text{if } i=j \wedge h=k \wedge G_{hh}=0 \\ \tau_{ij}\Pr(l_{ij}|\mu_{ij},\sigma_{ij}^2) & \text{if } i \neq j \wedge h \neq k \wedge G_{hk}=1 \\ 1-\tau_{ij} & \text{if } i \neq j \wedge h \neq k \wedge G_{hh}=1 \wedge \\ & G_{kk}=1 \wedge G_{hk}=0 , \\ 1 & \text{otherwise.} \end{cases}$$

$$\overline{\Theta}_{ij}^{hk} = \begin{cases} 0 & \text{if } i=j \wedge h \neq k \text{ or } i \neq j \wedge h = k \\ \bar{\theta} & \text{if } h=k \wedge i=j=n+1 \\ \bar{\tau}\Pr(l_{ij}|\bar{\mu},\bar{\sigma}^2) & \text{if } (i=n+1 \vee j=n+1) \wedge G_{hk}=1 \\ 1-\bar{\tau} & \text{if } (i=n+1 \vee j=n+1) \wedge G_{hk}=0 \\ 1 & \text{otherwise.} \end{cases}$$

3 Model Estimation

Key to the estimation of the structural model is the realization that, condition on a given set of node correspondences, the node observation processes are independent from one another. Hence, since the structural component of the model is always a complete graph and node/edge observation is dictated by the model parameters, knowing the set of correspondences would effectively decouple parameters and structure.

Here we make the simplifying assumption that the likelihood of the set of correspondences σ_g between graph g and model $\mathcal{G}$ is strongly peaked, i.e., we have $P(g|\mathcal{G}) \approx \max_{\sigma_g} P(g|\mathcal{G},\sigma_g)$. With this assumption the estimation of the structural model can be achieved with an EM-like process by alternating the estimation of the correspondences σ_g of every graph $g \in S$ with a fixed set of model parameters Θ, and the estimation of Θ given the correspondences.

While this EM-like approach solves the problem of estimating the structural model of a given size, the problem of model order selection remains open. We have

chosen to use Minimum Message Length (MML) criterion [7] which allows us to address parameter estimation and model order selection within a single framework, solidly based on information theory.

3.1 Correspondence Estimation

The estimation of the set of correspondences σ is an instance of a graph matching problem., where, for each graph g, we are looking for the set of correspondences that maximizes $P(g|\mathcal{G},\sigma)$. To do this we relax the space of partial correspondences, where a relaxed state is represented by a $(n+1)\times(m+1)$ matrix $P=(p_{ih})$ satisfying the constraints $p_{ih}\geq 0$, $\sum_{h=1}^{m+1} p_{ih}=1$, and $\sum_{i=1}^{n+1} p_{ih}=1$. The matrix P is almost doubly stochastic, with the exception for the last row and column that are not normalized. The probability $P(g,\mathcal{G},\sigma)$ can be extended to the relaxed assignment space using as:

$$E(g,\mathcal{G},P)=(1-\bar{\theta})\Big(\prod_{i=1}^{n}\prod_{j=i}^{n}\sum_{h=1}^{m+1}\sum_{k=h}^{m+1} p_{ih}\Theta_{ij}^{hk}p_{jk}\Big)\cdot \Big(\prod_{h=1}^{m}\prod_{k=i}^{m}\sum_{i=1}^{n+1}\sum_{j=i}^{n+1} p_{ih}\bar{\Theta}_{ij}^{hk}p_{jk}\Big). \quad (1)$$

In an approach similar to Graduated assignment [3], we maximize the energy function E by iterating the recurrence $P^{t+1}=\mu\big(DE^t\big)$, where DE^t is the differential of E with respect to P^t and μ is a function projecting DE^t to the relaxed assignment space.

3.2 Parameter Estimation

MML is a Bayesian method of point estimation based on an information-theoretic formalization of Occam's razor. Here, simplicity of an explanation is formalized as the joint cost of describing a probabilistic model for the data and describing the data given the model. Hence, to estimate a model class and the model parameters, MML minimizes a two-part message. The first encodes the model class/order and the parameters, while the second assumes a Shannon-optimal encoding of the data given the model. MML is closely related to the Kolmogorov complexity, is invariant under 1-to-1 parameter transformations [7], and has general statistical consistency properties [1].

The cost of describing a fully specified model (in the first part of the message) with a parameter vector $\theta_{\mathcal{G}}$ of dimension D is approximately

$$-\log\left[\frac{h(\theta_{\mathcal{G}})}{\sqrt{F(\theta_{\mathcal{G}})}}\right]+\frac{D}{2}\log k_D,$$

where k_D are the lattice constants specifying how tightly unit spheres can be packed in a D-dimensional space, $h(\theta)$ is the prior of the parameters θ, $F(\theta)$ is the Fisher information matrix (Wallace and Freeman, 1987) and the term $1/(\sqrt{k_D^D F(\theta_{\mathcal{G}})})$ is the optimal round-off in the parameter estimates.

According to Shannon's theorem, the cost of encoding the data (in the second part of the message) has a tight lower bound in the negative log-likelihood function, to which we add the roundoff error $D/2$.

In this work we have opted for a standard non-informative Jeffreys's prior for the model parameters which will push the parameters towards the edges of their range forcing each node/edge to be observed either very frequently or very rarely. A consequence of this choice is that the MML point estimates of the parameters are the same as the maximum likelihood estimates, leaving the MML criterion only for model-order selection. In fact, the use of Jeffrey's prior implies $h(\theta)=\sqrt{F_1(\theta)}$, where $F_1(\theta)$ is the single datum Fisher information matrix and $F(\theta)=|S|^D F_1(\theta)$. Hence, the final description length, approximating $\log(k_D)$ as described in [7], is

$$I_1=\frac{D}{2}\log\left(\frac{|S|}{2\pi}\right)+\frac{1}{2}\log(\pi D)-1-\sum_{g\in S}\log P(g|\mathcal{G},\sigma_g),$$

where $|S|$ is the number of samples and the number of parameters for a n-node structural model is $D=3\binom{n-1}{2}+n+2$.

Recalling the maximum likelihood estimates for binomial distribution, we have $\theta_i=\frac{a_i}{|S|}$, where a_i is the number of graphs that observe model node i, i.e., that have a node mapped to i, $\tau_{ij}=\frac{|\{g\in S|(\sigma_g(i),\sigma_g(j))\in E_g\}|}{a_{ij}}$, where a_{ij} is the number of graphs that observe both nodes i and j, and $\bar{\theta}=\frac{u}{u+|S|}$, where u is the set of external nodes that do not map to any node in the model.

Concluding, given a set of observation graphs S and a model dimension n, we jointly estimate node correspondences and model parameters by alternating the two estimation processes in an EM-like approach, and then we chose the model order that minimizes the message length I_1.

4 Experimental Evaluation

We tested our structural learning approach on shock-graphs [4], a skeletal-based representation of the differential structure of the boundary of a 2D shape. We have used a database consisting of 72 shapes from the MPEG7 database, divided into 6 classes of 12 shapes each. The shape classes where composed of bottles, children, hands, glasses, horses, and tools. The size of the shock-graphs varied from 4 to 20 nodes. We have learned a model for each shape class and computed the sampling probability of each graph from each model.

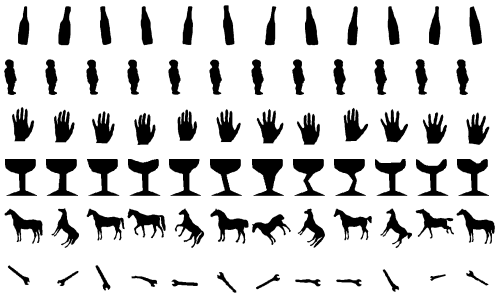

Figure 2. The shape classes in the database.

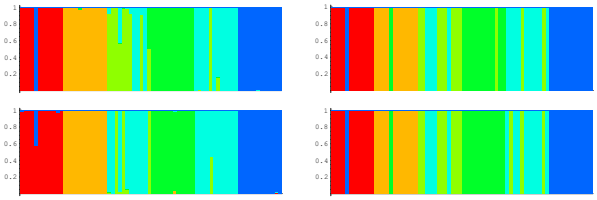

Figure 3. Assignment probability of the graphs to the learned models. Top: full database, bottom: reduced database. Left: the proposed model, right: NN rule.

For comparison, we also computed the structural similarities using Graduated assignment [3].

Figure 3 plots the model-assignment probability for each graph, i. e., a stacked histogram of the model probabilities normalized over the sum of all model probabilities associated with each graph, where he colors of the classes are as follows: Bottle (red), Child (orange), Hand (light green), Glass (dark green), Horse (light bue), and Tool (dark blue). Figure 3 shows on the top left the assignment of graphs to classes according to the proposed approach, while on the top right it plots the assignments obtained using the nearest neighbor (NN) rule based on the distances obtained using Graduated assignment. Here we can see that in most cases shock-graphs are predominantly assigned to the correct class,while NN has a significantly higher rate of misclassifications of 14% versus the 7% misclassification we obtained with our approach. Furthermore, it should be noted that NN classification is computationally more demanding than the classification using our structural models, as NN requires computing the similarity against each training graph, while our approach requires computing the probabilities only against the learned models. Clearly our approach requires the models to be learned ahead of time, but that can be performed offline. To assess the generalization capabilities of the approach we have repeated the experiment using only 6 shapes to learn the models. The bottom row of Figure 3 plot the model assignments obtained using our approach and the NN rule. We can clearly see that the approach generalizes fairly well in both cases, with the probabilities approximately distributed in the same way as those obtained from the full training set, resulting in a 13% misclassification for our approach and 14% for NN classification.

5 Conclusions

In this paper we have proposed an approach to the problem of learning a generative model of edge-weighted graphs from examples in a situation where the correspondences between the nodes in the data graphs and the nodes in the model are not not known *ab initio* and must be estimated from local structure. To this end, we defined a structural model which is learned with an EM-like approach where we alternate the estimation of the node correspondences with the estimation of the model parameters. Parameter estimation and model order selection are jointly addressed within a Minimum Message Length (MML) framework. Experiments on a shape recognition task show that the approach is effective in characterizing the modes of structural variation present in a set of graphs.

Acknowledgment

We acknowledge the financial support of the Future and Emerging Technology (FET) Programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open project SIMBAD grant no. 213250.

References

[1] D. L. Dowe, S. Gardner, and G. R. Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, 58(4):709–754, 2007.

[2] N. Friedman and D. Koller. Being Bayesian about Network Structure, *Machine Learning*, 50:95–125, 2003.

[3] S. Gold and A. Rangarajan. A graduated Assignment Algorithm for Graph Matching. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(4):377–388, 1995.

[4] B. B. Kimia, A. R. Tannenbaum, and S. W. Zucker. Shapes, shocks, and deformations I: the components of shape and the reaction-diffusion space. *Int. J. Computer Vision*, vol. 15, no. 3, pp. 189–224, 1995.

[5] A. Torsello. An Importance Sampling Approach to Learning Structural Representations of Shape. In *IEEE Int Conf. on Computer Vision and Pattern Recognition*, 2008.

[6] A. Torsello and E. R. Hancock. Learning Shape-Classes Using a Mixture of Tree-Unions. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(6):954-967, 2006.

[7] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer Verlag, ISBN 0-387-23795-X, 2005.