ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Visualisation of Tree-structured Data through Generative Probabilistic Modelling

Nikolaos Gianniotis, Peter Tiňo

School of Computer Science - University of Birmingham
Birmingham B15 2TT - UK

Abstract.
We present a generative probabilistic model for the topographic mapping of tree structured data. The model is formulated as constrained mixture of hidden Markov tree models. A natural measure of likelihood arises as a cost function that guides the model fitting. We compare our approach with an existing neural-based methodology for constructing topographic maps of directed acyclic graphs. We argue that the probabilistic nature of our model brings several advantages, such as principled interpretation of the visualisation plots.

1 Introduction

The Self Organising Map (SOM) [1] has inspired numerous extensions for the topographic organisation of non-vectorial forms of data, such as sequences or trees [2, 3]. Such approaches attempt to introduce a notion of *context* that is updated in an recursive manner and is supposed to represent data items processed until the current competition step. This is realised with additional feed-back connections that allow for natural processing of recursive data types. Typical examples of such models are e.g. merge SOM [4] and SOM for structured data [5].

The heuristic nature of SOM and its extensions inherently brings about certain limitations. One of the major limitations is the lack of a principled cost function that quantifies the topographic organisation of the map (although see developments in e.g. [6]). This introduces difficulties in the comparison of map formations resulting from different initialisations, parameter settings, or optimisation algorithms.

The second major limitation is the inability of such approaches to deliver a model-based interpretation of the visualisation result. Clusters may be formed on the map that indicate some close relationship between the concerned structured data items, but there is no explanation on what the characteristics of the cluster are. Of course one can inspect the individual data points to deduce those relationships once the map has been formed, but reasoning about mapping of new data items (not used for model fitting) is still quite problematic.

To address those limitations, we introduce a model based approach to constructing topographic maps of tree-structured data formulated in a principled framework of probability theory. Generative probabilistic modelling brings a number of advantages, e.g. greater explanatory power, principled handling of missing data and hierarchy construction.

2 An overview of Hidden Markov Tree Models

A tree $\boldsymbol{y}$ is an acyclic directed graph and as such it consists of a set of nodes $u \in \mathcal{U}_{\boldsymbol{y}} = \{1, 2, ..., U_{\boldsymbol{y}}\}$, a set of directed edges between the nodes (each edge connects a parent node to a child node) and a set of labels $\boldsymbol{o}_u \in \mathbb{R}^d$ on nodes u. Each node u has a single parent $\rho(u)$, apart from node number one, the root node. Conversely each node has a set of children, apart from the leaf nodes.

Density over tree-structured data can be modelled e.g. by a hidden Markov tree Model (HMTM) [7] (analogous to hidden Markov model (HMM) [8] for sequential data). Each node u can be in one of K discrete states $q_u \in \{1, 2, ..., K\}$. A HMTM is defined by three sets of parameters: the initial probability distribution that describes the state q_1 of the root, the transition probability distribution that describes the transitions between parent and child states, $p(q_u|q_{\rho(u)})$, and the emission parameters that parametrise Gaussian distributions, $f(.; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, one for each state k. which emit the labels. Here, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrix, respectively, of the Gaussian associated with emission process in state k.

The HMTM distribution factorises as follows:

$$P(\boldsymbol{y}) = \sum_{\boldsymbol{q} \in \{1,2,...,K\}^{U_{\boldsymbol{y}}}} P(q_1) \prod_{u \in \mathcal{U}_{\boldsymbol{y}}, u \neq 1} P(q_u|q_{\rho(u)}) \prod_{u \in \mathcal{U}_{\boldsymbol{y}}} P(\boldsymbol{o}_u|q_u), \qquad (1)$$

where $\boldsymbol{q} \in \{1, 2, ..., K\}^{U_{\boldsymbol{y}}}$ is the set of all $U_{\boldsymbol{y}}$-tuples over K hidden states. Similarly to the forward-backward algorithm for HMM [8], the likelihood of an HMTM can be efficiently computed by the upward-downward algorithm [7].

3 HMTMs as noise models for GTM

This section presents an extension of GTM from vectorial to tree structured data in the spirit of [9], where GTM is extended to visualise sequential data. Due to space limitations only the model formulation will be presented, detailed derivations and more involved developments will be presented elsewhere. Our model, GTM-HMTM, is a mixture of HMTMs. In order to have the HMTM components topologically organised we constrain the mixture of HMTMs, by requiring that the HMTM parameters be generated through a parameterised *smooth* nonlinear mapping from the latent space $[-1, +1]^2$ into the HMTM parameter space. As in GTM, we discretise the latent space into a regular grid of points $\boldsymbol{x}_c$, $c = 1, 2, ..., C$. Each latent centre $\boldsymbol{x}_c$ will correspond to a component HMTM $p(\boldsymbol{y}|\boldsymbol{x}_c)$ with flat mixing coefficient $1/C$.

Given a dataset $\mathcal{T} = \{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(N)}\}$ of N independently generated trees the model likelihood is proportional to:

$$\mathcal{L} \propto \prod_{n=1}^{N} \sum_{c=1}^{C} \sum_{\boldsymbol{q} \in \{1,2,...,K\}^{U_n}} p(q_1|\boldsymbol{x}_c) \prod_{u=2}^{U_n} p(q_u|q_{\rho(u)}, \boldsymbol{x}_c) \prod_{u=1}^{U_n} P(\boldsymbol{o}_u^{(n)}|q_u, \boldsymbol{x}_c), \qquad (2)$$

where U_n stands for the number $U_{\boldsymbol{y}^{(n)}}$ of nodes of tree $\boldsymbol{y}^{(n)}$.

The smooth nonlinear mapping yields the following sets of parameters for each latent point $\boldsymbol{x}_c$:

$$\begin{aligned}
\boldsymbol{\pi}_c &= \{p(q_1 = k|\boldsymbol{x}_c)\}_{k=1:K} = \{g_k(\boldsymbol{A}^{(\boldsymbol{\pi})}\boldsymbol{\phi}(\boldsymbol{x}_c))\}_{k=1:K} \\
\boldsymbol{T}_c &= \{p(q_u = k|q_{\rho(u)} = l, \boldsymbol{x}_c)\}_{k,l=1:K} = \{g_k(\boldsymbol{A}^{(\boldsymbol{T}_l)}\boldsymbol{\phi}(\boldsymbol{x}_c))\}_{k,l=1:K} \\
\boldsymbol{B}_c &= \{\boldsymbol{\mu}_k^{(c)}\}_{k=1:K} = \{\boldsymbol{A}^{(\boldsymbol{B}_k)}\boldsymbol{\phi}(\boldsymbol{x}_c)\}_{k=1:K}, \quad \text{where}
\end{aligned}$$

- the function $g(.)$ is the softmax function, which is the canonical inverse link function of multinomial distributions and $g_k(.)$ denotes the k-th component returned by the softmax.
- $\boldsymbol{x}_c \in \mathbb{R}^2$ is the c-th grid point,
- $\boldsymbol{\phi}(.) = (\phi_1(.), ..., \phi_M(.))^T, \phi_m(.) : \mathbb{R}^2 \rightarrow \mathbb{R}$ is an ordered set of M non-parametric nonlinear smooth basis functions (typically RBFs),
- the matrices $\boldsymbol{A}^{(\boldsymbol{\pi})} \in \mathbb{R}^{K \times M}$, $\boldsymbol{A}^{(\boldsymbol{T}_l)} \in \mathbb{R}^{K \times M}$ and $\boldsymbol{A}^{(\boldsymbol{B}_k)} \in \mathbb{R}^{d \times M}$ are the free parameters of the model.

The model likelihood is maximised using the expectation-maximisation (EM) algorithm.

Regarding the covariance of the emission distribution, we noticed that higher quality models were obtained when instead of direct modelling of the covariance, the covariance was calculated, in the spirit of [10], at the end of each M-step using standard update equations. Having trained the model, we can then represent each data item $\boldsymbol{y}^{(n)}$ with a point $\mathfrak{p}^{(n)}$ in the latent space given by the expectation of the posterior distribution over all latent space: $\mathfrak{p}^{(n)} = \sum_{c=1}^{C} p(\boldsymbol{x}_c|\boldsymbol{y}^{(n)})\boldsymbol{x}_c$.

4 Experiments

We have used two datasets in our experiments. The first is an artificial toy dataset produced by sampling from 4 HMTMs with 2 hidden states with 2-dimensional Gaussian emissions of fixed spherical variance, each corresponding to one artificial class. All patterns have the topology of a binary tree with 15 nodes. The parameters of the models were set so as to ensure that it would be impossible to distinguish the classes from the observations alone, without taking into account the underlying tree structure.

The second dataset consists of images produced by the *Traffic Policeman Benchmark* (TPB) used to demonstrate the functionality of SOM for Structured Data (SOMSD) in [5]. The images resemble traffic policemen, houses and ships of different shape and size. Connected components in each image have a parent-child relationship[1]. Nodes are labelled with a 2-dimensional vectors denoting the centre of gravity of the component that node stands for. The dataset defines 12 classes that are represented on the plots with 12 different markers.

[1] we have restricted TPB to produce only images expressed as trees

Both datasets were normalised in each dimension to zero mean and unit standard deviation. The lattice was a 10x10 regular grid (i.e. $C = 100$) and the RBF network consisted of $M = 17$ basis functions; 16 of them were Gaussian radial basis functions of variance $\sigma^2 = 1$ centred on a 4x4 regular grid, and one was a constant function $\phi_{17}(\boldsymbol{x}_c) = 1$ (for a bias term). Training starts with random parameters initialised with uniform distribution in $[-1, 1]$.

Figure 1(a) shows the GTM-HMTM topographic organisation of the toy dataset[2] for $K = 2$. Each point on the plot represents an input pattern (tree). Four different markers correspond to the four generative classes used to construct the data set. Training is completely unsupervised and class markers are used only after the training when plotting the projections. A clear topographic organisation of classes has been achieved - there is an evident trend of patterns of the same class to belong to the same cluster.

Figure 2(a) shows the visualisation of the TPB dataset[3] produced by GTM-HMTM with $K = 2$. In figure 2(a), next to each cluster a representative image is displayed. The model has clearly achieved a level of topographic organisation. It is interesting to note the emerging sub-clusters. Class $\times$ has been split into two sub-clusters, one with policemen with the right arm lowered and one with the right arm raised. The same has happened for class $\bigcirc$ which has been divided into policeman with the right arm lowered and policemen with the arm raised. Also for the class of ships with two masts, class $*$, it is interesting to note that it has been divided into three sub-clusters. Nevertheless, the model has not been successful in the visualisation of the classes representing houses. No clusters have been formed as all classes have been merged into one big cluster representing a superclass of all the images of houses. Moreover, we attempted training for $K = 3, 4$, but with suboptimal results.

As a comparison, in figures 1(b),2(b) we also present the results obtained by using SOMSD on the two datasets. We tried numerous parameter settings for SOMSD and picked the best results[4]. On the toy dataset GTM-HMTM performs better (but note that the dataset may be biased toward GTM-HMTM), while SOMSD is better at the TPB dataset. It manages to distinguish between all of the classes, especially the classes of houses that are problematic in GTM-HMTM. On the other hand, SOMSD does not discover the subclasses that GTM-HMTM does for the policemen and ships.

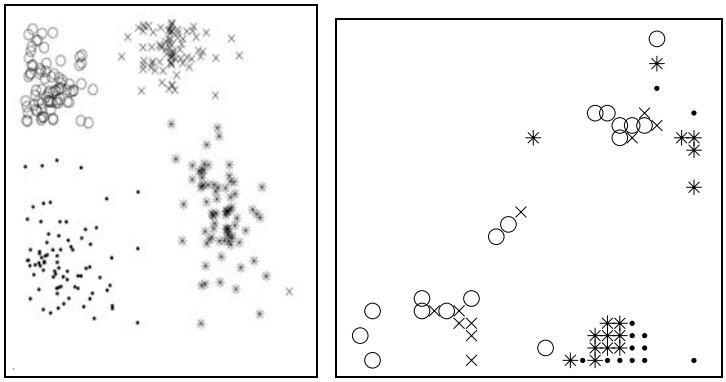
5 Discussion

Because of the absence of a clear cost function, the performance of SOMSD was measured in [5] as the accuracy of classification of data. After the map

[2]Covariance of the emission distribution was initially set to $\Sigma_k = 2I$ for both states $k = 1, 2$ (I stands for the identity matrix). We also tried initialising it with $\Sigma_k = 2I, 3I, 5I$ with similar success.

[3]Initial covariance matrix for the emission distribution was set to $\Sigma_k = 2I$ for both states $k = 1, 2$. We also tried initialising the covariance matrix with $\Sigma_k = 1I, 3I$ which yielded similar results and $\Sigma_k = 0.5I$ which failed to achieve the same level of topographic organisation.

[4]according to the measure in [5]

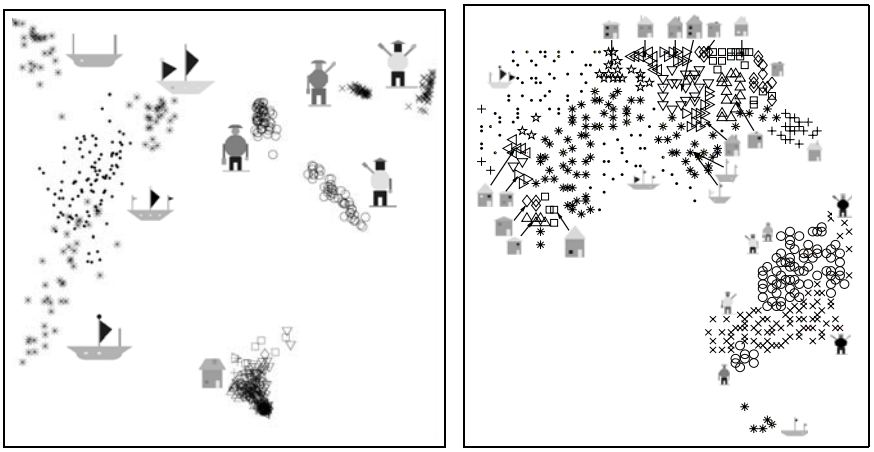


(a) (b)

Fig. 1: Visualisation of toy dataset using GTM-HMTM (a) and SOMSD (b).

formation, a secondary hold-out test dataset was used. Items from the test set were represented on the trained map and each test item was predicted to have the class label of its closest neighbour from the training set on the map. The accuracy was then defined as the percentage of correctly classified test points. The results of this measure on the toy dataset were 90% and 60% for GMT-HMTM and SOMSD respectively. The results were reversed as for the TPB dataset GMT-HMTM and SOMSD achieved 55% and 95% of accuracy respectively.

We argue that such a procedure makes sense only when the class organisation of the data correlates with the driving force behind topographic map formation. If for example, the classes of trees are organised along the lines that cannot be reasonably captured by HMTM modelling, there is simply no reason why the achieved classification accuracy of GMT-HMTM should be high. But low classification rate would just mean that our model-driven topographic map formation does not correlate well with the particular class labelling scheme. In such cases one can simply switch to local noise models that are more correlated with the class labelling. Alternatively, one might say that he/she wanted to see topographically organised data representations driven by aspects captured by HMTM (or any other noise model employed) and stick with the obtained topographic maps, irrespective of the class labels. This is an unsupervised learning setting after all... Again, without knowing the exact mechanism behind the topographic map formation, it is problematic to assign any performance-related interpretation to the classification rate obtained on the trained map. On the other hand, the probabilistic nature of GTM-HMTM allows us to objectively and in a principled manner evaluate and compare the models as density estimators.



(a) (b)

Fig. 2: Visualisation of TPB dataset using GTM-HMTM (a) and SOMSD (b)

References

[1] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, September 1990.

[2] G de A. Barreto, A.F.R. Araújo, and S.C. Kremer. A taxanomy of spatiotemporal connectionist networks revisited: The unsupervised case. *Neural Computation*, 15:1255–1320, 2003.

[3] B. Hammer, A. Micheli, M. Strickert, and A. Sperduti. A general framework for unsupervised processing of structured data. *Neurocomputing*, 57:3–35, 2004.

[4] M. Strickert and B. Hammer. Merge SOM for temporal data. *Neurocomputing*, 64:39–72, 2005.

[5] M. Hagenbuchner, A. Sperduti, and Ah C. Tsoi. A self-organizing map for adaptive processing of structured data. *Neural Networks, IEEE Transactions on*, 14(3):491–505, 2003.

[6] T. Heskes. Energy functions for self-organizing maps. In S. Oja and E. Kaski, editors, *Kohonen Maps*, pages 303–315. Elsevier, Amsterdam, 1999.

[7] J.-B Durand, P. Goncalves, and Y. Guedon. Computational methods for hidden markov tree models-an application to wavelet trees. *Signal Processing, IEEE Transactions on*, 52(9):2552–2560, 2004.

[8] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[9] Peter Tino, Ata Kaban, and Yi Sun. A generative probabilistic approach to visualizing sets of symbolic sequences. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–706. ACM Press, 2004.

[10] Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.