ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 459-464

Visualization and clustering with Categorical Topological Map

Mustapha. LEBBAH[a], Fouad. BADRAN[b], Sylvie. THIRIA[a,b]
a- CEDERIC, Conservatoire National des Arts et Mtiers,
292 rue Saint Martin, 75003 Paris, France
b- Laboratoire LODYC, Université Paris 6, Tour 26-4[e] étage,
4 place Jussieu 75252 Paris cedex 05 France

Abstract

This paper introduces a topological map dedicated to cluster analysis and visualization of categorical data. Usually, when dealing with symbolic data, topological maps use an encoding stage: symbolic data are changed into numerical vectors and traditional numerical algorithms are run. In the present paper, we propose a probabilistic formalism where neurons are now represented by probability tables. Two examples using actual and synthetic data allow to validate the approach. The results show the good quality of the topological order obtained as well as its performances in classification.

1 Introduction

The topological map proposed by Kohonen [10] use a self-organization algorithm (SOM) which provides quantification and clustering of the observation space. Bishop et al [2] have recently introduced a latent-variable density model called Generative Topographic Mapping (GTM), which is closely related to the SOM. Kaban et al [9] proposed a method for analysis and visualisation of binary data based on GTM. In the paper of Lebbah et al [5], we presented specific topological maps dedicated to binary data. In this paper we generalize the proposed approach to categorical data [6]. The model we propose named Categorial Topological Map (CTM), uses a probabilistic formalism and a learning procedure to maximize the likelihood function of the data set. In section 2, we present the CTM algorithm and its learning procedure based on the EM algorithm. In section 3, we show how estimated probabilities allow to compute a-posteriori probabilities, allowing CTM to act as soft classifier. The validation of CTM approach is presented in section 4: two different applications on synthetic and real data show the ability of CTM to deal with categorial data.

2 Categorical Topological Map (CTM)

Let $A = \{\mathbf{z}_i, i = 1..N\}$ the learning data set. We assume that a given observation $\mathbf{z}_i$ is a M dimentional vector $\mathbf{z}_i = (z_i^1, z_i^2, ..., z_i^k, ..., z_i^M)$ where the kth component z_i^k is a categorical variable with n_k modalities taking its value in

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 459-464

$\mathcal{A}_k = \{\xi_1^k, \xi_2^k, ..., \xi_{n_k}^k\}$, such as each observation $\mathbf{z}_i$ is thus, a realization of a random variable which belongs to $\mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_M$. We assume also that the M categorical components of a given observation $\mathbf{z}$ are independant. Under this assumption $p(\mathbf{z}) = \prod_{k=1}^{M} p(z^k)$ where each $p(z^k)$ is a one dimensional table which represents the probabilities of the n_k modalities of $\mathcal{A}_k$. The independency assumption is necessary for computation purpose, nevertheless results obtained on real data for with this assumption is usally false, show the robustness of CTM. The topological order and the estimated probabilities have high quality. The goal of this paper is to present a new topological map dedicated to these categorical data. As for the traditional topological maps, the map $\mathcal{C}$ has a discrete topology defined by an undirect graph. Usually this graph is a regular grid in one or two dimensions. We denote N_{neuron} the number of neurons in $\mathcal{C}$. For each pair of neurons (c,r) on the map, the distance $\delta(c,r)$ is defined as the shortest path between c and r on the graph. In the following, we introduce a kernel positive function K ($\lim_{|\delta|\to\infty} K(\delta) = 0$) and its associated family K_T parametrized by T : $K_T(\delta) = [1/T]K(\delta/T)$, which controls the size of the neighborhood. Following Luttrel [8], Kaban [9] and Anouar et al [1], we introduce a probabilistic formalism to deal with topological map dedicated to categorical data. We assume that the map $\mathcal{C}$ is duplicated in two similar maps $\mathcal{C}_1$ and $\mathcal{C}_2$ provided with the same topology as $\mathcal{C}$. At each neuron $c_1 \in \mathcal{C}_1$, we associate M probability tables denoted by $p(z^k/c_1)$ where $(k = 1 \dots M)$. The kth probability table is defined by the n_k values of $p(z^k = \xi_j^k/c_1)$, $(j = 1..n_k)$. As before, the probability $p(\mathbf{z}/c_1)$ can be expressed under the independency hypothesis of its component as : $p(\mathbf{z}/c_1) = \prod_{k=1}^{M} p(z^k/c_1)$. For each neuron c_1, this expression describes the observations generated by c_1. We assume that each neuron c_1 of $\mathcal{C}_1$ is a distortion of a neuron c_2 of $\mathcal{C}_2$ and this distortion is described by the probability $p(c_1/c_2)$. In order to introduce a topological order we assume that: $p(c_1/c_2) = [1/T_{c_2}]K_T(\delta(c_1, c_2))$, where $T_{c_2} = \sum_{r \in C_2} K_T(\delta(c_2, r))$. Under the "Markov" property : $p(\mathbf{z}/c_1, c_2) = p(\mathbf{z}/c_1)$, the probability distribution of the observations generated by a neuron c_2 of $\mathcal{C}_2$ is a mixture of probabilities completely defined from the map and the probability tables $p(\mathbf{z}/c_1)$, $p(\mathbf{z}/c_2) = \sum_{c_1 \in C_1} p(c_1/c_2)p(\mathbf{z}/c_1)$. Finally, the probability $p(\mathbf{z})$, is defined on $\mathcal{C}_2$ as : $p(\mathbf{z}) = \sum_{c_2 \in C_2} p(c_2)p(\mathbf{z}/c_2)$. In this expression, $p(c_2)$ represents the a priori probability of the neuron c_2. The aim of the learning algorithm is to estimate all the parameters of the model. These parameters are: the N_{neuron} parameters $\theta^{c_2} = p(c_2)$ and for each neuron c_1 the different values $\theta_j^{k,c_1} = p(z^k = \xi_j^k/c_1)$. In the following we denote : $\theta^{k,c_1} = \{\theta_j^{k,c_1}, j = 1 \dots n_k\}$ and $\theta^{c_1} = \cup_k \theta^{k,c_1}$ the parameters which define neuron c_1. We assume that the N observations of the learning set $\mathcal{A}$ are independant and generated according to $p(\mathbf{z})$. The learning procedure estimates the model parameters by maximizing the likelihood of the observations: $p(\mathbf{z}_1, \mathbf{z}_2, ...\mathbf{z}_N) = \prod_i p(\mathbf{z}_i)$, for this purpose we use the EM algorithm (Dempster et al (1977)). Since our model suggests that the observations are generated by two cells c_1 and c_2, we introduce, as hidden variables, the boolean variables $\chi_i^{c_1,c_2} = 1$, if $\mathbf{z}_i$ is generated by c_1 and c_2 and is equal to 0 otherwize. So each observation $\mathbf{z}_i$ is associated a hidden variable χ_i whose components are the $\chi_i^{c_1,c_2}$. The application of EM gives rise to the iterative algorihtm of CTM, which for a given value of T maximizes the likelihood of

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 459-464

$\mathcal{A}$. In order to present this algorithm, we define the conditional probabilities $p(c_2/\mathbf{z}_i)$ and $p(c_1, c_2/\mathbf{z}_i)$, which can be expressed as:

$$p(c_1, c_2/\mathbf{z}_i) = \frac{\theta^{c_2} K_T(\delta(c_1, c_2)) p(\mathbf{z}_i/c_1)}{\sum_{r \in C_2} \theta^r p(\mathbf{z}_i/r)} \tag{1}$$

$$p(c_2/\mathbf{z}_i) = \sum_{c_1 \in C_1} p(c_1, c_2/\mathbf{z}_i) \tag{2}$$

The 2 steps CTM algorithm is expressed as follow :

- **Intialisation Step :** Choose an initial values θ_0^{c1} and θ_0^{c2} for the parameters.
- **The iteration step :** Compute the current values of $p(c_1, c_2/\mathbf{z}_i)$ and $p(c_2/\mathbf{z}_i)$ by applying equations (1) and (2) and the new parameters using the equations 3 and 4:

$$\theta^{c_2} = \frac{\sum_{i=1}^{N} p(c_2/\mathbf{z}_i)}{N} \tag{3}$$

$$\theta_j^{k,c_1} = \frac{\sum_{i \in \tau_{k,j_0}} p(c_1/\mathbf{z}_i)}{\sum_{j=1..n_k} \sum_{i \in \tau_{k,j}} p(c_1/\mathbf{z}_i)} \tag{4}$$

  In the last equation $\tau_{k,j} = \{i \text{ such that } z_i^k = \xi_j^k\}$ represents the set of observations $\mathbf{z}_i$ for whom the kth component takes the modality ξ_j^k.
- **Repeat** the iterative procedure until convergence.

In this presentation T is kept fixed, but T is also a parameter of the model since it is used to compute the probabilities $p(c_1/c_2)$. In order to minimize the likelihood with respect to T we can, as in the Kohonen algorithm [10], repeat the preceding procedure for different values of T decreasing it from an initilal value T_{max} to a final value T_{min}. The required value T^* is the one which minimizes the likelihood function.

3 CTM-Classifier

The map provided by CTM can be used in classification tasks, combining supervised and unsupervised learning. If we denote by $L = \{l_i, i = 1..S\}$ the labels used for different classes, and if each observation $\mathbf{z}$ is assigned to a particular class l_i; the CTM algorithm can be used as a "soft" classifier computing the a-posterior probaility of each label as :

$$p(l_i/\mathbf{z}) = \sum_{c_1 \in C_1} p(l_i/c_1) p(c_1/\mathbf{z}) \tag{5}$$

where $p(c_1/\mathbf{z}) = \sum_{c_2 \in C_2} p(c_1, c_2/\mathbf{z})$. and $p(l_i/c_1) = \frac{n_{c_1}^{l_i}}{n_{c_1}}$; n_{c_1} represents the number of observations of the learning data set assigned to neuron c_1 by the assignement function $\chi(\mathbf{z}) = \text{argmax}_{c_1} p(c_1/\mathbf{z})$ from which $n_{c_1}^{l_i}$ have label l_i. The accuracy of these probabilities depend both on the size of the set learning data set and on the topological order.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 459-464

4 Experiments

In the following we used Categorical Topological Maps (CTM) for an automatic classification of two distinct samples. The first experiment deals with artificial data, which have been created for comparison purposes (Leich et al [7]). Using this data base allows to cmpare CTM with the performances provided by several cluster algorithm dedicated to binary data. The second experiment is dedicated to a behavioral survey. This example clearly shows the adequacy of CTM to perform an accurate analysis of high dimentional data. **In the first experiment**, the comparison of CTM with other clustering algorithms has been made using binary data distributed on the web site www.wu-wien.ac.at/am, (Dolinicar et al [4]). We extract from the benchmark two different data bases made of 6000 individuals. These data are artificial data simulated for comparison purposes proposed by authors; the simulation mimic typical situations from tourism marketing. The tourists are classified in 6 classes according to their answer ("Yes" or "No") to twelve questions. Six different synthetic data bases with increasing difficulties are available; we select the two most difficult problems according to their performances computed from Bayes classifier (scenario 5ind, scenario 5dep). Two 2-D topological maps with 5×5 neurons were trained using CTM. At the end, we assign each observation of the data bases on the corresponding map using CTM-classifier (formula 5). The comparison with seven different clustering algorithms (Hard Competitive Learning with Euclidian Distance (HCL-ED) or Absolute Distance (HCL-AD), Neural Gas NGAS-ED or NGAS-AD, k-Means and Self Organizing Map SOM) are presented in table 1. The authors in [4] estimated the classification rate using the learning data set, for comparison purpose we do the same. Table 1 provides the classification rate provided by each algorithm together with the theoritical Bayes rate. Clearly CTM allows approach the theoritical Bayes classification rate. **In the second experiment** we process a behavioral survey of very large dimension. The survey consists in answers given by individuals and concerning their feeling about 70 words (as death, war, flower, to charm, to buy, gifts...). Each individual gives a note between "1" and "7" for each word; this note represents the feeling he/she has about the word ("1" stands for a very bad feeling). The data base is made of 1228 individuals. Each one is represented by a vector of 70 categorical variables; each variable has 7 modalities. A CTM map with 7×7 neurons is trained using the whole data set. In the following we present some possible way

	HCL-ED	HCL-AD	k-means	NGAS-ED	NGAS-AD	SOM	**CTM**	**TBR**
5ind	71%	83%	51%	71%	74%	51%	**85%**	**89%**
5dep	49%	58%	48%	52%	59%	49%	**71%**	**79%**

Table 1: Comparison of the classification perfomances reached by CTM and seven clustering algorithms on the two simulated data sets (scenario 5ind and scenario 5dep). TBR represent the Theoretical Bayes Rate

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 459-464

to analyze the results of the clustering in order to show the consistency of the CTM approach. At the end of the learning phase, each neuron is associated

	War	War	War	War,Honest	War,Honest	War,Honest
	War	War	War	Stranger,War, Immobility, Souvereign	War,Honest	Gift,Courage,War, Honest, Infinit,Soft, Death, Politeness
War	War		War,Death	War, Immobility	Flower, War, House	Gift, Courage,Dynamic, Flower,War,Honest, Death,Respect, Victory
Immobility				Distress,War, Politeness	Gift,War, Death	Distress,Gift,Courage, Glory, Humour,House, Death, Ocean, Perfume,Wealth, Charm,
				War	Gift,Courage, Flower,War, Honest,House, House, Death, Charm.	Money,Jewel,Gift, Courage,Dynamic, Flower,War,Honest, Humour,Intimate, House,Death, Politeness,Protect, Respect,Wealt,Charm ,Victory
			War	Flower,War	Flower,War	Charity,Courage, Discipline,Dynamic, Flower,War,Honest, Law,House,Politness, Protect,Respect,Work ,Victory
Dynamic,Humour, Immobility, Intimate, Skin		War, Humour	Flower,War, Humour, Intimate,		Flower,War	Glory, War

Figure 1: 7×7 Map. Each cell of the grid represents a neuron of the map, in ecach one we plot the words which have notes with a probability greater than 0.8

with 70 probability tables with seven values corresponding to the probabilities of the seven notes. In Figure 1, we plot for each neuron the words which have notes with a probability greater than 0.8. So, for each neuron, the selected words represent a shared feeling and can give an interpretation of the map. Looking carefully at the different neurons, reveals that close neurons have similar feelings; the topological order presents a coherent clustering. If we focus the analysis to note "1", (for each neuron , we look for words whose modality "1" is greater than 0.8, this words are undelined in figure 1). It can be seen that there is a concensus of opinion with regard to the word "war" which is badly noted by must of the pepole. If the probability is decreased up to 0.6, the word "war" appear for each neuron of the map and the word "death" is mostly associated to it, this enlighten the general feeling. We now present for the two words "to

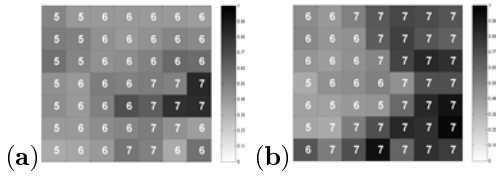

Figure 2: Topological map giving a posterior probability of the most probable note.(a): for the word "**to charm**", (b):for the word "**Flower**". For each cell the neuron, the number indicates the most probabble note and the gray scale stands for its probability $p("note"/c_1)$ (white =0, black =1)

charm" and "Flower", the probability associated with the most probable note.

(see figure 2). Figure 2.a and 2.b, show a the spacial coherence with respect to the note. The two maps present similar patterns enlignting the strong correlation linking the two words. In the present paper we choose some few words which show the ability of CTM to extract some information embedded in the survey. As the CTM map summarize a large amount of information, more complex analysis can be made using computed probabilities.

5 Conclusion

In this paper, we presented a new algorithm dedicated to topological maps and categorical data. For this purpose, we used a probabilistic formalism which allows to maximize the likelihood of the data. This formalism allows to define an assignement function based on a-posteriori probability. We applied CTM on two data bases with different level of complexity; these experiments prove the ability of CTM to deals with classification and visualization tasks. If we look at these data bases used, we can see that one is made of independant categorical variables when the two others used dependant categorical variables. Though the assumption is necessary for computation purpose, it is interesting, looking at results, to see that good results (topological order, estimated probabilities) can be obtained when this assumption is not verified.

Acknowledgement: The authors would like to acknowledge Ludovic Lebart (Director of research at CNRS) for providing us the second data base, and making useful comments about the results.

References

[1] Anouar, F. Badran, F. Thiria, S. (1998): Probabilistic self-organizing map and radial basis function networks. Neurocomputing 20, 83-96.

[2] Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998): GTM: The Generative Topographic Mapping. Neural Computation, 10(1), 215-234.

[3] Dempster,A. P. Laird, N. M. Rubin, D. B. Maximum liklihoode from incomplete data via the EM algorithm. Journal of royal Statistic Society, Series B, 39, 1-38

[4] Dolinicar, Weingessel , A. Buchta, C. (1998): Dimitriadou, E. A Comparaison of several cluster algorithms on artificial binary data, scenarios from travel market segmentation. Working paper series 19, SFB (adaptive information systems and modelling in economics and management science).

[5] Lebbah, M. Badran, F. (2000): Thiria, S. Topological Map for Binary Data, ESANN 2000, Bruges, April 26-27-28, Proceedings

[6] Lebbah. M , Thiria. S, Badran. F, Chabanon. C. ICANN 2002, Categorical Topological map, Madrid 2002.

[7] Leich, F. Weingessel, A. Dimitriadou, E. (1998): Competitive Learning for Binary Data. Proc of ICANN'98, septembre 2-4. Springer Verlag.

[8] Luttrel S. P. (1994). A Bayesian Ananlysis of Self-Organizing Maps, Neural Computing vol 6.

[9] Kaban, A and Girolami, M. (2001): A Combined Latent Class and Trait Model for the Analysis and Visualisation of Discrete Data. I.E.E.E Transactions on Pattern Analysis and Machine Intelligence.23(8), pp859 -.872.

[10] Kohonen, T. (1994): Self-Organizing Map. Springer, Berlin.