ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

Representing hierarchical relationships using a modified Asymmetric SOM Algorithm

Manuel Martin-Merino [1] and Alberto Muñoz [2]

[1] University Pontificia de Salamanca, C/Compañía 5, 37002 Salamanca, Spain
Email: manuel@upsa.es
[2] University Carlos III de Madrid, C/Madrid 126, 28903 Getafe, Spain
Email: albmun@est-econ.uc3m.es

Abstract. Self organizing maps (SOM) are useful visualization techniques that have been successfully applied to the analysis of multivariate data. However most of the algorithms proposed in the literature are not able to handle asymmetric similarities. Therefore they are not able to deal with hierarchical relationships in an appropriate manner although this feature is crucial for many practical applications.

In this paper we propose a new asymmetric SOM batch algorithm (ASOM) suitable to derive hierarchical relationships from an asymmetric similarity measure. Object relations are shown by a trapezoidal grid of neurons. The first coordinate represents the object proximities while the second one visualizes the object's degree of generality. A kernel version is also proposed that improves the network organization by transforming nonlinearly the original dissimilarity. The new models have been applied to the challenging problem of thesaurus generation with remarkable results.

1 Introduction

Self organizing maps (SOM) are nonlinear visualization techniques that show in an intuitive way multidimensional object relationships [6]. They have been applied successfully to many practical problems [6]. However, most of the algorithms proposed in the literature are not able to deal with asymmetric proximities although this kind of measures allow to model hierarchical relationships in a straightforward way [7, 11].

Several authors have pointed out that the skew symmetric component of a similarity matrix conveys valuable information to organize the objects hierarchically [7, 11, 8]. To get a deeper understanding of this property, let examine the problem of thesaurus generation. Consider for instance the relation between a broad term such as "Mathematics" and a specific term such as "Bayesian".

Financial support from DGICYT grant BEC2000-0167 (Spain) is gratefully appreciated.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

Obviously, the first term contains the semantic meaning of the second one while the reverse statement is weaker. This asymmetric relation suggests that both terms are related but the first one (broad term) should be assigned to the upper levels of the hierarchy while the second one (specific term) should belong to a lower level. Therefore an important goal (not previously considered in the literature) is the development of asymmetric algorithms suitable to visualize hierarchical relationships.

In this paper we first present a batch version of the ASOM algorithm [9]. This model is a generalization of the SOM algorithm [6] that incorporates asymmetric proximities and that is able to represent hierarchical relationships. Therefore our algorithm differs from other models presented in the literature (see for instance [10]) that are based on symmetric distances. Next we propose a kernel version of the ASOM that transforms nonlinearly the original dissimilarities to improve the network organization. The minimization of the error function in feature space is done efficiently with no additional computational burden over the classic algorithm.

This paper is organized as follows. In section 2 we present the batch version of the ASOM algorithm. In section 3 a kernelized version of previous algorithm is presented. In section 4 the new models are applied to the generation of topic hierarchies and finally in section 5 we get conclusions and outline future research trends.

2 The ASOM algorithm

In this section we present the batch version of the ASOM algorithm. This is a new SOM based topographic mapping algorithm that incorporates asymmetric measures to induce automatically the object hierarchy. The model is based in the work presented in [9] and therefore the necessary background will be explained concisely.

The ASOM algorithm represents input vectors by neurons arranged along a trapezoidal grid (see figure 1). Neighboring neurons along the X axis will represent objects spatially close in input space. Likewise, neighboring neurons along the Y coordinate will represent objects of similar degree of (semantic) generality.

To achieve the previous goal the ASOM algorithm proceeds in two steps:
1. Voronoi tessellation: A quantization algorithm is run using the following dissimilarity measure,

$$D'(\boldsymbol{x}_\mu, \boldsymbol{w}_s) = \beta(\mid \boldsymbol{x}_\mu \mid - \mid \boldsymbol{w}_s \mid)^2 + (1-\beta)(\boldsymbol{x}_\mu - \boldsymbol{w}_s)^T(\boldsymbol{x}_\mu - \boldsymbol{w}_s). \qquad (1)$$

The first term in equation (1) is proportional to the skew symmetric component of the fuzzy logic similarity measure [8] and will become smaller for objects with similar L_1 norm. In the application considered in this paper the L_1 norm is proportional to the number of documents indexed by the corresponding term (see [8] for details). Obviously this coefficient will become large for broad terms,

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

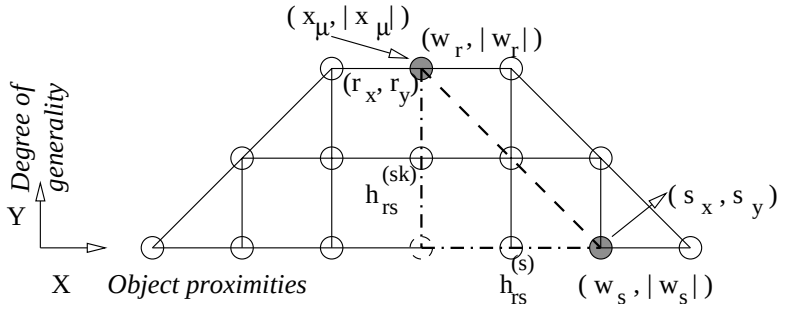

Figure 1: ASOM topology. $h_{rs}^{(s)}$ and $h_{rs}^{(sk)}$ denote the neighborhood functions that transform the symmetric (X) and skew-symmetric (Y) components.

which appear in a large number of documents. Therefore, intuitively speaking, the first term in equation (1) accounts for differences in the object degree of generality.
The second term in equation (1) measures the object proximities according to the Euclidean distance. The β parameter acts as a weighting factor between both terms. Values of about 0.6 gives good results for the problem at hand. Therefore, the dissimilarity (1) induces Voronoi regions that group together objects that are spatially close and also share a similar degree of generality.

2. Error quantization optimization: In this step the prototypes are organized along the trapezoidal grid by the optimization of the following quantization error:

$$E(\mathcal{W}) = \sum_r \sum_{\hat{x}_\mu \in V_r'} \left[\sum_s h_{rs}^{'(sk)} (\mid \boldsymbol{x}_\mu \mid - \mid \boldsymbol{w}_s \mid)^2 + \sum_s h_{rs}^{(s)} (\boldsymbol{x}_\mu - \boldsymbol{w}_s)^T (\boldsymbol{x}_\mu - \boldsymbol{w}_s) \right], \tag{2}$$

where V_r' denotes the Voronoi region computed using the dissimilarity (1) and $h_{rs}^{'(sk)}$, $h_{rs}^{(s)}$ are the neighborhood Gaussian functions for the Y and X coordinates on the trapezoidal grid (see figure 1). The parameters of both functions are updated iteratively using the rule $[\sigma(t) = \sigma_i(\sigma_f/\sigma_i)^{t/t_{max}}]$ proposed by [5] for the SOM batch algorithm.
The first term in equation (1) will be minimized if neighboring neurons along the Y coordinate ($h_{rs}^{'(sk)}$ large) represent objects with similar L_1 norm. Therefore, after convergence, consecutive levels of the hierarchy will represent objects of similar degree of specificity. The second term in (1) organizes the terms along the X axis according to their semantic relations, minimizing the same error that the classic SOM [6].

Assuming that the whole dataset is available, the error function (2) can be easily optimized by solving the set of linear equations $\partial E(\mathcal{W})/\partial w_{sk} = 0$. After that, a simple updating solution for the prototype coordinates is obtained.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

$$\omega_{sk} = \frac{\sum_r \sum_{\hat{x}_\mu \in V'_r} h^{(s)}_{rs} x_{\mu k}}{\sum_r \sum_{\hat{x}_\mu \in V'_r} h^{(s)}_{rs}}. \quad (3)$$

$$\mid \boldsymbol{w}_s \mid = \frac{\sum_r \sum_{\hat{x}_\mu \in V'_r} h^{'(a)}_{rs} \mid \boldsymbol{x}_\mu \mid}{\sum_r \sum_{\hat{x}_\mu \in V'_r} h^{'(a)}_{rs}}. \quad (4)$$

Notice that the solution maintains the same level of simplicity than the SOM batch algorithm originally proposed by [6].

3 A new Kernel ASOM algorithm

Object L_1 norm histogram is very skew and follows a Zipf law [1] for several applications, including those concerning textual data analysis. Consequently, the L_1 norm difference in equation (1) gets close to 0 frequently. As several authors have explained [4, 8] this feature significantly degrades the performance of any algorithm based on distances. In particular, the organization of the Y ASOM coordinate will be negatively affected. Moreover, the Voronoi regions induced by the dissimilarity (1) will be distorted by the same problem. In this section we propose to transform nonlinearly the dissimilarity (1) using the kernel trick [12]. The nonlinear transformation of the dissimilarities will help to improve both, the organization of the Y component and the quality of the Voronoi regions.

Let $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ be a Mercer kernel [12]. That is, there exists a nonlinear map ϕ such that $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$. Let $\hat{\boldsymbol{x}}_\mu = (\boldsymbol{x}_\mu, \mid \boldsymbol{x}_\mu \mid)$ denotes the input patterns represented in $\mathbb{R}^{d+1}$, where d is the input space dimension, $\boldsymbol{x}_\mu$ denotes the vectorial representation in $\mathbb{R}^d$ and $\mid \boldsymbol{x}_\mu \mid$ the L_1 norm. The mapping ϕ can be extended to the $\hat{\boldsymbol{x}}_\mu$ vectors by letting $\Phi(\hat{\boldsymbol{x}}_\mu) = (\phi(\boldsymbol{x}_\mu), \phi'(\mid \boldsymbol{x}_\mu \mid))$. Finally let $\hat{\boldsymbol{w}}_t = (\boldsymbol{w}_t, \mid \boldsymbol{w}_t \mid)$ denotes the prototypes of ASOM in feature space.
The kernel ASOM algorithm proceeds in two steps:
1. Voronoi tessellation: A quantization algorithm is run in feature space. The dissimilarity (1) in feature space can be written as:

$$D'(\Phi(\hat{\boldsymbol{x}}_\mu), \hat{\boldsymbol{w}}_t) = \beta\,[\phi'(\mid \boldsymbol{x}_\mu \mid) - \mid \boldsymbol{w}_t \mid]^T [\phi'(\mid \boldsymbol{x}_\mu \mid) - \mid \boldsymbol{w}_t \mid] \quad (5)$$

$$+ \quad (1-\beta)\,[\phi(\boldsymbol{x}_\mu) - \boldsymbol{w}_t]^T [\phi(\boldsymbol{x}_\mu) - \boldsymbol{w}_t], \quad (6)$$

where ϕ and ϕ' are the nonlinear mapping for the object coordinates and the L_1 norm respectively. The Euclidean distance in feature space induces an alternative dissimilarity in input space that is expected to model better the

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

object proximities. Using the Mercer's property of the kernel, the previous dissimilarity can be rewritten exclusively in terms of kernel evaluations:

$$\begin{aligned} D'(\Phi(\hat{\boldsymbol{x}}_\mu), \hat{\boldsymbol{w}}_t) &= k(\boldsymbol{x}_\mu, \boldsymbol{x}_\mu) + k'(\mid \boldsymbol{x}_\mu \mid, \mid \boldsymbol{x}_\mu \mid) \\ &- 2\sum_i \alpha_{ti} k(\boldsymbol{x}_\mu, \boldsymbol{x}_i) + \alpha'_{ti} k'(\mid \boldsymbol{x}_\mu \mid, \mid \boldsymbol{x}_i \mid) \\ &+ \sum_{ij} \alpha_{ti}\alpha_{tj} k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \alpha'_{ti}\alpha'_{tj} k'(\mid \boldsymbol{x}_i \mid, \mid \boldsymbol{x}_j \mid), \end{aligned} \tag{7}$$

where k, k' are the kernels associated to the mappings ϕ, ϕ'.
2. Quantization error minimization: The quantization error in feature space is obtained just transforming nonlinearly the input patterns via the mapping Φ.

Considering that the prototypes in feature space can be written as $\boldsymbol{w}_t = \sum_i \alpha_{ti}\phi(\boldsymbol{x}_i)$ and $\mid \boldsymbol{w}_t \mid = \sum_i \alpha'_{ti}\phi'(\mid \boldsymbol{x}_i \mid)$ [12] the quantization error can be expressed in terms of scalar products only. Therefore the function error can be rewritten in terms of kernel evaluations as:

$$\begin{aligned} E(\mathcal{W}) &= \sum_r \sum_{\Phi(\hat{x}_\mu)\in V'_r} \sum_t h_{rt}^{'(a)} \Bigg[k'(\mid \boldsymbol{x}_\mu \mid, \mid \boldsymbol{x}_\mu \mid) - 2\sum_i \alpha'_{ti} k'(\mid \boldsymbol{x}_\mu \mid, \mid \boldsymbol{x}_i \mid) \\ &+ \sum_{ij} \alpha'_{ti}\alpha'_{tj} k'(\mid \boldsymbol{x}_i \mid, \mid \boldsymbol{x}_j \mid) \Bigg] + \sum_t h_{rt}^{(s)} \Bigg[k(\boldsymbol{x}_\mu, \boldsymbol{x}_\mu) - 2\sum_i \alpha_{ti} k(\boldsymbol{x}_\mu, \boldsymbol{x}_i) \\ &+ \sum_{ij} \alpha_{ti}\alpha_{tj} k(\boldsymbol{x}_i, \boldsymbol{x}_j) \Bigg]. \end{aligned} \tag{8}$$

The above function error can be easily optimized by solving the set of linear equations $[\partial E(\mathcal{W})/\partial \alpha_{ti} = 0]$ in dual space. The matrix $\boldsymbol{\alpha} = (\alpha_{ti})$ of coefficients are given by

$$\boldsymbol{\alpha} = 2\mathbf{k}^{+}\mathbf{a}\mathbf{N}^{-1} \tag{9}$$

$$\boldsymbol{\alpha}' = 2\mathbf{k}^{'+}\mathbf{a}'\mathbf{N}^{'-1}, \tag{10}$$

where $\mathbf{k}^{+}$, $\mathbf{k}^{'+}$ are the pseudoinverse of the kernel matrices, $\mathbf{N} = diag(\sum_r N_r h_{rt}^{(s)})$ and $\mathbf{a} = \sum_r h_{rt}^{(s)} \sum_{\phi(x_\mu)\in V'_r} k(\boldsymbol{x}_i, \boldsymbol{x}_\mu)$. Finally, $\mathbf{a}'$, and $\mathbf{N}'$ are defined in the same way, substituting $h_{rt}^{(s)}$ by $h_{rt}^{(sk)}$.
Notice that the solution requires only the computation of an SVD that can be done efficiently for sparse data [2].

4 Experimental results

In this section our algorithms are applied to the visualization of topic hierarchies in which terms should be organized attending their semantic meaning

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 231-238

and their degree or generality. To this aim, we have built a collection made up of 2000 documents recovered from three commercial databases "LISA", "INSPEC" and "Sociological abstracts". The documents group in 7 topics according to the thesaurus available.

The mapping algorithms are evaluated through several objective functions. First we check if the neurons of larger Y coordinate represent broader terms than neurons of smaller Y coordinate. To this aim, we plot the average L_1 norm for every neuron of each level versus the index neuron. Each neuron is labeled with the level of hierarchy to which it belong.
Next we determine, for each level of the hierarchy, if words belonging to the same group according to the ASOM network, are related in the thesaurus. To this aim, the prototypes for each level of the ASOM network are clustered using a $k-$means based algorithm [5]. Next the word clusters are evaluated through the following objective measures: The F measure [1] shows if words related in the thesaurus are clustered together by the network. The entropy measure [1] gives the uncertainty for the classification of words assigned to the same clustered. Small values suggest little overlapping in the map between words belonging to different topics. Finally, the mutual information [1] is a nonlinear correlation measure between the classifications induced by the network and the thesaurus. Larger values are preferred.
The most relevant results are the following:

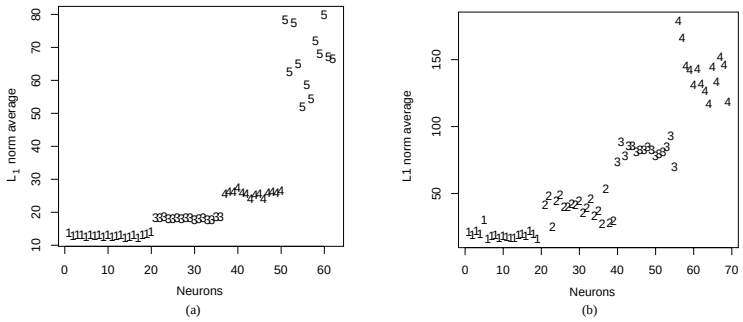

Figure 2: L_1 norm average for neurons corresponding to 5 levels.(a) ASOM. (b) Polynomial kernel ASOM of degree 3.

Figure 2 shows that the organization of the Y component for the ASOM algorithms is excellent. In fact, each pair of neurons verifies that whenever $\mid \boldsymbol{w}_k \mid > \mid \boldsymbol{w}_l \mid$ then $Y_k > Y_l$. This property, allow us to interpret the network as a hierarchy in which broad terms appears always above more specific terms in the grid of neurons. On the other hand, the experiments show that the linear ASOM algorithm represent the non specific terms ($\mid \boldsymbol{x}_i \mid > 30$) just by one level in the hierarchy. Otherwise, the kernel alternative represents better this group by two levels. Consequently, the kernel version achieves a smoother variation of the term degree of generality with the Y coordinate.

Tables 1(a)-1(b) suggest that the semantic relations induced by each level 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. 231-238

Level	F	E	I
5	0.73	0.41	0.30
4	0.63	0.47	0.23
3	0.59	0.40	0.22
1	0.49	0.39	0.16
Average	0.61	0.41	0.23

(a)

Level	F	E	I
4	0.83	0.16	0.31
3	0.72	0.22	0.30
2	0.54	0.37	0.17
1	0.63	0.30	0.22
Average	0.68	0.26	0.25

(b)

Parameters:Nneur$_x = 20$, Nneur$_y = 5$, niter $= 20$, $\beta = 0.65$; (a) $\sigma_{xi} = 11$, $\sigma_{xf} = 1$, $\sigma_{yi} = 4$, $\sigma_{yf} = 0.25$. (b) $\sigma_{xi} = 9$, $\sigma_{xf} = 1$, $\sigma_{yi} = 4$, $\sigma_{yf} = 0.3$.

SOM Batch	F	E	I
	0.71	0.36	0.24

(c)

Table 1: Objective measures for each level of the hierarchies. (a) ASOM batch. (b) Kernel ASOM with polynomial kernels of degrees 1 and 3 for the symmetric and skew symmetric components. (c) SOM batch.

the hierarchy are good according to the thesaurus. In particular, the average measures for the whole hierarchy are only slightly worse than for the SOM batch [6] algorithm. This behavior may be justified because the number of neurons per level (20) is significantly smaller than the number of neurons for the linear SOM (80) considered in this paper.

The kernel ASOM algorithm outperforms the linear version. The position of specific terms is improved in average ($\Delta I = 9\%$) as well as the general word map quality ($\Delta F = 11\%$). We remark that the overlapping between diferent topics in the grid of neurons is in average significantly reduced ($\Delta E = 36\%$) by the kernel version. This trend becomes stronger for the first levels of the hierarchy improving significantly the organization of terms according to their semantic relations. Notice that this behavior supports that the kernel version proposed improves the Voronoi regions and the X network organization by transforming nonlinearly the dissimilarity considered.

5 Conclusions and future research trends

In this paper we have introduced a new asymmetric self organizing map (ASOM) suitable to visualize hierarchical relationships in an intuitive way. Next a kernel version has been developed to improve the network organization by nonlinearly transforming the original dissimilarity. The algorithms have been evaluated in a challenging problem such as thesaurus generation through several objective error functions.

We conclude that the ASOM generate successfully visual representations of the term hierarchy. In particular, broad terms are almost always positioned above specific terms in the hierarchy. On the other hand, term relations induced by each level are as good as for the classic SOM algorithm. The kernel version

outperforms the linear version achieving smoother transitions between different levels of the hierarchy and improving significantly the term relations induced by each level of the hierarchy.
Future research, will focus on the study of new dissimilarities that allows to reproduce the hierarchical organization of the objects.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, UK, 1999.

[2] M. W. Berry and Z. Drmac and E. R. Jessup. Matrices, Vector Spaces and Information Retrieval. SIAM review. 41(2), 335-362, 1999.

[3] J. C. Bezdek and N. R. Pal . An Index of Topological Preservation for Feature Extraction, Pattern Recognition, 28(3):381-391, 1995.

[4] A. Buja and B. F. Logan and J. A. Reeds and L. A. Shepp. Inequalities and Positive-Definite Functions Arising from a Problem in Multidimensional Scaling. Annals of Statistics, 22(1), 406-438, 1994.

[5] V. Cherkassky and F. Mulier. Learning from Data. Wiley, New York, 1998.

[6] T. Kohonen. Self-Organizing Maps. Second Edition, Springer Verlag, 1997.

[7] D. Lawrie, W. B. Croft and A. Rosenberg. Finding Topic Words for Hierarchical Summarization, SIGIR'01, New Orleans, ACM, 349-357, 2001.

[8] M. Martin-Merino and A. Muñoz. Self Organizing Map and Sammon Mapping for Asymmetric Proximities, ICANN, LNCS 2130, 429-435, Springer Verlag, 2001.

[9] M. Martin-Merino and A. Muñoz. A New Asymmetric Topographic Mapping Algorithm. Proc. of the International Conference on Knowledge-Based Intelligent Information. 873-878, IOS Press, 2002.

[10] D. Merkl, Text Classification with Self-organizing Maps: Some Lessons Learned, Neurocomputing, 21, 61-77, 1998.

[11] A. Muñoz. Compound key word generation from document databases using a hierarchical clustering ART model. Journal of Intelligent Data Analysis, 1(1), 1997.

[12] B. Schölkopf and A. Smola. Learning with Kernels, MIT Press, Cambridge, 2002.