SELF ORGANIZING MAPS FOR REDUCING THE NUMBER OF CLUSTERS BY ONE ON SIMPLEX SUBSPACES

Constantine Kotropoulos and Vassiliki Moschou

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {costas, vmoshou}@aiia.csd.auth.gr

ABSTRACT

This paper deals with N-dimensional patterns that are represented as points on the $(N-1)$-dimensional simplex. The elements of such patterns could be the posterior class probabilities for N classes, given a feature vector derived by the Bayes classifier for example. Such patterns form N clusters on the $(N-1)$-dimensional simplex. We are interested in reducing the number of clusters to $N-1$ in order to redistribute the features assigned to a particular class in the $N-1$ simplex over the remaining $N-1$ classes in an optimal manner by using a self-organizing map. An application of the proposed solution to the re-assignment of emotional speech features classified as neutral into the emotional states of anger, happiness, surprise, and sadness on the Danish Emotional Speech database is presented.

1. INTRODUCTION

Self organizing maps (SOMs) establish a mapping from an input data space $\mathbb{R}^N$ onto a two-dimensional array of nodes, which are associated with a weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_N)^T$ [1]. The nodes (i.e., neurons) are organized on a map and they compete in order to be activated by the input patterns. The SOM is one of the most popular neural networks. A number of 5384 papers related to the SOM are listed in [2, 3] and cited at `www.cis.hut.fi/research/som-bibl/`. Recent applications of the SOM include the classification of human body postures from images [4], the grouping and visualizing human endogenous retroviruses [5], speaker clustering [6], to mention a few.

This paper deals with N-dimensional patterns that are represented as points on the $(N-1)$-dimensional simplex. The elements of such patterns could be the posterior class probabilities for N classes given a feature vector derived by the Bayes classifier for example. Such patterns form clusters on the $(N-1)$-dimensional simplex. We are interested in reducing the number of clusters to $N-1$ in order to redistribute the features assigned into a particular class in the $N-1$ simplex according to the maximum a posteriori probability principle, over the remaining $N-1$ classes in an optimal manner by using the SOM. The motivations for the work reported in this paper are the following: 1) There are facial expression databases such as Action-Unit coded Cohn-Kanade database [7] where the neutral emotional class is not represented adequately. Therefore, facial expression recognition experiments may not report classification rates for the neutral emotional class as in [8]. For the emotional speech databases, utterances are regularly classified as neutral. Accordingly, when the neutral class is not represented in one modality, it is difficult to develop multimodal emotion recognition algorithms (e.g., decision fusion algorithms). 2) Frequently, the ground-truth information related to emotions that is provided by the human evaluators is biased towards the neutral class. Therefore, the patterns classified to this class might be redistributed among the non-neutral classes. 3) In general, although there are criteria that could be used for selecting the number of clusters, such as the Akaike Information Criterion (also known as Bayesian Information Criterion), the number of clusters is frequently selected arbitrarily. Furthermore, the number of clusters can be either increased as in divisive algorithms or decreased as in agglomerative algorithms. 4) Moreover, when two clustering algorithms are to be compared (e.g. cluster validity), they should refer to the same number of clusters. In this paper, we are interested in reducing the number of clusters when the patterns belong to simplex spaces.

The novelty of this paper is in the mathematical derivation of the training algorithm for a SOM that reduces the number of clusters by one on a simplex subspace. An application of the proposed solution to the re-assignment of emotional speech features classified as neutral into the emotional states of anger, happiness, surprise, and sadness on the Danish Emotional Speech database is presented.

The rest of the paper is organized as follows. Section 2 describes the mathematical derivation of the training algorithm. It is split into two subsections; Subsection 2.1 studies the constrained optimization problem for a single neuron and the generalization to a map is made in subsection 2.2. Section 3 demonstrates an application of the developed theory to emotional speech classification. Finally, conclusions are drawn in Section 4.

2. SOM TRAINING ALGORITHM FOR REDUCING THE NUMBER OF CLUSTERS BY ONE ON SIMPLEX SUBSPACES

2.1. Single neuron case

Let $\mathbf{w} \in \mathbb{R}^N$ be the weight vector of a single neuron. This neuron could be the best matching neuron to input pattern $\mathbf{x}$. For the sake of simplicity let us treat this neuron as an adaptive filter that processes a random input vector $\mathbf{x}$ whose elements are the posterior class probabilities derived by a classifier (i.e. the Bayes classifier) given a feature vector $\mathbf{z}$

$$\mathbf{x} = (P(\omega_1|\mathbf{z}), P(\omega_2|\mathbf{z}), \ldots, P(\omega_N|\mathbf{z}))^T \tag{1}$$

where T is the transposition operator and $\omega_i, i = 1, 2, \ldots, N$ denote the classes. Since the elements of $\mathbf{x}$ are probabilities they satisfy the

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and LEarning" (FP6-507752).

properties

$$x_1 \geq 0, x_2 \geq 0, \ldots, x_N \geq 0 \tag{2}$$

$$\mathbf{1}^T \mathbf{x} = 1 \tag{3}$$

where $\mathbf{1}$ is the $N \times 1$ vector of ones. Let us define the operator $\tilde{}$ that is applied to any vector $\mathbf{z} \in \mathbb{R}^N$ and discards its Nth element, i.e.

$$\mathbf{z} = (z_1, z_2, \ldots, z_{N-1} | z_N)^T = \left(\widetilde{\mathbf{z}}^T | z_N\right)^T. \tag{4}$$

We would like to update the weight vector $\mathbf{w}$ so that it minimizes the mean-squared error between $\mathbf{x}$ and $\mathbf{w}$

$$J = \mathrm{E}\left\{||\mathbf{x} - \mathbf{w}||^2\right\} \tag{5}$$

subject to the constraints

$$w_1 \geq 0, w_2 \geq 0, \ldots, w_{N-1} \geq 0 \tag{6}$$

$$\widetilde{\mathbf{1}}^T \, \widetilde{\mathbf{w}} = 1. \tag{7}$$

The optimization problem (5)-(7) is interpreted as follows. We require the reduced weight vector $\widetilde{\mathbf{w}}$ to lie on the $(N-2)$-dimensional simplex, so that its elements could be treated as posterior class probabilities. The input vectors $\mathbf{x}$ lie on an $(N-1)$ simplex and create a cluster on this simplex. The optimization problem (5)-(7) enforces the random input vectors $\mathbf{x}$ to create a single cluster on the $(N-2)$-dimensional simplex whose parametric reference vector is $\widetilde{\mathbf{w}}$. The aforementioned problem can be solved by applying the steepest-descent algorithm and the theory of constrained optimization [9].

By dropping the expectation operator and using the instantaneous squared error, the unconstrained minimization of (5) yields a Least-Mean-Square (LMS) adaptation of $\mathbf{w}$ [1, 10, 11]

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \beta\left[\mathbf{x}(n) - \mathbf{w}(n)\right] \tag{8}$$

where n denotes discrete time (i.e., iteration) and β is the adaptation step-size. For the time-being the adaptation step-size is considered to be constant.

Next, we elaborate on the constrained optimization problem (5)-(7). The sum of the N elements of $\mathbf{w}$, taking into account (7), becomes

$$\mathbf{1}^T \, \mathbf{w}(n+1) = \left(\widetilde{\mathbf{1}}^T | 1\right) \begin{bmatrix} \widetilde{\mathbf{w}}(n+1) \\ w_N(n+1) \end{bmatrix} = 1 + w_N(n+1) \tag{9}$$

where $w_N(n+1)$ is the last element of $\mathbf{w}$ in the $(n+1)$th iteration. By replacing (8) into (9) we obtain

$$\mathbf{1}^T \mathbf{w}(n+1) \quad = \quad 1 + w_N(n+1) = \delta(n) \tag{10}$$

$$w_N(n+1) \quad = \quad (1-\beta) w_N(n) = \gamma(n) \tag{11}$$

with $\delta(n) = 1 + \gamma(n)$. In the following, we drop the dependence on n for notation simplicity. To minimize the instantaneous squared error (i.e., the objective function) subject to the equality constraint (10) and the $N-1$ inequality constraints (6) we introduce the Lagrangian function

$$\mathcal{L}(\mathbf{w}, \lambda, \widetilde{\boldsymbol{\xi}}) = ||\mathbf{x} - \mathbf{w}||^2 - \lambda\left(\mathbf{1}^T \mathbf{w} - \delta\right) - \widetilde{\boldsymbol{\xi}}^T \widetilde{\mathbf{w}} \tag{12}$$

where λ and $\widetilde{\boldsymbol{\xi}}$ are the Lagrange multipliers. By equating the gradient of the Lagrangian function with respect to $\mathbf{w}$ with the zero vector we obtain

$$\mathbf{w} = \mathbf{x} + \frac{\lambda}{2}\mathbf{1} + \frac{1}{2}[\frac{\widetilde{\boldsymbol{\xi}}}{0}]. \tag{13}$$

The substitution of (13) into (10) yields

$$\lambda = \frac{2}{N}\left[(\delta - 1) - \frac{1}{2}\widetilde{\mathbf{1}}^T \widetilde{\boldsymbol{\xi}}\right]. \tag{14}$$

Let $\mathbf{J} = \widetilde{\mathbf{1}}\widetilde{\mathbf{1}}^T$, $\mathbf{I}$ be the $(N-1) \times (N-1)$ identity matrix, and $\boldsymbol{\Theta} = \mathbf{J} - N\mathbf{I}$. Putting (13) into the Lagrangian (12) and taking into account the Kuhn-Tucker (KT) conditions [9], we obtain the Wolf dual functional

$$\mathcal{W}(\widetilde{\boldsymbol{\xi}}) = \frac{(\delta-1)^2}{N} + \frac{1}{4N}\widetilde{\boldsymbol{\xi}}^T \boldsymbol{\Theta} \widetilde{\boldsymbol{\xi}} - \widetilde{\boldsymbol{\xi}}^T \left(\widetilde{\mathbf{x}} + \frac{\delta - 1}{N}\widetilde{\mathbf{1}}\right) \tag{15}$$

that should be maximized in the nonnegative quadrant of ξ_i, $i = 1, 2, \ldots, N-1$. The solution of the aforementioned maximization problem is given by

$$\widetilde{\boldsymbol{\xi}} = 2N\boldsymbol{\Theta}^{-1}\left(\widetilde{\mathbf{x}} + \frac{\delta - 1}{N}\widetilde{\mathbf{1}}\right). \tag{16}$$

2.2. Generalization for a map

Let $d(\mathbf{x}(n), \mathbf{w}_l(n))$ denote a generic distance measure between an input pattern $\mathbf{x}(n)$ and a neuron of the map $\mathbf{w}_l(n)$. In this paper, $d(\cdot)$ is the Euclidean distance. The index of the best matching neuron to input pattern $\mathbf{x}(n)$ is given by

$$c(\mathbf{x}(n)) = \arg\min_l \{d(\mathbf{x}(n), \mathbf{w}_l(n))\}, \qquad l = 1, 2, \ldots, M \tag{17}$$

where M is the number of neurons. Let the adaptation step-size and the kernel function used in the map be merged in the term $\widetilde{h}_{cl}(n)$ [11]. When an input pattern is presented to the map at the nth iteration, the following computations take place at the lth neuron of the map:

Step 1: Determine $\delta_l(n)$:

$$\delta_l(n) = 1 + (1 - \widetilde{h}_{cl}(n)) w_{lN}(n). \tag{18}$$

Step 2: Determine the Lagrange multipliers for the inequalities $\widetilde{\boldsymbol{\xi}}\,(n)$:

$$\widetilde{\boldsymbol{\xi}}_l(n) = 2N\boldsymbol{\Theta}^{-1}\left(\widetilde{\mathbf{x}}(n) + \frac{\delta_l(n) - 1}{N}\widetilde{\mathbf{1}}\right). \tag{19}$$

Step 3: Determine the Lagrange multiplier for the equality constraint $\lambda(n)$:

$$\lambda_l(n) = \frac{2}{N}\left[(\delta_l(n) - 1) - \frac{1}{2}\widetilde{\mathbf{1}}^T \widetilde{\boldsymbol{\xi}}_l(n)\right]. \tag{20}$$

Step 4: Update the weight vector $\mathbf{w}_l$:

$$\mathbf{w}_l(n+1) = \mathbf{x}(n) + \frac{\lambda_l(n)}{2}\mathbf{1} + \frac{1}{2}[\frac{\widetilde{\boldsymbol{\xi}}_l(n)}{0}]. \tag{21}$$

3. EXPERIMENTAL RESULTS

The Danish Emotional Speech (DES) database [12] was used in order to demonstrate the proposed training algorithm for a SOM that is able to reduce the number of clusters by one on a simplex subspace. In particular, a SOM variant was trained to re-assign emotional speech features classified as neutral into four emotional states, namely hot anger, happiness, surprise, and sadness indicated by the

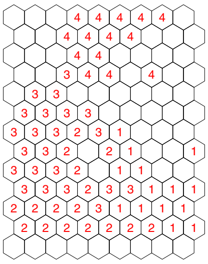

Fig. 1. Class label assignment of the map neurons.

labels 1–4, respectively. The training algorithm was integrated with the SOM Toolbox [14].

The input patterns $\mathbf{x}$ are posterior class probabilities of emotional speech features for 5 emotional states, including the neutral class. They were derived by applying the Bayes classifier that employs features optimally selected by the sequential floating forward selection algorithm so that they minimize the classification error [13]. The training set consists of patterns that are not classified as neutral. However, the patterns do include the Nth element $x_N = P(\omega_N|\mathbf{z})$ that is the a posteriori probability the feature vector $\mathbf{z}$ to be classified as neutral determined by the Bayes classifier. One pattern is fed to the SOM at each iteration. The map topology is depicted in Figure 1 where the class labels assigned to each neuron is also indicated. The unlabeled neurons are not assigned enough patterns.

A number of 400000 iterations were needed for the training algorithm to converge. To assess convergence, two measures were employed, namely

$$J(l,n) = \frac{1}{n}\sum_{n'=1}^{n}\left(\tilde{\mathbf{1}}^T\tilde{\mathbf{w}}_l(n') - 1\right)^2 \tag{22}$$

$$J'(n) = \frac{1}{M}\sum_{l=1}^{M}\left(\tilde{\mathbf{1}}^T\tilde{\mathbf{w}}_l(n) - 1\right)^2 \tag{23}$$

where $M = 13 \times 9 = 117$ is the number of map neurons. $J(l,n)$ is plotted for several iteration numbers n=1, 100000, 200000, 300000, and 400000 in Figure 2. It is seen that $J(l,n)$ admits values of the order of 10^{-8} after 200000 iterations. The peaks in the plots of Figure 2 show a periodic behavior. Their position corresponds to the boundary neurons of the map. This fact can be explained by taking into account the scanning order for computing the kernel function between any neuron and the best matching neuron for an input pattern [1] during training.

Figure 3 demonstrates that $J'(n)$ converges to a steady-state value of the order 10^{-3}. It should be noted that the generalization to a map employs a variable equality constraint $\delta(n)$ and not a constant δ as in the simple case of a single neuron. The plot in Figure 3 plays the role of the learning curve of the training algorithm.

The neurons can be characterized as good, modest, and bad, depending on the number of patterns they "win" during the competition. Let ν be the number of patterns a neuron wins. Let $\eta = \mathrm{E}\{\nu\}$ be the mean and $\sigma = \sqrt{\mathrm{var}\{\nu\}}$ be the standard deviation of the random variable ν for a neuron. The neuron under discussion is characterized as good if it wins at least $\eta + \sigma$ patterns and bad if it wins less than $\eta - \sigma$ patterns. Otherwise, it is characterized as modest. Figure 4 depicts the category of each neuron. The red color indicates the good neurons, the green color indicates the modest neurons, and blue the bad ones. From this perspective, as can be noticed from the top row of the map in Figure 4, 4 boundary neurons of the map are bad, 3 are found to be modest, and only 2 are categorized as good.

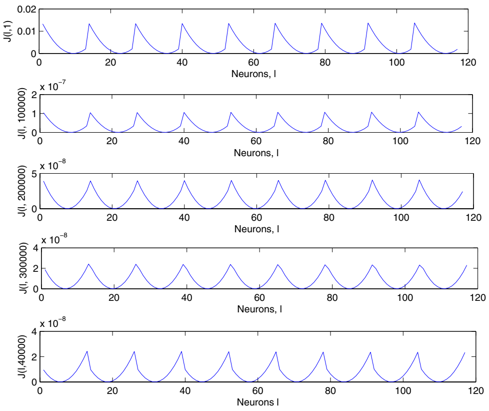

Fig. 2. $J(l,n)$ for several iteration numbers n versus the neuron index l.

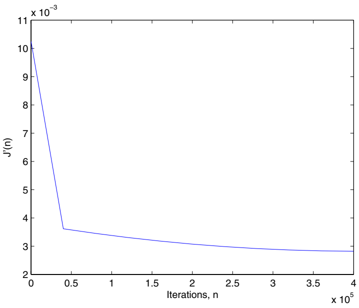

Fig. 3. $J'(n)$ versus the iteration index n collectively for the map.

Figure 5 plots the ensemble-averaged values of the weights for a good neuron [10]. It is seen that each weight converges to a steady state value and the sum of the 4 weights upon convergence equals 1. A very few number of weights, after the training is complete, are found to be negative. As can be seen from Table 1, this happens mostly for bad neurons. The situation is better for modest neurons, and there is not any problem for the good neurons. Table 1 suggests that we should use the sign of the Lagrange multipliers $\tilde{\boldsymbol{\xi}}_l(n)$ as a control mechanism to avoid negative weights during the optimization.

Table 1. Percentage of negative weights found for each neuron category

Neuron category	Percentage (%)
Good	**0.17**
Modest	0.51
Bad	3.93

In the test phase, the trained SOM variant was applied to unseen patterns that were originally classified as neutral. Figure 6 shows the percentage of neutral patterns that are re-assigned to each non-

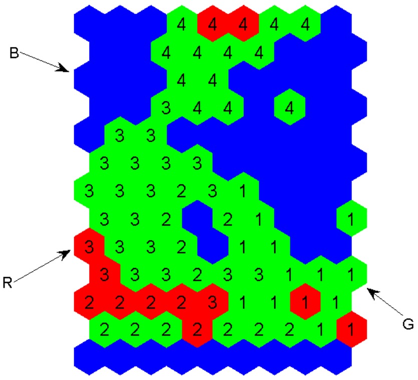

Fig. 4. Characterization of map neurons as good, modest,and bad according to the number of patterns they win. (R: red, B: blue, G: green)

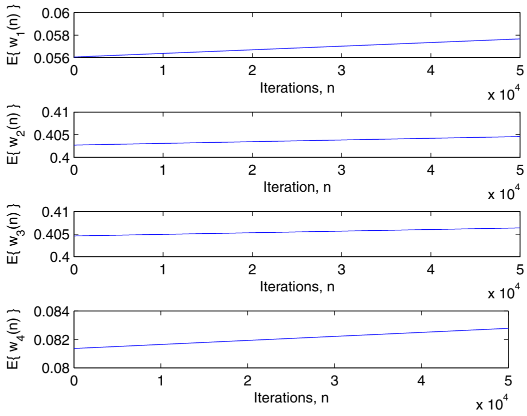

Fig. 5. Ensemble-averaged values of the weights for a good neuron.

neutral emotional class by the trained SOM variant. As can be seen, surprise earns the majority of the neutral patterns, while anger is found to be the least close to the neutral state.

4. CONCLUSIONS

A SOM that is able to reduce the number of clusters by one on simplex subspaces has been proposed. The training algorithm for such a SOM has been derived theoretically. The theoretical developments were successfully applied to re-assign emotional speech features originally classified as neutral to four non-neutral emotional classes in the DES database. The convergence properties of the developed SOM variant have been demonstrated by experiments. The experimental results have validated the theoretical developments.

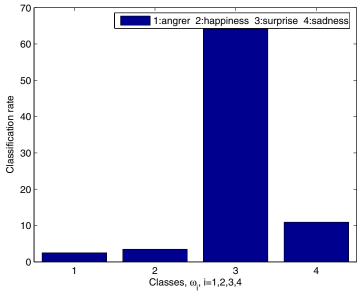

Fig. 6. Percentage of neutral patterns that each non-neutral emotional class wins.

5. REFERENCES

[1] T. Kohonen, *Self-Organizating Maps*, 3/e. Berlin, Germany: Springer-Verlag, 2000.

[2] S. Kaski, J. Kangas, and T. Kohonen, “Bibliography of Self-Organizing Map (SOM) Papers: 1981-1997,” *Neural Computing Surveys*, vol. 1, pp. 102-350, 1998.

[3] M. Oja, S. Kaski, and T. Kohonen, “Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum,” *Neural Computing Surveys*, vol. 3, pp. 1-156, 2003.

[4] K. Takahashi and S. Sugakawa, “Remarks on human posture classification using self-organizing map,” in Proc. *IEEE Int. Conf. Systems, Man, and Cybernetics*, vol. 3, pp. 2623-2628, October 2004.

[5] M. Oja, G. Sperber, J. Blomberg and S. Kaski, “Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps”, in Proc. *2004 IEEE Symp.Computational Intelligence in Bioinformatics and Computational Biology*, pp. 95-101, 2004.

[6] I. Lapidot, H. Guterman, and A. Cohen, “Unsurpervised speaker recognition based on competition between self-organizing maps,” *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 877-887, July 2002.

[7] J. C. T. Kanade and Y. Tian, “Comprehensive database for facial expression analysis,” in *Proc. IEEE Int. Conf. Face and Gesture Recognition*, pp. 46-53, March 2000.

[8] I. Kotsia and I. Pitas, “Real-time facial expression recognition from image sequences using support vector machines,” in *Proc. Conf. Visual Communications Image Processing*, Beijing, China, July 12-15, 2005.

[9] R. Fletcher, *Practical Methods of Optimization*, 2/e. London: John Wiley & Sons, 1987.

[10] S. Haykin, *Adaptive Filter Theory*, 4/e. Englewood Cliffs, N.J.: Prentice-Hall, 2001.

[11] C. Kotropoulos and I. Pitas,“Self-organizing maps and their applications in image processing, information organization, and retrieval,” in *Nonlinear Signal and Image Processing: Theory, Methods, and Applications* (K. E. Barner and G. R. Arce, Eds.), Boca Raton, FL: CRC Press, 2004.

[12] I. S. Engberg and A. V. Hansen, “Documentation of the Danish Emotional Speech Database (DES),” Internal report, Center for Person Kommunikation, Aalborg University, 1996.

[13] D. Ververidis and C. Kotropoulos,“Sequential forward feature selection with low computational cost,” in *Proc. XIII European Signal Processing Conf.*, Antalya, Turkey, September 2005.

[14] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, *SOM Toolbox for Matlab 5*, Finland, 2000, `www.cis.hut.fi`.