6th. International Work Conference on Artificial and Natural Neural Networks
(IWANN2001), Granada, June 13-15 2001

Comparison of supervised self-organizing maps using Euclidian or Mahalanobis distance in classification context

F. Fessant[1], P. Aknin[1], L. Oukhellou[2], S. Midenet[1],

[1] INRETS, 2 av. Général Malleret Joinville,
F-94114 Arcueil, France
{fessant, aknin, midenet}@inrets.fr
[2] LETIEF, Université Paris 12, 61 av. Général de Gaulle,
F-94010 Créteil, France
oukhellou@univ-paris12.fr

Abstract. The supervised self-organizing map consists in associating output vectors to input vectors through a map, after self-organizing it on the basis of both input and desired output given altogether. This paper compares the use of Euclidian distance and Mahalanobis distance for this model. The distance comparison is made on a data classification application with either global approach or partitioning approach. The Mahalanobis distance in conjunction with the partitioning approach leads to interesting classification results.

1 Introduction

The self-organizing map -or SOM- is a well-known and quite widely used model that belongs to the unsupervised neural network category concerned with classification processes. The LASSO model (Learning ASsociations by Self-Organization) used in this work, can be considered as an extension of self-organizing maps and allows the classification process in a supervised way [1]. The LASSO model had been tested on pattern recognition tasks [1,2] and it has been shown that the encoding and use of supervision data during the learning phase improve the well-classification rate compared to the standard SOM results.

This paper focuses on the metric choice for the prototype-to-observation distance estimation required during the self-organization and exploitation phases. The distance most commonly used in SOM is the Euclidian distance that considers each observation dimension with the same significance whatever the observation distribution inside classes.

Obviously, if the data set variances are not uniformly shared out among the input dimensions, the use of Mahalanobis distance becomes interesting.

After the presentation of supervised SOM and metrics, the article introduces two different classification strategies : a global approach and a partitioning approach. We give classification performances for the two metrics on a data classification application concerning non-destructive evaluation of rail. We show that the choice of a Mahalanobis metric can clearly improve the classification results, specially in the context of partitioning approach.

2 The supervised self-organizing map

2.1 Standard SOM

A SOM is a neural network model made up of a set of prototypes (or nodes) organized into a 2-dimension grid. Each prototype *j* has fixed coordinates in the map and adaptive coordinates *W(j)* (or weights) in the input space. The input space is relative to the variables setting up the observations. Two distances are defined, one in the original input space and one in the 2-dimension space of the map for which Euclidian distance and integer map coordinates are always used. The definition choice of first distance will be discussed in section 3. The self-organizing process slightly moves the prototype coordinates in the data definition space -i.e. adjusts *W*- according to the data distribution. The chosen learning algorithm uses a learning rate and a smooth neighbouring function continuously decreasing with iterations [3].

The SOM is traditionally used for classification purpose. The exploitation phase consists in associating an observation with its closest prototype, called the image-node. The SOMs have the well known ability [4] that the observation projection on the map preserves the proximities : close observations in the original input space are associated with close prototypes in the map.

2.2 LASSO model

Midenet and Grumbach [1] proposed the exploitation of SOM in a supervised way through the LASSO model. As SOM, the LASSO model associates a 2-dimension map with a communication layer. However, 2 sets of nodes are distinguished on the communication layer : observation nodes and class coding nodes. During the learning phase, the whole communication layer (observation nodes and class nodes) is used as input layer (fig. 1). This phase is done as in the classical Kohonen model. The LASSO map self-organizes with observations and associated classes being presented altogether at the network input.

On the other hand, during the exploitation phase, only the observation nodes are used as

inputs. The map is used to associate missing data (class estimation) with partial information (observation) in the following way :

1. presentation of an observation *X* on the input layer
2. selection of the image-node by minimizing the distance between observation and prototypes. The class coding dimensions are simply ignored during the selection
3. the estimation of the class is computed from the weights of the image-node in the class coding dimensions (fig. 1)

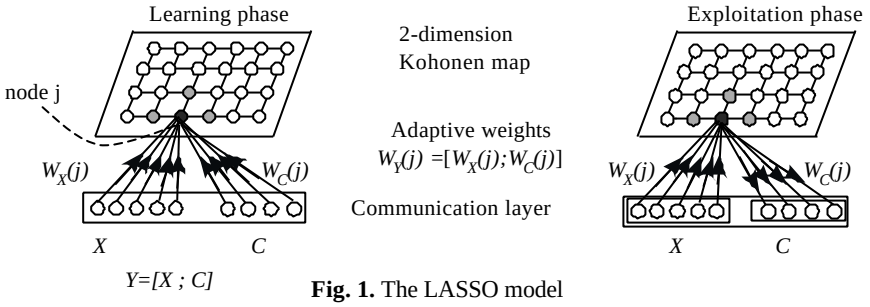

Fig. 1. The LASSO model

3 Metric choice

The original contribution of this paper lies in the use of a non-euclidian metric in learning phase as well as exploitation phase. The distance d_N in a N-dimension space between 2 patterns *R* and *S* is given in its generalized form with the following expression :

$$d_N^2(R,S) = (R-S)^T \mathbf{A}^{-1}(R-S) \ . \qquad (1)$$

with **A** being a N by N positive normalisation matrix. If **A** is the identity matrix, the distance is Euclidian. If **A** is the variance-covariance matrix, this distance is refereed as the Mahalanobis distance. In a multi-classes problem, the usual choice of **A** is given by the average of inner class variance matrices [5].

$$\mathbf{A} = \sum_{i=1}^{K} n_{\Omega_i} \mathbf{V}_{\Omega_i} \ . \qquad (2)$$

with $n_{\Omega i}$ the observation number in class Ω_i, $\mathbf{V}_{\Omega i}$ the variance-covariance matrix of class Ω_i and K the class number.

3.1 Use of Mahalanobis distance in LASSO learning phase

The distance d_N in the N-dimension input space between a complete sample *Y* (observation *X* + associated class *C*) and a prototype *j* is given by :

$$d_N^2(Y, j) = (Y - W_Y(j))^T \mathbf{A}_Y^{-1} (Y - W_Y(j)) \ . \qquad (3)$$

$\mathbf{A}_Y$ (dimension N x N) being estimated from learning set with $Y=[X\,;\,C]$. C has N_C dimensions. According to equation 2, $\mathbf{A}_Y$ can be written as :

$$\mathbf{A}_Y = \begin{bmatrix} [\mathbf{A}_X] & 0 \cdots & 0 \\ \ddots & & \ddots \\ 0 & \cdots\ 0 \cdots & 0 \end{bmatrix} . \qquad (4)$$

The rank of $\mathbf{A}_Y$ is the dimension of X (i.e. N-N_C). The components corresponding to class information in $\mathbf{A}_Y$ are equal to zero. However, it is important to point out that the class information is implicitly taken into account in $\mathbf{A}_X$ and therefore in d_N . Otherwise the class coding is explicitly used in the weight updating equation.

3.2 Use of Mahalanobis distance in LASSO exploitation phase

During the exploitation phase, the distance is only computed in the observation dimensions (N-N_C dimensions). The class coding dimensions are ignored.

$$d^2_{\text{N-Nc}}(X, j) = (X - W_X(j))^T \mathbf{A}_X{}^{-1} (X - W_X(j)) \ . \qquad (5)$$

Despite the dimension reduction, the part of $\mathbf{A}$ that is used in equation 5 is the same as in learning phase. Distances used in learning and in exploitation phases are consistent.

4 The classification task

4.1 Global classification approach

A single LASSO model is designed for the simultaneous discrimination of the whole set of classes. Inducing by a disjunctive coding, there are as many class nodes as classes to be discriminated. Each output delivers a level that can be considered as an estimation of class belonging probability of the observation (fig. 2 presents the architecture used to solve a global K=4 classes problem).

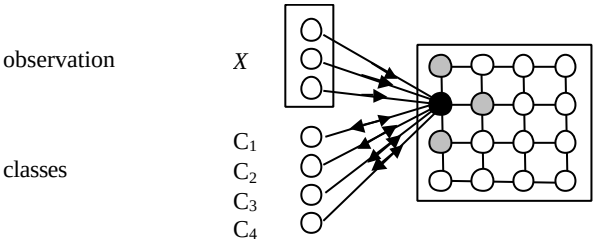

Fig. 2. A single LASSO model for the simultaneous discrimination of 4 classes

The decision rule for the global classifier is simple : during the exploitation phase, the

assigned class corresponds to the class node component with the highest weight level.

4.2 Partitioning classification approach

The classification task is now split into several elementary tasks : K sub-problems are generated ; each of them is dedicated to the separation of one class among the others. Each sub-problem is solved with a specific sub-LASSO model. Fig. 3 presents the architecture used to solve a 4 classes classification problem, with 4 sub-networks which deliver some kind of estimation of posterior probability class membership. The decision rule for the final classification corresponds to the maximum posterior probability estimation.

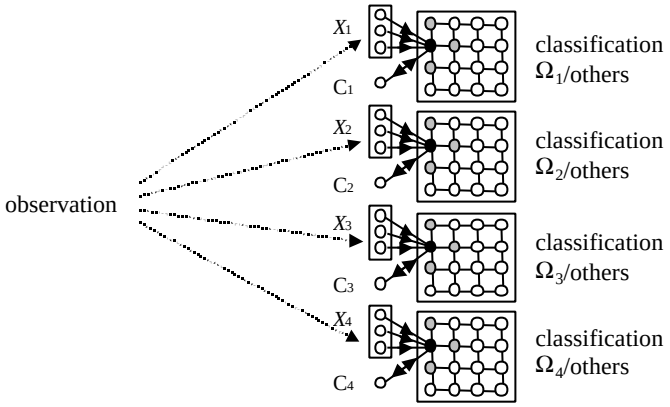

Fig. 3. Complete architecture for a 1 by 4 partitioning approach

The partitioning approach allows us to independently design each sub-network and the input vectors can be adjusted in order to be more relevant for the separability of each particular class ; sub-network input dimension can differ one from another. Moreover, the learning phase of each sub-classifier is independent, which gives us a better control than a global learning.

5 Rail defect classification

5.1 Application description

The application concerns the classification of rail defect signatures measured by a specific eddy current sensor [6]. After a raw signal parametrisation and a reduction of the number of descriptors [7], a classification procedure is carried out in order to assign each defect to one of the 4 defined classes of defects or singularities : external crack, welded joint, joint before crossing, shelling. A database of 140 defects is available for the design and the evaluation of the classifiers (64 from class 1, 39 from class 2, 16 from class 3 and 21 from class 4).

5.2 Practical settings

For the global classification approach, a square map has been used. Its size has been experimentally determined by testing a whole range of values between 3 by 3 and 10 by 10. The best size was found to be 8 by 8. We adopted the disjunctive coding scheme for the class information : all the components of the 4-dimension class vectors are worth zero except the one associated with the represented class. This component is set to one. The observation variables are normalised. The communication layer has the following characteristics : observation dimension=22, class dimension=4.

For the partitioning classification approach, 4 square maps have been implemented, one for each sub-classifier. Their sizes have been determined for each of them by testing a range of values between 3 by 3 and 10 by 10. The best sizes were found to be respectively {7 by 7, 5 by 5, 5 by 5, 5 by 5}. These sizes correspond to the best well-classified rate of each sub-problem. The single class node of each sub-classifier is set to one or zero depending on the observation class. The observation variables are normalised. The communication layers have the following characteristics : observation dimensions={15,15,8,9}, class dimension=1.

5.3 Global classification results

Fig. 4 depicts the four class nodes weight maps for the LASSO with Mahalanobis metric, after self-organization (observation nodes weight maps are not represented). On the figure, a weight value is represented by a dot in gray scale : dark dot means low value whereas light dot means high value.

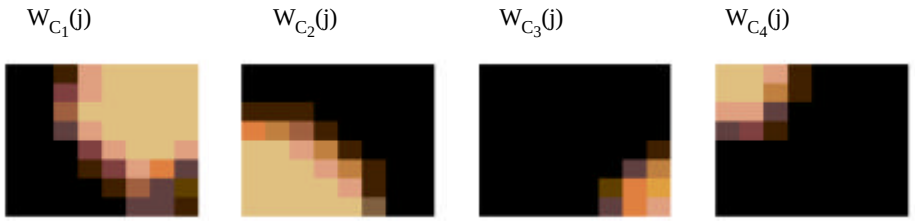

Fig. 4. Class node weight maps of one global classifier with Mahalanobis metric

The prototypes that specialize in one particular class recognition cluster in the map into a related neighbourhood. The areas are compact and do not overlap. The area associated to one class is more or less important in the map according to its occurrence frequency in the learning set ; we can observe that defects of classes 3 and 4 are treated with very few prototypes. This is a consequence of their low representation in the learning set. The class node map configurations reflect some similarity between classes ; for instance class 1 (crack) and 3 (joint before crossing) are lightly close. On the contrary, class 3 and class 4 are far away, indeed they stand in the opposite corners of the map.

The classification performances are summarized table 1. Because of the small number of observations, a “leave-one-out procedure” was applied for the classifiers evaluation [5] :

one observation is removed from the complete data base ; then the remaining observations are used for training and the selected observation is used for testing. The procedure is repeated until all the observations of the initial base are listed. The classification rates have been obtained from the average of 20 results given by 20 complete leave-one-out procedures with different initial weights. Standard deviations σ are also pointed out table 1.

Table 1. Global classification performances (in percentage)

	Euclidian metric	Mahalanobis metric
global classifier	**92.8%** (σ=1.6)	**94.1%** (σ=1.6)

The use of the Mahalanobis metric lightly improves the classification results. We can assume that this benefit is due to the processing operated by means of the variance-covariance matrix, which enables to take into account the variances and correlation among components.

5.4 Partitioning classification results

Table 2 summarizes the classifications rates for each sub-classifier studied independently, as well as the classification performance for the complete classifier.

Table 2. Classification performances for partitioning approach (in percentage)

	Euclidian metric	Mahalanobis metric
Class 1/others	89.7%	93.3%
Class 2/others	96.4%	97.9%
Class 3/others	93.7%	94.5%
Class 4/others	98.2%	97.8%
Complete classifier	**90%** (σ=1.5)	**95%** (σ=1)

A partitioning approach in conjunction with the Mahalanobis metric leads to improved classification performances. We can notice that some sub-problems remain rather difficult to solve, like the discrimination of class 1 and 3. The benefit of the Mahalanobis metric is about 5% in partitioning context. This result is quite worthwhile compared to the 1.3% in global context.

6 Conclusion

This paper proposed the comparison of supervised self-organizing maps designed with different distance measures : an Euclidian distance and a Mahalanobis one. This comparison is achieved on a classification task, a particular non destructive evaluation problem involving the discrimination of four classes of defects. Classification performances are given for two classification strategies : a global approach in which the whole classes are

discriminated simultaneously and a partitioning one in which the multiclass problem is split into 2-class sub-problems.

Concerning our classification problem, the Mahalanobis distance turns out to be more effective because of the large range of data components variations. In fact, the giving up of Euclidian distance is advisable when the variance of input vector components are highly different.

The best classification results are obtained for a LASSO model using a Mahalanobis metric in conjunction with the partitioning approach. This approach is more efficient than the global one essentially for the reason that the preprocessings and network architecture can be better adapted to each sub-problem.

These results compare with those achieved on the same database with supervised neural networks like multilayer perceptrons or radial basis functions [8].

Acknowledgments. This research is supported by the French Research Ministry within the framework of the PREDIT program (Research program and Innovation in Transportation Systems). The RATP company coordinates this project.

References

1. Midenet, S., Grumbach, A.: Learning Associations by Self-Organization : the LASSO model. Neurocomputing, vol. 6. Elsevier, (1994) 343-361
2. Fessant, F., Aknin, P.: Classification de défauts de rail à l'aide de cartes auto-organisatrices supervisées. NSI'2000, Dinard (2000) 75-79
3. Ritter, H., Martinetz, T., Schulten, K.: Topology conserving maps for learning visuo motor coordination. Neural networks, vol. 2. (1989) 159-168
4. Kohonen, T.: Self-Organizing Maps. Springer, Heidelberg (1995)
5. Bishop, C. M.: Neural Networks for pattern recognition. Clarendon Press, Oxford (1997)
6. Oukhellou, L., Aknin, P., Perrin, J.P.: Dedicated sensor and classifier of rail head defects for railway systems. 8th IFAC Int. Symp. On Transportation Systems, Chania (1997)
7. Oukhellou, L., Aknin, P., Stoppiglia, H., Dreyfus, G.: A new decision criterion for feature selection. Application to the classification of non destructive testing signatures. EUSIPCO, vol. 1 Rhodes (1998) 411-414
8. Oukhellou, L., Aknin, P.: Hybrid training of radial basis function networks in a partitioning context of classification. Neurocomputing, vol. 28 (1999) 165-175