FEATURE SPACE PARAMETRIZATION BY MEANS SELF-ORGANIZING MAPS AND BOX-COX TRANSFORMATIONS

Platero,C., Crespo, C., García, J., Campoy, P., Aracil, R.
División de Automática (DISAM - UPM)
c/José Gutiérrez Abascal, 2, 28006 Madrid (España)
cplatero@disam.upm.es

Abstract

This paper shows a new method for implement hybrid classifiers, they are supported by self-organizing maps and parametric approach. The first are used to data analysis by means SOM, U-Matrix and approximation to typical density functions. The space parametrization is based on marginal Box-Cox transformations. The optimal feature space is determined by maximizing the Bhattacharyyaɜs distance. The effect of size samples is considered. This method has been employed in the classification of visual defects in cast aluminium. Traditional classifiers as LVQ, MLP, and systems based on rules had been implemented.

Key Words: SOM, Box-Cox transformations, Bayes error.

1 Introduction

In the classifier design, it should be considered at least two matters. First that the majority of times the size of the samples is limited. Second, the design should allow the introduction of priori knowledge.
From the first point, it is possible to say that the problem consist on how difficult is to get samples. The finite samples set will affect the parametric approach, and also its conclusions. For example, from the theory point of view, in the Bayes error, it has been demonstrated that if the space dimension increases then, the classifier's power of discrimination also increases, considering the same relation signal-noise. However in practice, when a fixed number of samples is used to design a classifier, the error of the classifier tends to increase as the dimension gets large. This trend is called the Hughes phenomena [1]. Conclusions, the optimal space must have the minor Bayes error considering the samples number.
The second point has the goal to improve the classifier performance. When knowledge is added, it is possible to locate data transformations, and it could assure a major separation among groups. But, in general, transformations are local [2]. This purpose suppose the regions localization, but, how to get them?. The main purpose of this paper is to show a total technique, from clustering to parametric approach, minimizing the Bayes error.

2 SOM as clustering techniques

There are several studies that show how to use the self-organizing maps as tools of data analysis [3] [4]. The Kohonenɜs maps tend to a supposed topological order, such as combine the interaction of neighbour neural with the approximation to density function of input data [5]. The $\Re^n$ projections on lower dimension (e.g. $\Re^2$) allow easily to define the regions into the input data, as any point group in $\Re^n$ has a region in the map. However, the use of the Kohonenɜs maps as analysis technique is not definitive, due to the Euclidean distance between data disappear.

Ultsh et al [6] propose a new and complementary method to the SOM algorithm called U-Matrix. This procedure allow to get an overlay map to the SOM, with the goal to mark again the distance into the data space. However, the constructive proofs did not give the desirable results, there were no clear borders between clusters. In the last paragraphs, the convergence of Kohonenэs maps has been commented and it is said that the reference vectors approach to represent the density function of the input data, p(X). Moreover, Kohonen indicates that the reference vectors tend to be the centroides of the Voronoiэs set, reaching the expression (1):

$$M_i = \frac{\int_{U_i} Xp(X)dX}{\int_{U_i} p(X)dX} \qquad (1)$$

If in this equation p(X) is represented by the conditional functions of clusters owned by Voronoiэs set, U_i, neuron i:

$$p(X) = \sum_{j=1}^{h_j} p(X|\omega_j)p(\omega_j) \qquad (2)$$

where h_j is the number of clusters in U_i. The resulting expression is:

$$M_i \int_{U_i} p(X)dX = \int_{U_i} X \sum_{j=1}^{h_j} p(X|\omega_j)p(\omega_j)dX \qquad (3)$$

So, M_i is the representative element in its domain, in other words, it is the best approximation of p(X) in this limited region. It is possible to search the conditioned density function which better reflect p(X). With this purpose the next expression is minimized:

$$min(\int_{U_i} (M_i - X) \sum_{j=1}^{m_j} p(X|\omega_j)p(\omega_j)dX) \qquad (4)$$

For this plan, it is necessary to know the number of the different classes that are in U_i and the function type with its parameters. Summarising, the problem is to know how many clusters has the system, and the type of the density functions. As it is said before, if each Vononoiэs set tend to p(X), the number of clusters is defined as the number of the neurons in the net. With reference to conditioned functions, it is necessary to notice how the fault neurons (winner neurons with more than one pattern), have sets with more than one conditional distribution function. This reason suggests to use a technique, which only label those winner neurons corresponding with the unique pattern. If the proposition is accepted, the expression (4) will result:

$$min(\int_{U_i} (M_i - X)p(X|\omega_j)dX) \qquad (5)$$

where ω_j is the only winner class in U_i. Finally, it is necessary to know the optimal parametric approach with its estimated parameters.

$$min(\int_{U_i} (M_i - X)p(X|\theta_{\omega_{j1}}, \theta_{\omega_{j2}}, \ldots, \theta_{\omega_{jr}})dX) \qquad (6)$$

This method has practice difficulties, as the extended integrate in U_i domine, and the proposition of an unique pattern inferred with a few samples, in general.

The maps construction proof show that similar patterns tend to be next to each other, typical characteristic of SOM. It is deduced that it is possible to join them in a cluster. But, how could be decided the join?. The process of cluster formation consist on, join neighbour neurons, with the same label, and suggesting typical distribution function. Then it is possible to express the location and parameters of cluster as:

$$min(\sum_{i=1}^{H} \int_{U_i} (M_i - X)p(X|\theta_{\omega_{j1}}, \theta_{\omega_{j2}}, .., \theta_{\omega_{rn}})dX) \qquad (7)$$

where H is the number of domines jointed in the building of cluster. This is the general expression of a parametric cluster. However, neither the number of neurons that constitute the cluster is known, nor the type of distribution function. For it, it is proposed the following iterative procedure: First, only the winner neurons with a unique pattern type should be labelled, in contrast with Kohonen, who use a voted method. Second, cluster design should be initialized with U-Matrix and the description form of patterns. Third, cluster should be refined by means of an approximation to known density function, for example, normality test in each cluster.

As it was said before, the classifier looses discrimination when the dimension is increased, it considers a fixed number of training samples. It seems that it is necessary to search a compromise between dimension and error. But also it is known that the discrete variables help to build regions. With this purpose, in first step, the vector must have discrete components able to generate regions. Then, in the second step, when the optimal local classifier will be designed, those components will be deleted. This process of classifiers design permit to introduce priori knowledge.
The construction proofs have demonstrated that discrete variables favour the regions formation into the self-organizing map. When the map has been built, this components are deleted and the space dimension of the local classifier is decreased. The classification problem is transformed in the design of m-classifiers, with different spaces.
At this moment, the result is a full clustering technique. The next paragraphs will try to introduce a new parametric approach, in which clusters are taken in pairs.

3 Parameter clusters

Each cluster modelling based on some density function could be tried from Box Cox transformations [8] [9]. These transformations try to convert the distribution into normal. The expression is:

$$x_{jk}^{\lambda_k} = \begin{cases} \dfrac{(x_{jk}+m)^{\lambda_k}-1}{\lambda_k} & \forall \lambda_k \neq 0 \\ \log(x_{jk}+m) & \lambda_k = 0 \end{cases} \qquad (8)$$

where x_{jk} is the sample j of the component k. For the determination of vector Λ, it is necessary to maximize

$$L_{max}(\Lambda) = -\frac{N}{2}\log\left|\Sigma^{(\Lambda)}\right| + \sum_{k=1}^{n}(\lambda_k - 1)\sum_{j=1}^{N}\log x_{jk} \qquad (9)$$

where N is the sample number, n is the space dimension, and $\Sigma^{(\Lambda)}$ the covariance matrix of transformated data. For Λ location, the procedure has been developed by *Velilla et al* [10]. However, this method has involved different transformations in each cluster, and therefore each cluster has different level of distance. Having this situation, the alternative plan consist on the unique transformation in the region. The objetive is to preserve the same space for each cluster into this region.
The base of the study it to find the optimal Box Cox marginal transformations. The bias of estimator $L(\Lambda)$ follows a χ^2 with n freedom degrees. This event permit to find what the marginal transformation is, if that exist, which is able to convert the density function into a normal one. The solution will be more difficult as the number of groups increases in the region. For that, this study is centred on parameter cluster taken in pairs. Considering only two classes, c = 1,2, it try to demonstrate the possibility of common marginal Box Cox transformation. The feature is i, i = 1,2,...,n and the optimal marginal solutions are called λ_{i1}^* and λ_{i2}^*, then it is possible to define a common λ_i^*, if it satisfies:

$$\begin{gathered} L_{max}(\lambda_{i1}) - \frac{1}{2}\chi_1^2(\alpha_1) \leq L(\lambda_i^*) \\ L_{max}(\lambda_{i2}) - \frac{1}{2}\chi_1^2(\alpha_2) = L(\lambda_i^*) \\ \alpha > \alpha_c \end{gathered} \qquad (10)$$

where α_i is the confidence interval, and it is able to build an interval for the likelihood function in the true value of λ, while α_c is the critical level. In addition, working with marginal distributions, χ^2 will be a one degree of freedom.
The optimal point between two distributions is imposed by the equality condition in the confidence levels. It is a nonsense to be different, because of it would mean to more approximate a population to normal than the other one. Without a loose of generality, groups are labelled as 1 and 2, in a way that $\lambda_{i2}^* > \lambda_{i1}^*$ would be true. It could be noticed that the solution will be in the $[\lambda_{i1}^* \ \lambda_{i2}^*]$ interval and, because $L(\lambda_i)$ is a continue and convex function then $L_{max}(\lambda_{i2}) - L(\lambda_{i2})$ is decreasing monotony, and $L_{max}(\lambda_{i1}) - L(\lambda_{i1})$ is increasing monotony, both of them are higher or equal than zero. It implicates a unique solution, which is:

$$L_{max}(\lambda_{i2}) - L_{max}(\lambda_{i1}) = -\frac{n_2}{2}\ln(\hat{\sigma}_2(\lambda_i^*)^2) + \frac{n_1}{2}\ln(\hat{\sigma}_1(\lambda_i^*)^2)$$
$$+ (\lambda_i^* - 1)\left[\sum_{j=1}^{n_2}\ln x_{ji} - \sum_{j=1}^{n_2}\ln x_{ji}\right] \qquad \lambda_i^* \in [\lambda_{i1}^*\lambda_{i2}^*] \qquad (11)$$

3.1 Feature selection

After the possible marginal transformations in the region have been located, defined by the two clusters in clash, then, it suggest to realize the first selection. Not all the components will have a satisfactory transformation, because they do not comply (10). Afterwards, a space is calculated. This space has to comply normality in the two clusters, it has to maximize the Bhattacharyyaəs distance, and the measure has to have the minimum bias and variance. This measure has been determined by a finite number of samples.

Although the search of the optimal space could be explosive, because of $\binom{n}{m}$ combinations with m = 1,...,n, when the necessary condition of normality sub-space is applied, then the possibilities of exploration are limited. For the location of normal possible spaces has been used the next algorithm:

1. Inicialization:
   - 1.1. Guide vector, V, built by features space.
   - 1.2. V = [0], vector dimension 1, dim(V) = 1.
   - 1.3. Definition of maximum space dimension, n.
2. While (dim(V) != 0)
   - 2.1. Increase plus 1 the first component of V, V(1).
   - 2.2. Normality test in two populations,)a?.
   - 2.3. If (min(α_1, α_2) > α_c) then 2.4, else 2.5.
   - 2.4. Increase the V dimension, V = [V(1) V].
   - 2.5. If (dim(V) == n), then delete the first component, V(1).
   - 2.6. Go to 2.

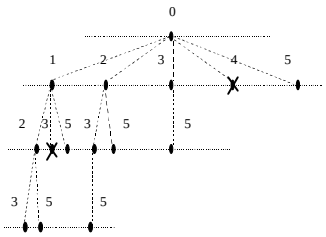

figure 1 Exploration tree of normal spaces

This procedure generates a tree of posible normal spaces. Figure 1 shows a graphic about the exploration with n = 5. The marks indicate that the normality test has been not passed. The different routes from the root to each node determines the possible solutions of the space. For fast computation, it is good to order the components from the smallest normality index to the biggest normality index, the smallest one will be called 1 and the biggest one n.

3.2 Bhattacharyyaəs distance, bias and variance.

Once the tree of possible normal spaces has been generated, it is possible to use the Bhattacharyyaəs distance as upper limit of Bayes error.

$$\mu = \frac{1}{8}(M_2 - M_1)^T\left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} + \frac{1}{2}\ln\frac{\left[\frac{\Sigma_1 + \Sigma_2}{2}\right]}{\sqrt{\Sigma_1\Sigma_2}} \qquad (12)$$

The first and second term, μ1 and μ2, measure the distance between two distributions, due to mean and covariance. When M_i and Σ_i are calculated by means of fixed samples, the result μ is different from its true value. For this reason, it is necessary to determinate its bias and variance.
Following Fukunagaəs method [11][12], about the effect of size samples in the Bhattacharyyaəs distance, expressions for bias and variance were got for every normal population $N_1(M_1, \Sigma_1)$, $N_2(M_2, \Sigma_2)$ inferred with N1 and N2 samples. Though, without a loose of generality, the distributions must be displaced and

transformed, in order to convert covariance-matrix in a I-Λ (whitening transformations). The distributions will look like N1(0,I) and N2(M,Λ). Bias and variance expressions are:

$$E\{\Delta\mu_1\} = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{2(1+\lambda_i)}\left(\frac{1}{N_1}+\frac{\lambda_i}{N_2}\right)+\frac{m_i^2}{(1+\lambda_i)^3}\left(\frac{1}{N_1-1}+\frac{\lambda_i^2}{N_2-1}\right)+$$
$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{m_i^2(1+\lambda_j)+m_j^2(1+\lambda_i)}{2(1+\lambda_i)^2(1+\lambda_j)^2}\left(\frac{1}{N_1}+\frac{\lambda_i\lambda_j}{N_2}\right) \quad (13)$$

$$E\{\Delta\mu_2\} = \frac{1}{4}\sum_{i=1}^{n}\left(1-\frac{2}{(1+\lambda_i)^2}\right)\frac{1}{N_1-1}+\left(\frac{1}{\lambda_i^2}-\frac{2}{(1+\lambda_i)^2}\right)\frac{\lambda_i^2}{N_2-1}+$$
$$\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\left(1-\frac{2}{(1+\lambda_i)(1+\lambda_j)}\right)\frac{1}{N_1}+\left(\frac{1}{\lambda_i\lambda_j}-\frac{2}{(1+\lambda_i)(1+\lambda_j)}\right)\frac{\lambda_i\lambda_j}{N_2} \quad (14)$$

$$Var\{\hat{\mu}_1\} = \frac{1}{4}\left[\sum_{i=1}^{n}\frac{m_i^2}{(1+\lambda_i)^2}\left(\frac{1}{N_1}+\frac{\lambda_i}{N_2}\right)++\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{m_i^2m_j^2}{2(1+\lambda_i)^2(1+\lambda_j)^2}\left(\frac{1}{N_1}+\frac{\lambda_i\lambda_j}{N_2}\right)\right] \quad (15)$$

$$Var\{\hat{\mu}_2\} = \frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{1}{(1+\lambda_i)}-\frac{1}{2}\right)^2\frac{1}{N_1-1}+\left(\frac{1}{(1+\lambda_i)}-\frac{1}{2\lambda_i}\right)^2\frac{\lambda_i^2}{N_2-1}\right] \quad (16)$$

The equations has been tested with Fukunagaэs conclusions. With this distance, its bias and variance, the optimum transformed space will be determined.

4 Defect classification result in aluminium cast

The commented techniques have been applied in the defect classification of cast aluminum. The classification is done with five or eight defect types. The algorithms used for defect segmentation are based on morphological processing, according to *Platero et al* [13]. For defect representation was used description techniques by means primitives of free grammar. The feature vector formation corresponds to zero, first and second moment, and its length is 18. When shape and size are measured, the vector should accomplish with Raoэs requisites [14]. Thus, the two principal components fix shape and size of defects.
The Kohonenэs maps was made with two dimensions (6 x 8) and minimal quantization error by hexagonal neighbourhood kernel. The U-Matrix was modified and the clustering result is figure 2. In the map appear 12 different clusters; neurons labelled with XX are those in which is impossible to infer, due to the short number of active samples. While YY labels correspond with fault neurons, there coexist samples of different patterns.
Discrete components were deleted by means “meda”. A preliminary selection was made with “manova” and “univar” [15]. Normality tests were based on Shapiroэs method, with signification levels higher than 0.01. It was accepted the normality with the condition of marginal distributions projected over principal components, besides, Bonferroniэs inequalities must be complied [16].
The frame and table show a method example applied to the region formed by clusters 1-4. At this moment, the classifier design is being worked with. The only thing that rests is to indicate that traditional classifiers (LVQ, MLP, systems based on rules) had not exceeded 90% of success.

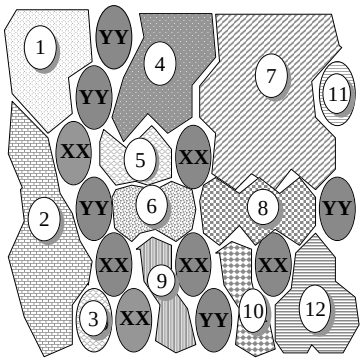

figure 2 Cluster: (1)Small dark transverse bands, (2)Medium dark transverse bands, (3) Dark stains, (4)Scratch , (5) (6) Medium dark longitudinal bands, (7)Small lacks (8)Big lacks (9)Big dark longitudinal bands, (10) Abnormal crystallize, (11)Bright bands, (12) Sticking

Clusters
1 4
Proposed components (meda)
2 5 13 11 1
Criterion value (Univar)
0.907 0.854 0.491 -1.10 -2.16
λ_{i1}^* **:**
1.42 -5.49 -7.19 -5.72 -4.31
Deleted components (by under α_c):
11
λ_{i2} **:**
-5.72 0.33 9.99 -0.14
Deleted components (by under α_c):
13
λ_i^*
-0.96 -0.39 -0.14

α_1	α_2	ε(%)	μ	$\Delta\mu_1$	$\Delta\mu_2$	Var(μ_1)	Var(μ_2)	Component
0.04	0.04	$5\ 10^{-4}$	11.4	1.0	0.01	11.41	$4\ 10^{-4}$	2
0.03	0.58	10^{-4}	12.9	1.7	0.04	1.23	$8\ 10^{-3}$	2 5
0.04	0.31	10^{-4}	13.2	1.8	0.07	1.26	$5\ 10^{-3}$	2 5 1
0.05	0.04	$3\ 10^{-4}$	12	1.3	0.05	0.86	$6\ 10^{-3}$	2 1
0.34	0.34	6	1.91	0.1	0.01	0.22	$8\ 10^{-3}$	5
0.41	0.30	1	3.25	0.3	0.04	0.38	$5\ 10^{-3}$	5 1
0.58	0.03	32	0.38	0.2	0.01	$6\ 10^{-3}$	$6\ 10^{-3}$	1

Table : α_1, α_2 confidence interval for cluster ; ε Bayes error (%) ; μ Bhattacharyya's distance ; $\Delta\mu$ bias ; Var(μ) variance ; transformed component

5 Reference

[1] Hughes, G.F, On the mean accuracy of statistical pattern recognizers, Trans. IEEE Inform. Theory, IT-14, pp. 55-63, 1968

[2] Patrick, E.A., Fundamentals of Pattern Recognition, Prentice Hall, 1972, pp 411-457

[3] Martín del Brio, B., Procesamiento Neuronal con Mapas Autoorganizados, Tesis Doctoral, Universidad de Zargoza, 1995

[4] Varfis, A., Versino, C., Clustering of socio-economic data with Kohonen maps, Neural Networks World, 2, 6, pp 813-833, 1992

[5] Kohonen, T., Self-Organizing Maps, Proceedings of the IEEE, vol 78, Nº9, pp 1464-1480

[6] Utlsh, A., Siemon, H.P., Kohonen´s self-organizing feature maps for exploratory data analysis, Proc. Int. Neural Networks Conf. INNC 90, Paris, pp 305-208, 1990

[7] Kohonen, T. Self-Organizing Maps. Springer-Verlag, Heidelberg, pp 95-112, 1995

[8] Andrews, D.F., Gnanadesikan, R., Warner, J.L., Transformation of multivariate data, Biometrics 27, pp. 825-840, 1971

[9] Box, G.E.P., Cox, D.R., An analysis of transformation, Journal of the Royal Statistical Society 26, pp 211-252, 1964

[10]Velilla, S., Barrio, J.A., A discriminant rule under transformation, Technometrics, vol.36, nº4, November 1994

[11]Fukunaga, K., Hayes, Effects of sample size in classifier performance, PAMI 11, pp.1087-1101, 1989

[12]Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, 1990

[13]Platero, C., Fernández, C., Campoy, Aracil, R., Hybrid system : Application to defect classification in cast aluminum, International Conference on Quality Control by Artificial Vision, pp 48-58, 1995

[14]Rao, C.R. Taxonomy in anthropology. Mathematics in the Archaeological and Historical Sciences, Edinburhg University Press, 1971

[15]Rauber,T.W., Barata, M.M., Steiger-Garcao, A.S., A toolbox for analysis and visualization of sensor data in supervision. Universidade Nova de Lisboa, 1995

[16]Cox, D.R., Small, N.H.J., Testing multivariate normality. Biometrika, 65(2), pp 263-272, 1978