Information criteria performance for feature selection

Mohamed Abadi, Olivier Alata, Christian Olivier,
Majdi Khoudeir
University of Poitiers
XLIM-SIC department UMR CNRS 6172
Chasseneuil-Futuroscope, France

Enguerran Grandchamp
University of Antilles and Guyana
LAMIA
Pointe-à-Pitre, Guadeloupe, France

Abstract—**This paper shows the information criteria (IC) performances in feature selection framework. Feature selection aims to select a representative subset among a wide set of features. We apply this approach to classify an hand segmented image. The performance is tested using various feature selection schemes (SFS, SBS, SFFS and SBFS) to select the candidate subsets. The accuracy of the approach is based on a good quality of the joint probability density approximation of the combined features. They are obtained using histogram optimized thanks to the adaptive arithmetic coding principles. Our approach is tested on different reference data. The subsets quality is evaluated using correct classification rate computed on multiple classifiers. Results show stability and convergence properties of this tool and its ability to select representative subsets (in the sense that the subset of feature is a good characterization of the classes in which the data belong). Information Criteria could be used for feature selection as a good alternative to other criteria.**

Keywords-component; information criteria, feature selection, histogram estimation and selection, feature scheme search.

I. Introduction

Feature selection is a thematic developed to give solutions to exploration and interpretation of large set of features. It allows reducing the dimension of the data by identifying subsets of representative's features. Representativeness is linked to an evaluation criterion. This approach has been used for many applications (such as text characterization [1], image indexation [2], bioinformatics [3], color image processing [4], etc.) to allow the execution in a reasonable time.

Three kinds of approaches have been developed for feature selection, called: wrapper [5], filter [6] and embedded [7].

- *Wrappers* methods use classifiers to generate and then evaluate the candidate's feature subset. This approach is probably the most pertinent but requires the training of a large number of classifiers, a manual specification of many parameters and an expertise in machine learning.
- *Filters* methods are based on a criterion function to measure the quality of the feature subsets candidates. These methods considerably reduce the computation time while previous ones lead to better results.
- *Embedded* methods try to combine the advantages of both approaches. They include the feature selection step in the learning phase of a classifier. Nevertheless, the computing time still remain important.

This is probably for all these reason that the *Filters* methods are the most popular. In order to have the best compromise between time and quality we will investigate these methods.

Several methods have been developed to generate the candidate's subsets because an exhaustive search over all the features is too time consuming [7]. We are interested in this study in methods based on sequential approaches and more particularly on Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). In this variety of approaches we can found the Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SFBS). Their advantage is that these methods are easy to implement, independent of the classification algorithms and decision criterion. So we can easily compare many criteria and underline their importance in the selection process.

Different criteria have been developed to evaluate candidate's subsets. They explore distance measures [8], statistical measures [4, 9] and more recently probabilistic measures based on the estimation of mutual information [10-12].

In this paper, we propose to study the Information Criteria (IC) based on the computation of the Maximum Likelihood (ML) with penalty [13]. We choose to compare the ICs with Wilk's criteria [4] because this criterion is easy to implement and not time consuming. All Criteria are integrated in SFS and SBS scheme. In fact, the computation of the ML requires estimating the feature distribution. We use an adaptive arithmetic coding method based to the probability density function (PDF) and with an optimal histogram computed for each feature [14, 15].

The information criteria are described in section 2. In section 3 we detail the way to estimate the histogram of each feature useful to compute the criteria. The evaluation of our approach and comparison with other methods is done in section 4. Then, conclusions and perspectives are given in section 5.

II. Information Criteria

In this study we propose an adaptation and integration of the Information Criterion AIC [13], BIC [16] et φ_β [17] (resp. for Akaike, Bayes Information Criterion and phi-beta) in a feature selection scheme.

The common way to write an Information Criterion for a parametric model Y is the following:

$$\mathrm{IC(Y)} = -2\log(\mathrm{ML}) + |\mathrm{Y}|\alpha(\mathrm{n}) \tag{1}$$

where $|\mathrm{Y}|$ is the number of free parameters which increases with the complexity of the model and $\alpha(\mathrm{n})$ is a penalty function which varies according to the criterion: $\alpha(\mathrm{n})$ is equal to 2, $\log \mathrm{n}$ and $\mathrm{n}^{\beta} \log\log \mathrm{n}$ for AIC, BIC and φ_{β} respectively. For φ_{β}, β must belong to $]0,1[$ to ensure good asymptotic properties. $|\mathrm{Y}|\alpha(\mathrm{n})$ allows to regularize the ML behavior to avoid over-parameterization. Then the minimization of the criterion allows a compromise between data fitting and model complexity.

In the other hand, the choice of β for φ_{β} criterion can to recover the usual penalty of the other criteria (AIC and BIC). Particular values of β, β_{AIC} and β_{BIC}, imply that φ_{β} criterion has the same penalty of AIC and BIC criteria respectively:

$$\beta_{\mathrm{AIC}} = \frac{\log(2) - \mathrm{logloglog}(\mathrm{n})}{\log(\mathrm{n})}$$

$$\beta_{\mathrm{BIC}} = \frac{\mathrm{loglog}(\mathrm{n}) - \mathrm{logloglog}(\mathrm{n})}{\log(\mathrm{n})}$$

Note that, when $\mathrm{n} > 1619$ then $\beta_{\mathrm{AIC}} < 0$. In [21], the authors also introduce the following particular values of β :

$$\beta_{\min} = \frac{\mathrm{loglog}(\mathrm{n})}{\log(\mathrm{n})} = 1 - \beta_{\max}$$

advise using φ_{β} criterion with β comprised between $\beta_{\min}$ and $\beta_{\max}$ $(\beta_{\min} < \beta < \beta_{\max})$. They even recommend to use φ_{β} with $\beta_{\min}$. As $\beta_{\mathrm{BIC}} < \beta_{\min}$, the $\varphi_{\beta_{\min}}$ criterion penalizes more than the BIC criterion.

To solve the feature selection problem using ICs we define the following context: let X be a set of M features and $\mathrm{Y} \subseteq \mathrm{X}$, subset of X with m features. We have n observations each ones assigned to a class k, $\mathrm{k} = 1, \cdots, \mathrm{K}$. K is the total number of classes. The joint probability of having class k and set X is expressed by $\mathrm{P(k,X)} = \frac{\mathrm{P(k,Y) \cdot P(X)}}{\mathrm{P(Y)}}$ because we have $P(k/X) = P(k/Y)$ when the classification is independent of the subset $\mathrm{X} \setminus \mathrm{Y}$ [13].

According to [13], then ICs are then defined by

$$\mathrm{IC(Y)} = -\sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{\mathrm{c} \in \mathrm{C_Y}} \mathrm{n_{k,c}} \log\left\{\frac{\mathrm{n_{k,c}}}{\mathrm{n(Y)}}\right\} + \left(|\mathrm{C_{k,Y}}| - |\mathrm{C_Y}|\right)\alpha(\mathrm{n}) \tag{2}$$

where $|\mathrm{C_{k,Y}}|$ and $|\mathrm{C_Y}|$ respectively denotes the number of areas (called "boxes" in the following) of the multidimensional histogram of non zero occurrences linked to the subset Y, taking into account classes or not. We suppose that $0 \times \log 0 = 0$, $\mathrm{n(Y)} = \mathrm{n}$ and $\mathrm{P(Y)} = 1$ for $\mathrm{Y} = \emptyset$ (empty subset). $\mathrm{n_{k,c}}$ is the number of observations of class k localized in the box c of the multidimensional histogram. The multidimensional histogram is obtained from an optimal histogram beforehand computed for each feature [14]. In the next section, we present the method used for histogram estimation.

III. Optimal partition and histogram estimation

Optimal Probability Density Function (PDF) estimation is a difficult problem. This estimate is based on a finite set of observations $x^n = (\mathrm{x_1}, \cdots, \mathrm{x_n})$ of n random variables $X^n = (\mathrm{X_1}, \cdots, \mathrm{X_n})$ of an unknown density f. Here we mention only the most popular of them: kernel methods [19] and the method by histogram [18].

We are particularly interested in non-parametric framework to estimate the density by histogram. Our goal is to find the optimal partition for each feature thus we focus on continuous piecewise estimators using arithmetic coding [14].

Let's define f an unknown density defined on an interval $\mathrm{I[a,b]} \subset \mathbb{R}$. Given $\mathrm{P} = \left(\mathrm{I_j}\right)_{\mathrm{j} \in 1, \cdots, \mathrm{q}}$ a partition of I into $\mathrm{q} \in \mathbb{N}^*$ intervals. Using $\mathrm{x^n}$ and under maximum likelihood, f can be estimate by:

$$\hat{\mathrm{f}}_{\mathrm{P}}(\mathrm{x}) = \sum_{\mathrm{j}=1}^{\mathrm{q}} \frac{\mathrm{n_j}}{\mathrm{n}|\mathrm{I_j}|} \mathbb{1}_{\mathrm{I_j}}(\mathrm{x}) \tag{3}$$

where $\mathbb{1}_{\mathrm{I_j}}$ denotes the indicator function of a set $\mathrm{I_j}$, $\mathrm{n_j}$ is the number of data $\mathrm{x_i}$ falling into $\mathrm{I_j}$ and $|\mathrm{I_j}|$ is the length of $\mathrm{I_j}$. The main problem of histogram selection is to determine which partition P must be chosen in order to estimate f by $\hat{\mathrm{f}}_{\mathrm{P}}$. Coq and al. [14] have developed a technique based on information criteria and the minimum length coding of data $\mathrm{x^n}$. This technique is based on two steps: (i) the arithmetic coding and (ii) fixed length coding. We recall here the standard result of these developments. For more details, the interested reader can consult [14].

The Minimum Description Length principle [14] suggests choosing the partition that minimizes this criterion to approximate f by $\hat{\mathrm{f}}_{\hat{\mathrm{P}}}$ as follows:

$$\widehat{\mathrm{P}} = \mathrm{argmin}\{\mathrm{CRIT}(\mathrm{x^n}, \mathrm{P}), \mathrm{P} \in \mathrm{P_{max}}\} \tag{4}$$

with

$$\mathrm{CRIT}(\mathrm{x^n}|\mathrm{P}) = -\sum_{\mathrm{j}=1}^{\mathrm{q}} \mathrm{n_j} \log \frac{\mathrm{n_j}}{\mathrm{n}|\mathrm{I_j}|} + \frac{(\mathrm{q}-1)}{2}\alpha(\mathrm{n}) - \mathrm{n}\log \mathrm{r} \tag{5}$$

where $\mathrm{P_{max}}$ is the maximum partition of *I* into *R* intervals all with length r.

According to (4) we estimate on each feature a density $\hat{\mathrm{f}}_{\hat{\mathrm{P}}}$ and an optimal partition $\widehat{\mathrm{P}}$.

Figure 1 illustrates two examples for PDF estimation using kernel density estimation method (KDE: based on the normal density function) [19], method used histogram and density estimation based on information criteria.

These methods are compared with theoretical PDF. The first plot (see fig. 1 a) shows a comparison on data generated by a uniform mixed law and the second plot (see fig. 1 b) on data generated by a Gaussian mixed law (each composed of three laws).

We have to define only the mixed coefficient parameters $\{0.25; 0.45; 0.3\}$ for the Uniform Mixed Law.

For the Gaussian Mixed Law, we have to define three parameters: *the mixed coefficient* $\{0.45; 0.2; 0.35\}$, *the means* $\{-1.5; 0.5; 2.5\}$ and *standard deviation* $\{0.75; 0.6; 0.45\}$.

We use 1000 realizations to generate the two examples. These plots show a good PDF estimation returned by the method using BIC criterion.

As we can see on Figure 1, approach used BIC criterion is comparable to the others and give a good approximation. She could be used to estimate the histogram.

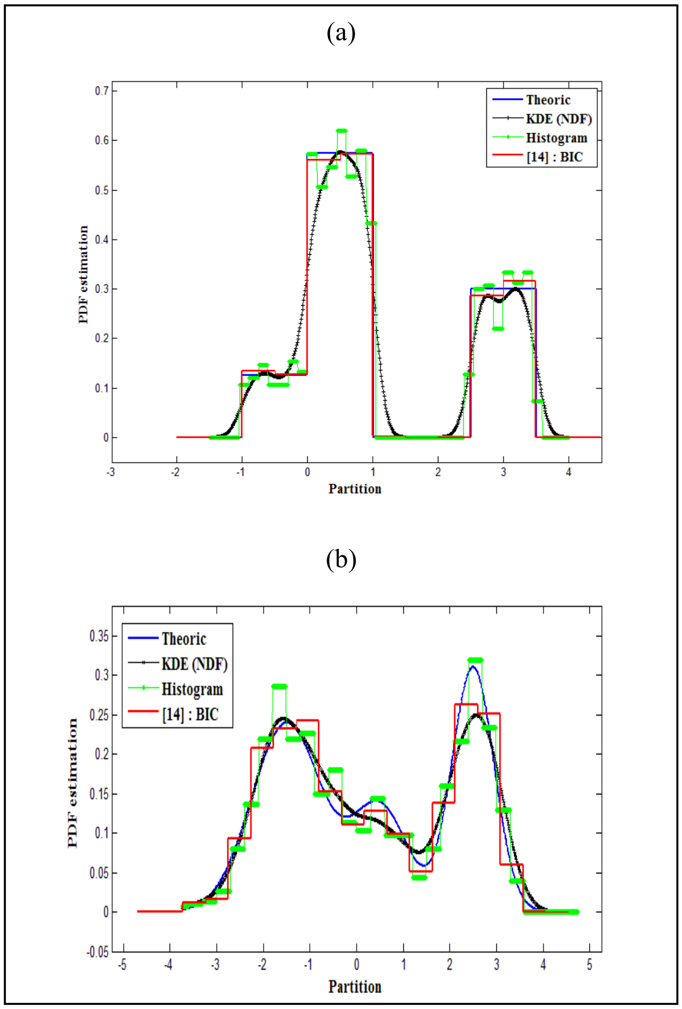

Figure 1. PDF estimation based on optimal partition obtained by BIC criterion and compared with other methods.

IV. Experiment

A. Data classification

The performances of our approach are evaluated and compared using the reference dataset [1] STATLOG (image segmentation). Table 1 shows the features of the data. The parameters are the number of realizations (N), features (M) and classes (K). We choose a cross validation type to build the learning and testing sets. The step is repeated n times (n=100) to obtain statistically representative results.

TABLE I. Characteristics of Features

Data	N	M	K	Cross Validation Type
Statlog	2130	18	10	10-Cross Validation

Furthermore, the quality of the results and the performance of the algorithms are measured using NB (Naive Bayes) and LDA (Linear Discriminant Analysis) classifiers. SFS, SFFS, SBS and SBFS schemes have been used to go through the features [20] and build the candidate's subsets.

Figure 2 and 3 represent a quantitative comparison between the results obtained using the information criterion φ_β (based on $\beta_{\min}$: he shows the highest good classification rates) and the Wilk's criterion (Wilks on the plots) as defined in [4]. Each figure represents the mean classification percentage for each subset dimension. It illustrates the behavior of the criteria depending on the schemes and classification algorithms used.

Results show a better stability of the information criteria compared to Wilk's criterion. Indeed, we remark that the Wilk's curve is not always increasing, i.e. the classification rate quickly decreases as the subset dimension increases (plots with SFS and SBS algorithms).

Generaly, the plots from ICs are often merged (especially with SBS and SFS algorithm), that is to say that the different criteria lead to the same subsets at each iteration. The different kind of information criteria didn't influence so much the feature selection. Here we have chosen to plot only the curve using $\varphi_{\beta_{\min}}$ criterion for easy reading and curves obtained have better visibility. In the same way, we also observe that plots are stabilized quicker with information criteria.

Plots show a good behavior and a quick convergence to optimal subsets for the $\varphi_{\beta_{\min}}$ approach. The dimension of the retained subsets using our approach is lower than with Wilk's criterion which is a good result in a feature selection context.

Furthermore, the use of the different classification algorithms to evaluate the quality of the feature selection has a low influence on the results (quite the same curve and classification rate). This result tends to prove that the selected features are robust. Nevertheless, we remark that the LDA classifier has a better behavior and the classification rate is better for all criteria and schemes.

[1] UCI Repository data base : http://archive.ics.uci.edu/ml/datasets.html

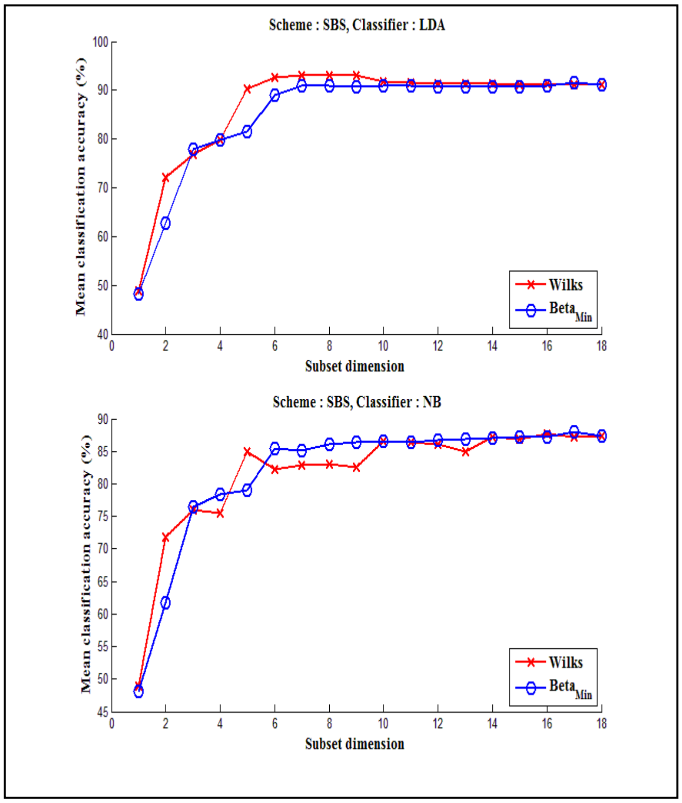

Figure 2. 10-fold CV means classification accuracy for STATLOG data with SFS scheme search and two classifiers (LDA, NB).

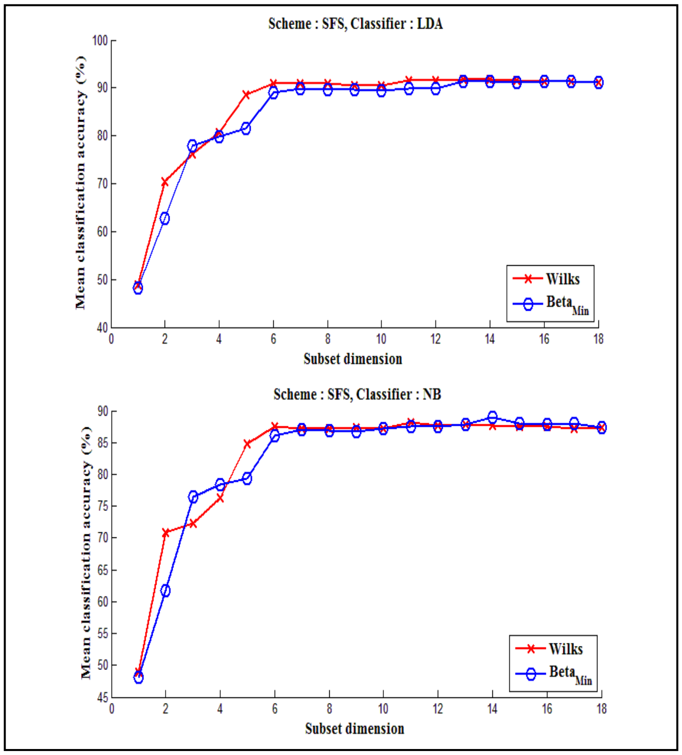

Figure 3. 10-fold CV means classification accuracy for STATLOG data with SBS scheme search and two classifiers (LDA, NB).

B. *Image segmentation*

We build an image (Figure 4.a) composed of 3 kind of forest (F1, F2, F3) and 3 kind of agricultural (A1, A2, A3) samples. This image is 1m/pixel spatial resolution. The repartition of the samples is given by Figure 4.b. The main goal is to illustrate representative configurations of forest organizations to evaluate our selection process performances.

Figure 4. shows the mosaic image (Figure 4.a) classified with LDA and NB algorithms using texture features based on co-occurences matrices (energy, contrast, homogeneity, dissimilarity, entropy, correlation) [22]. These values are computed on each band in the RGB color space in a 16x16 window using 32 gray level. There is 18 features for each pixel.

The second line of the figure 4. presents the segmentation result using LDA (column 1) classifier and NB (column 2) with all texture features. When applying Wilk's features selection approach we obtain the same results. Indeed, the minimum of the Wilk's criterion (computed on each feature subsets) is reached for the complete set.

However, the same approach using IC criterion (based on β_{min}) reaches an optimal subset composed of 6 features. The third line of figure 4. shows the corresponding segmentation using SFS scheme.

The results are visually close (Figure 4. line 2 and 3). It shows that the feature selection process reaches a good quality subset and allow to keep a good trade-off between quality and number of features (more generally between quality and data size).

V. Conclusion

We present in this paper a feature selection algorithm based on information criteria. We use continuous piecewise estimators with an adaptive arithmetic coding to match the probability density function (PDF) and to define the optimal partition for each feature.

Tests on the reference datasets STATLOG and comparisons with Wilk's criterion show good results for our approach for different classification algorithm.

In this paper we estimate for each feature an optimal partition. In the future we expect to estimate a common partition for all features or at least for all features composing a candidate's subset.

To apply this approach to wide feature sets (more than 2000 features) we are confronted to the problem of dominance of the penalty part of the information criteria. Indeed, the complexity of the model exponentially growth when adding features. To solve the problem we will adopt a multi-objective approach to throw the information criteria in a (Maximum Likelihood, Penalty) space. This will allow the optimization of the two parts of the criteria. This approach has been successfully applied with Mutual Information criteria and will be extended to Information Criteria.

Finally we also have to demonstrate that features selected by the information criteria approach are more robust to classifier evaluation than the features selected by Wilk's criterion. In this case, this approach will be a better alternative to select features in other contexts than classification (data simplification, data visualization, etc.).

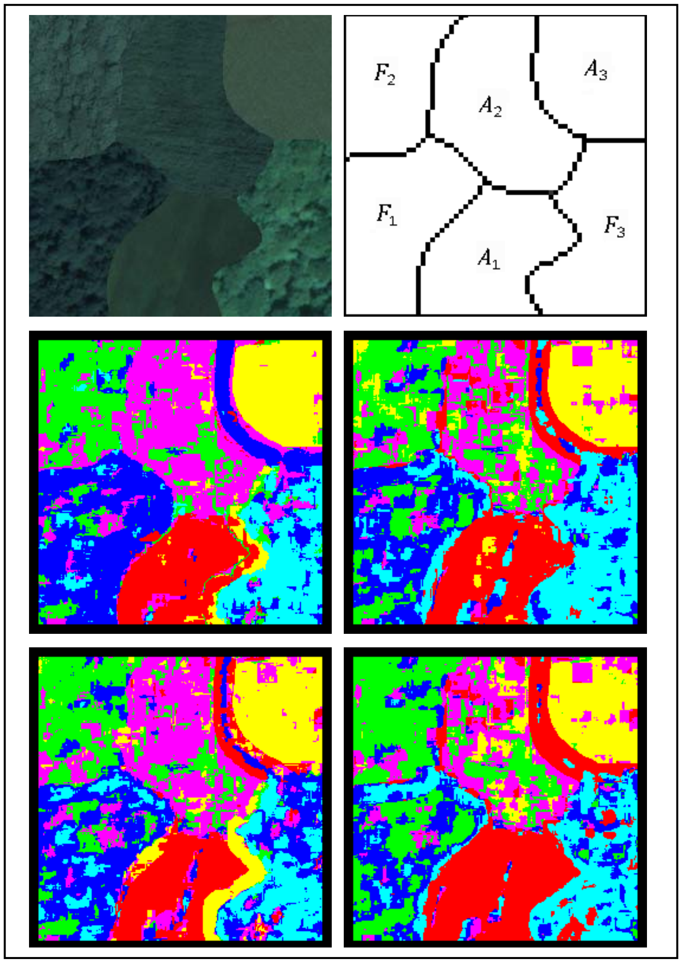

Figure 4. Line 1 : mosaic image (256 × 256) and labels repartition. Line 2 : image segmentation based on feature selection using all features and two classifiers (LDA: column 1, NB: column 2). Line 3 : segmentation based on feature selection used $\beta_{\min}$ and SFS schems with the same classifiers.

Acknowledgment

The authors want to thanks the European institutions and more precisely the INTERREG IV program within the *CESAR Part II* project leads by the University of Antilles and Guyana and the Poitou-Charentes region.

References

[1] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research 3 (2003) 1289-1305.

[2] J.G. Dy, C.E. Brodley, A.C. Kak, L.S. Broderick, A.M. Asien, Unsupervised feature selection applied to content-based retrieval of lung images, IEEE TPAMI 25 (3) (2003) 373–378.

[3] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[4] A. Porebski, N. Vandenbroucke, L. Macaire, Comparison of feature selection schemes for color texture classification, Int. Conf. on IPTA Paris, July 2010.

[5] R. Kohavi and G. John, Wrapper for Feature Subset Selection, Artificial Intelligence 97 (1-2) (1997) 273-324.

[6] P. Pudil and J. Novovicova, Novel methods for subset selection with respect to problem knowledge, IEEE Intell. Syst. 13 (2) (1998) 66–74.

[7] Y. Sun, S. Todorovic, and S. Goodison, Local-Learning-Based Feature Selection for High-Dimensional Data Analysis, IEEE trans. PAMI 32 (9) (2010) 1610-1626.

[8] Jianning Liang, Su Yang, Adam C. Winstanley, Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition 41 (5) (2008) 1429-1439.

[9] G. Qu, S. Hariri, M. Yousif, A new dependency and correlation analysis for features, IEEE Transactions on Knowledge and Data Engineering 17 (9) (2005) 1199-1207.

[10] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE trans. PAMI 27 (8) (2005) 1226-1238.

[11] A. Al-Ani, M. Deriche, J. Chebil, A new mutual information based measure for feature selection, Intelligent Data Analysis 7 (1) (2003) 43-57.

[12] H. Liu, J. Sun, L. Liuand, H. Zhang, Feature selection with dynamic mutual information, Pattern Recognition 42 (7) (2009) 1330-1339.

[13] Y. Sakamoto and H. Akaike. Analysis of cross classified data by AIC. Ann. Inst. Statist. Math, 30(B) (1978) 185-197.

[14] G. Coq, O. Alata, Y. Pousset, X. Li and C. Olivier. Law recognition via histogram-based estimation. IEEE ICASSP, Taïpei (Taïwan), (2009) 3425-3428.

[15] J. Rissanen, T. P. Speed, and B. Yu, "Density estimation by stochastic complexity.," IEEE Transactions on Information Theory, vol. 38, no. 2, pp. 315–323, 1992.

[16] C. Olivier and O. Alata. Optimisation in Signal and Image Processing. Chapter 4, Information Criteria : Examples of Applications in signal and Image Processing. ISTE, Wiley, (2009) 79-110.

[17] D.A. El-Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. 115 (1996) 291-299.

[18] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. ESAIM Probab. Stat., 10, (2006) 24-45.

[19] D. Scott. Density Estimation. Theory, Practice and Visualization. Wiley, 1992.

[20] H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering, 17 (3) (2005) 491-502.

[21] C. Olivier, F. Jouzel and A. El Matouat. Choice of the number of component clusters in mixtures models by information criterion. Proc. Vision Interface, (1999) 74-81.

[22] R. M. Haralick, I. DINSTEIN, and K. SHANMUGAM. Textural features for image classification. IEEE Transaction on Systems, Man, and Cybernetics SMC-3 (1973), 610–621.