Consensual clustering for unsupervised feature selection. Application to SPOT5 satellite images indexing

Marine Campedel MARINE.CAMPEDEL@TELECOM-PARISTECH.FR
Ivan Kyrgyzov IVAN.KYRGYZOV@TELECOM-PARISTECH.FR
Henri Maître HENRI.MAITRE@TELECOM-PARISTECH.FR
*TELECOM ParisTech *, CNRS LTCI*
Department Signal and Image Processing, TSI
46 rue Barrault,
75013, Paris, France.

Editor: Saeys et al.

Abstract

Satellite images are numerous and weakly exploited: it is urgent to develop efficient and fast indexing algorithms to facilitate their access. In order to determinate the best features to be extracted, we propose a methodology based on automatic feature selection algorithms, applied unsupervisingly on a strongly redundant features set. In this article we also demonstrate the usefulness of consensus clustering as a feature selection algorithm, allowing selected number of features estimation and exploration facilities. The efficiency of our approach is demonstrated on SPOT5 images.

1. Introduction

Despite the huge amount of works dealing with image indexing, dealing with satellite images is still an open issue. These large and content-rich images are generally manually exploited by human experts in specific application domains. Depending on the application, the objects of interest are not the same: for example meteorologists are interested in clouds while urbanist photointerpreters are concerned with roads, buildings and green spaces. The diversity of remote sensors allows people to work with the images most adapted to their application contexts.

Moreover these images are numerous (around 100 Gigabytes each day for the SPOT satellite alone) and weakly exploited. New satellites, like Pleiades [1] will soonly be launched and provide very high resolution images (450 images each day with resolution around 70cm per pixel). The manual exploitation of such images is untractable: (semi-)automatic and efficient processings are urgently required to facilitate the access to their content.

Hence we are interested in high resolution images indexing and particularly in the question: what are the best features to be extracted for such images? Are classical color (spectral), shape and texture features suitable? We propose to answer the first question by using a methodology based on automatic feature selection algorithms. The idea is simply to reuse features proposed in the literature, concatenate them and study the redundancy of

1. http://www.cnes.fr/web/print-3227-pleiades.php

the resulting features vector using adequate feature selection algorithms. The goal is then to identify a subset of these features being able to represent the informative content of the images while reducing the storage cost. Since the objects of interest are different according to the pointed applications, we aim at developing unsupervised approaches to perform the automatic selection, based on purely objective criteria.

Automatic feature selection (FS) algorithms are related to an abundant literature since 15 years Kohavi and John (1997); Blum and Langley (1997); Guyon and Elisseeff (2003). In our previous works Campedel and Moulines (2004); Campedel et al. (2005, 2007), we exploited the tool called Spider Weston et al. (2004) and compared both supervised and unsupervised feature selection approaches using supervised classification performances. Our application to satellite images demonstrated the usefulness of simple unsupervised filter methods to reduce the redundancy introduced on purpose in the features vectors. These methods are composed by two steps : i) group features using their similarity, ii) identify representant for each features group. This is a classical way to perform data selection, but in our study we apply it to the features. The most simple example (and the most efficient through our experiments) consists in applying k-Means algorithm on the features and then keep the best representant of each cluster (the nearest to the centroid): we call it k-Means-FS. This can easily be extended to Kernel k-Means clustering (Kk-Means-FS). We also derived a similar algorithm called SVC-FS based on Support Vector Clustering (SVC) Ben-Hur et al. (2001): in this case the selected features are directly identified by the support vectors and the clusters are then represented by their contours.

In this article we propose to introduce a consensus clustering method that will overcome the three main problems we still have:

- the estimation of the number of features to be selected;
- the influence of the clustering procedure initialisation;
- the visual analysis of the selected features.

In the next Section 2, we briefly describe our methodology and give the best results obtained when comparing the supervised Fisher selection process and our unsupervised k-Means-FS. We next introduce the consensus clustering algorithm and the associated feature selection algorithm in Section 3. Section 4 demonstrates the efficiency of this method in the context of satellite image texture classification. We finally propose conclusion and further works in Section 5.

2. Unsupervised feature selection

2.1 Methodology

The main idea of our methodology is to exploit objectively evaluated features proposed in the literature through concatenation followed by automatic selection. We make the hypothesis that the different features sets are highly redundant and relevant to our final task. In order to evaluate both supervised and unsupervised algorithms, we evaluate the selected features set with the classification of a labelled database; mean error rate and standard deviation estimated using cross-validation approach are used as evaluation criteria.

Figure 1 illustrates this process. Representation entropy Mitra et al. (2002) can also be used as a redundancy measure but we do not use it in this article. It is worthy to note that this methodology can be applied to anykind of data (numerical, symbolic, graphs, ...) but in our simulations we only deal with vectorial and numerical values.

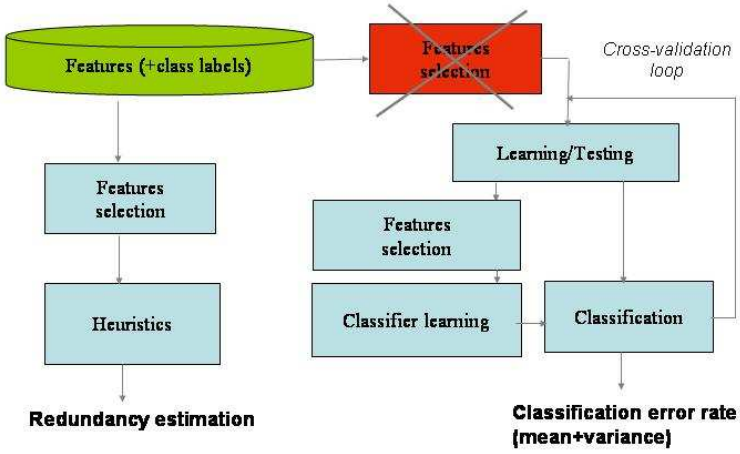

Figure 1: Evaluation of feature selection algorithms using a supervised classification task or heuristics. When only one database is available, selection must be performed on the training set, in the cross-validation loop, whereas when another database is available, feature selection is performed before the cross-validation loop.

2.2 Results and limits

The labelled database is composed by small 64×64 sub-images croped from SPOT 5 HMA panchromatic images with resolution 5m per pixel. These small images were manually selected in order to illustrate unambiguously 6 texture classes illustrated in Figure 2 (city, forest, sea, fields, desert and clouds). Each class is populated by 600 images, which results in a 3600 labelled database. We propose to compare several texture features proposed in the literature (cf Table 1) with few geometrical features resulting from an edge analysis. These attributes are classically normalised with 0 mean and variance 1, individually, over the 3600 data.

As a synthetic presentation of our preceding works Campedel and Moulines (2004); Campedel et al. (2005, 2007), we propose a comparative result of two feature selection algorithms. Practically these two methods were shown to be the lonely tractable ones on big image databases and to always produce better recognition results than the other ones. The first one is supervised based on the Fisher Discriminant Analysis as implemented in the Spider tool Weston et al. (2004). The second is unsupervised, called k-Means-FS; it

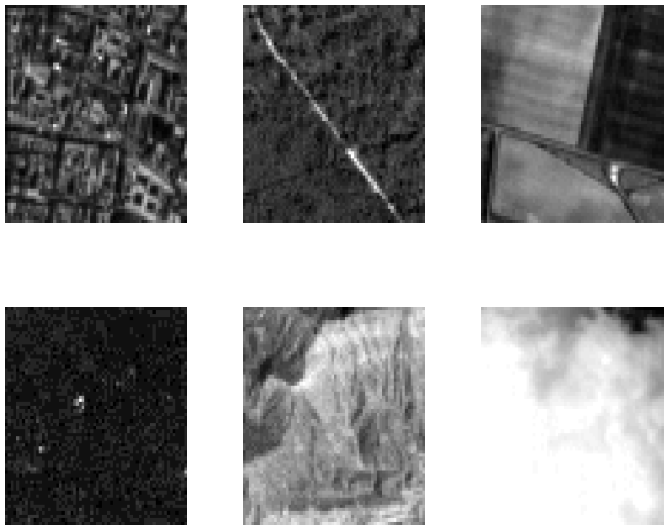

Figure 2: Six texture classes. From left to right, up to bottom: city, forest, field, sea, desert and clouds.

Model	Nb	Reference	Extracted attributes
Haralick	78	Haralick et al. (1973)	Cooccurrence matrices. 13 coefficients and 4 orientations + means and amplitudes.
Gabor	40	Ontrup and Ritter (1998)	mean+std on filtered images with 5 scales and 4 orientations filterbank.
Qmf	8	Mallat (2003)	Std computed on each subbands for a 2 stages decomposition + mean the last subband.
Go	15	Kyrgyzov (2008)	Geometrical attributes.
MV	2		Mean and std on the whole image.
Total	143		

Table 1: Features computed on the images. [] indicates publication references. We finally produce highly redundant features vectors of dimension 143.

relies on a k-Means feature clustering and then the selection of the features that are the closest to the clusters centroids. Figure 3 illustrates the classification performances (using a Gaussian SVM classifier) as a function of the number of selected features. Without any selection, the mean error rate is 1.7% $\pm$ 0.6%. We observe that:

- the unsupervised algorithm is as good as the supervised one;

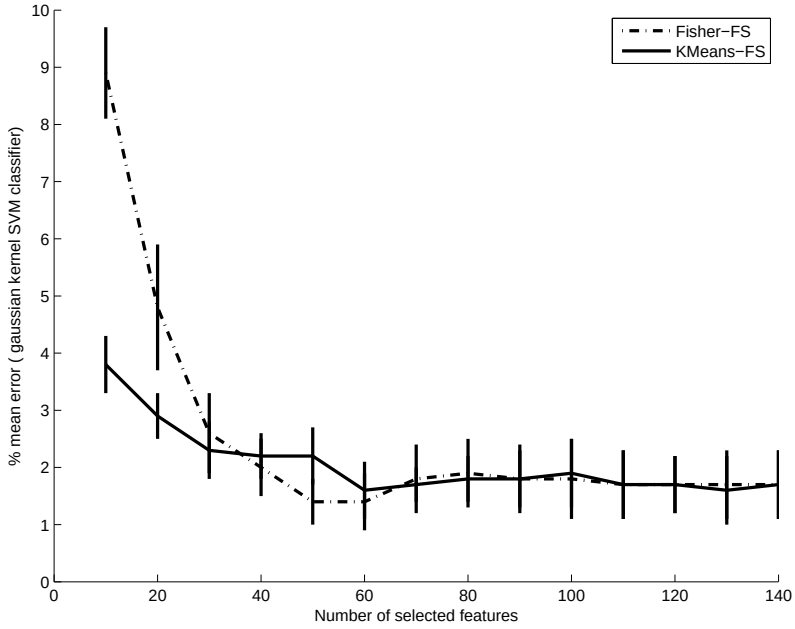

Figure 3: Mean classification error rate (5 cross validation loops) obtained with a Gaussian kernel SVM after selection by two methods: Fisher-FS (supervised) and k-Means-FS (unsupervised), as a function of the number of selected features. The vertical lines represent one standard deviation apart from the mean error. Without selection the results are 1.7% $\pm$ 0.6%. The best result with Fisher-FS is 1.4% $\pm$ 0.4% (50 selected features) and with k-Means-FS 1.6% $\pm$ 0.5% (60 selected features).

- both algorithms behave similarly: we observe a high decreasing slope from 10 to 30-40 selected features and then a stable region after a statistically non significant minimum;
- the classification results are similar with and without selection in the stable region.

These results are encouraging since they demonstrate the ability of simple unsupervised algorithm to capture the informative features in a redundant set. However there are still drawbacks related to k-Means clustering like: the initialisation procedure and the number of features to be selected. We hence propose a new unsupervised algorithm based on a consensus approach.

3. Consensual clustering

3.1 State of the art

The idea of combining classifiers is not new, especially in the supervised classification community. The goals are generally to enhance classification performances, identify common classes from different classification algorithms, classify partially labelled or distributed data, ... However when dealing with unlabelled data, this problem is not obvious (no predefined

classes, no well-established error measures) and the literature is less prolific and old Diday (1979); Michaud and Marcotorchino (1979); Marcotorchino and Michaud (1982); Fred and Jain (2005); Kuncheva (2004).

There exist many different methods to aggregate information pieces issued from different clustering techniques. One of the most attractive is based on the use of a co-association matrix Diday (1979); Michaud and Marcotorchino (1979); Marcotorchino and Michaud (1982). This matrix simply illustrates the fact that two data are or are not classified in the same cluster; this approach does not depend on a cluster numbering process nor reflect features spaces (which enable to deal with incomplete data). A typical problem is how to combine clustering results using k-Means algorithm with random initialisations and random number of clusters? This problem has been considered as the detection of "formes fortes" (strong shapes) proposed in Diday (1979), and solved using a mean co-association matrix. Several approaches have been proposed to issue a consensus clustering from the set of given clusterings Michaud and Marcotorchino (1979); Marcotorchino and Michaud (1982).

In the next section, we propose a well-defined mathematical objective function to solve the consensus problem.

3.2 Objective function

For each clustering result $p = 1 \dots P$, the co-association matrix A is a symmetric binary square matrix of size $N \times N$, N being the number of data to be classified. Each element A^p_{uv} is 1 if u and v are in the same cluster and 0 otherwise.We may also describe this p^{th} result using the allocation matrix B^p with N lines and J_p columns, with J_p being the corresponding number of clusters. B^p_{uj} is 1 when u belongs to cluster C_j. We have the relation:

$$A^p = B^p B'^p$$

where ' denotes the matrix transposition operation. For P different results, the mean co-association matrix A is given by:

$$A = \frac{1}{P}\sum_{p=1}^{P} A^p = \frac{1}{P}\sum_{p=1}^{P} B^p B^{p\prime} \tag{1}$$

For high P, A_{uv} estimates the probability of u and v to be in the same cluster. Our goal is to obtain the consensual allocation matrix B^s from A in order that $D = B^s \cdot B^{s\prime}$ be the most similar as A as possible. From the knowledge of A, we then have to minimize the following error:

$$\begin{aligned} E &= \sum_{u=1}^{N}\sum_{v=1}^{N}\left(\sum_{r=1}^{N}(B^s_{ur}B^s_{rv}{}') - A_{uv}\right)^2 = \sum_{u=1}^{N}\sum_{v=1}^{N}(D_{uv} - A_{uv})^2 \\ &\textit{under the constraint} \qquad B^{s\prime}B^s = \mathbf{I} \\ &\qquad Tr(\mathbf{I}) = \sum_{i}^{N}\mathbf{I}_{ii} = N \\ &\qquad B^s_{uv} \in \{0,1\} \end{aligned} \tag{2}$$

where $\mathbf{I}$ is a diagonal matrix $N \times N$ whith diagonal elements corresponding to final cluster sizes. This formulation is similar to the Kernel k-Means objective function, with A the Kernel. However we avoid initialisation problems and a priori knowledge about the number of clusters by using our efficient implementation.

3.3 Proposed solution

The complete exploration of all possibilities in untractable; we then propose a heuristic solution based on single-link algorithm. This new algorithm, called LSEC (Least Square Error Combination), procedes iteratively while reducing the error E(Eq. 2):

Algorithm 2: Pseudo code of *LSEC*-algorithm

1: Set B^s as the identity matrix, $J \leftarrow N$, $i \leftarrow 1$ and $E^{(i)} \leftarrow N^2$.
2: Find clusters' indexes $(j,k) = \arg \max_{u \in C_j, v \in C_k} A_{uv}$; $j,k = 1,...,J$, $j \neq k$.
3: Set $B^* \leftarrow B^s$.
4:Merge two clusters j and k by $B^s_{uj} \leftarrow \left(B^s_{uj} + B^s_{uk}\right)$, with $u = 1,...,N$.
5:Remove column k from matrix B^s.
6:$E^{(i+1)} \leftarrow \sum_{u=1}^{N} \sum_{v=1}^{N} \left(\sum_{j=1}^{J} (B^s_{uj} B^s_{vj}) - A_{uv} \right)^2$.
7:**if** $E^{(i+1)} \leq E^{(i)}$, **then**
8: $i \leftarrow i + 1$,
9: $J \leftarrow J - 1$,
10: go to **Step 2**;
11:**else** $B^s \leftarrow B^*$, B^s is the optimal partition, stop.

The optimal number of clusters J is found when the error E in Eq. (2) is minimum. At the first step we initialise B^s as the identity matrix supposing that each cluster has only one sample. Error $E^{(1)} = N^2$ is initialised to have its maximal value. The second step is an exploration heuristic choosing the clusters C_j and C_k that contain maximally connected examples (u and v). Merging is going on until $E^{(i)}$ reaches a minimum value.

Practically it is possible to avoid the storage of A and to efficiently initialize B^s using neighbour graphs. We do not detail these results here, please refer to Kyrgyzov et al. (2007); Kyrgyzov (2008) for more explanations. The main point is that this algorithm can be applied to a high number of data. In our context, we deal with images with size 12000×12000 pixels, that are cut into 64×64 overlapping sub-images: hence we get around 130 000 vectorial signatures for each big image. To test our algorithms we restrict ourselves to small databases but in practice we are dealing with a very huge amount of data.

3.4 Application to feature selection

The basic idea is the same as before: i) group similar features i.e. apply clustering algorithms and find the consensus, ii) select the cluster representatives. Considering the promising results we obtained with k-Means-FS, we applied k-Means algorithm with different random

initialisations and all possible values for k (in $2 \dots 143$). The cluster representant f_j is chosen as the more stable feature in the clusters. To be more precise, the analysis of the consensus result reveals different kinds of data:

- stable features: they lie in stable clusters.
- frontier features: they have unneglectable connectedness probability with at least two different stable clusters and can appear as singleton.
- outliers: they correspond to singleton clusters with very weak connectedness to anything.

Let remind that our method is meant to select features among a redundant and a priori relevant set of features. Hence we keep outliers in the final selection and remove frontier clusters. Other choices for the representatives are possible; this simple choice already gives interesting results as presented below. The stability criterion is computed, using the co-association matrix as well as the result of the consensual clustering. For a unique feature i, the stability is computed as:

$$S_i = \frac{1}{\#C_k - 1} \sum_{i,j \in C_k} A_{ij}$$

with $\#C_k$ being the number of elements in cluster C_k. For a cluster C_k, stability is computed as:

$$S^k_{intra} = \frac{1}{\#C_k(\#C_k - 1)} \sum_{(i,j) \in C_k} A_{ij}$$

We also derive a stability degree (called also degree of connectivity) between two clusters as:

$$S^{k,l}_{inter} = \frac{1}{\#C_k \#C_l} \sum_{i \in C_k} \sum_{j \in C_l} A_{ij}$$

S_{intra} and S_{inter} are very usefull to visualize cluster interactions using for example a graph representation (cf Figure 4).

4. Experiments

Unsupervised feature selection does not need any labelled database. Hence we defined a randomly constructed dataset called SpotRdn composed by 25000 64×64 randomly croped sub-images from 32 SPOT5 panchromatic scenes all over the world. We use it to perform the selection and then we will use the labelled database Sat3600 to evaluate our performances. The same features are extracted (cf Table 1) from the two datasets.

Because of the curse of dimensionality we select 100 data (vectorial signatures) among the 25000 available using SVC clustering (cf Section 1) and keeping all support vectors. We choose to apply this data selection method because we do not know anything about data distribution and SVC can adapt to any cluster shape. The number of selected data is arbitrarily chosen close to the number of features. We do not present here a complete evaluation for this data selection.

4.1 Efficiency

Let consider that our final application is the classification task defined on the Sat3600 database. We now compare the classification performance in terms of mean error rate using cross-validation, obtained using or not the selection algorithm. Results are presented in Table 3, using two different classifiers (3-NN and SVM). Note that contrarily to the

	3-NN	SVM (rbf 10)	D
All attributes	3.5±0.8	1.7±0.6	143
Consensus selection	3.7±0.6	1.5±0.4	28

Table 3: Mean± std classification error (%) obtained on sat3600 database, with and without consensual selection, using two different classifiers: k-NN using 3 neighbors and SVM with Gaussian kernel.

experiment shown in Figure 3, we do not have to set any parameter. The consensual selection process is able to determine the best number of features (here estimated to 28). Moreover the discriminative power obtained by the consensual set is as good as the original set, and as the previously tested approaches KMeans-FS and Fisher-FS. This observation does not depend on the classifier type.

4.2 Feature mining

Not only this consensus-based selection algorithm is providing discriminative features but it also enables us to mine the features set. The consensual clustering identified 76 clusters (including 48 unitary clusters corresponding to frontier features). Hence, as mentionned before, the selected set contains only the more stable feature of each big cluster. In the current experiment we select 28 features.

We observe that most of the clusters are small (3 elements), except one of them with size 12 (Cluster 21). This big cluster is semantically consistent since it contains only "mean" features ie mean of the image, as well as "mean" corresponding to Gabor or Qmf ouputs. This confirms the well-known idea that "mean" features corresponding to several scales and orientations do not convey discriminant information.

Using the values of S_{intra} and S_{inter} (cf Figure 4), we can also put in evidence peculiar clusters and produce semantic interpretations. For example Cluster 2 contains homogeneous features (same Haralick feature type concerning gray level inverse difference) computed with different orientations; similarly Cluster 5 groups features corresponding to correlation measures obtained with different orientations; Cluster 27 only contains geometrical attributes related to the length of the linear segments contained in images. Moreover the high stability of these clusters let us think that these types of information are particularly interesting in SPOT5 images and not represented by other feature types.

Similarly we can study relations between clusters. For example we observe that Cluster 17 is connected to clusters 16, 22 and 24 with a score above 20%. All these clusters contain features computed as "variances" (from QMF decomposition, Gabor filtering or from the original image), corresponding to different scales but the same orientation. We conclude

here that the scale granularity we used for the Gabor filtering is too fine and a simple dyadic decomposition (as used by the QMF decomposition) is sufficient; we also note that the 4 orientations are needed.

Finally, we defined experimental parameters to accelerate the process (number of k-Means initializations, step for the k value, number of selected data). When varying these parameters, the estimated number of clusters is quite stable ($\pm$ 2) and the cluster nature is similar, but the selected features are different. In fact we did not try different selection strategies and the choice of the "more stable" representatives should be discussed. This is part of our perspectives.

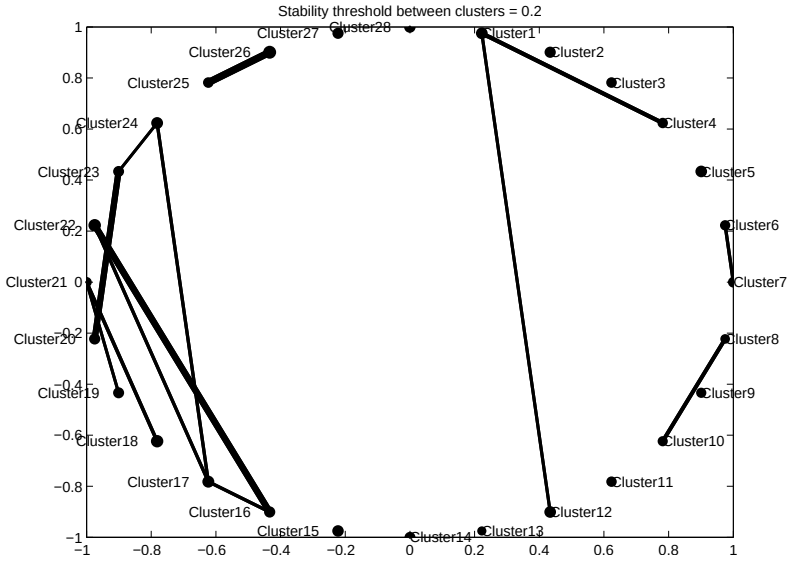

Figure 4: Graph representing clusters and their connections. The thickness of each link is proportional to the value S_{inter}. Isolated clusters correspond to high S_{intra} values.

5. Conclusion

In this article we proposed to use a simple unsupervised strategy to explore SPOT5 images features. We demonstrated the interest of a consensus clustering approach to solve the problems of i) number of selected features estimation ii) influence of initial parameters, iii) features mining through stability analysis. We observed that consensus selection makes semantically consistent clusters emerge, which helps interpretation. Moreover, when considering a simple landcover classification task, the selected features set has proven to be as discriminative as the original (redundant and a priori relevant) features set, as well as a set selected by Fisher supervised algorithm.

From a machine learning point of view, we are now interested in selection methodologies combining simultaneously feature and data selection approaches. In fact our problem is

symmetric (towards examples and features) and the main problem is still how to choose the best representative data (among all available data or inside estimated clusters).

From an applicative perspective, the relevant features are generally not a priori known (for example in the context of new sensors); hence the study of outliers identified using the stability measure will be of great help.

References

A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, dec 2001.

A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271, 1997. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5.

M. Campedel and E. Moulines. Classification et sélection automatique de caractéristiques de textures. *RNTI*, C-1:25–37, 2004.

M. Campedel, E. Moulines, H. Maître, and M. Datcu. Feature selection for satellite image indexing. In *ESA-EUSC: Image Information Mining*, Frascati (Italy), oct 2005.

M. Campedel, I. Kyrgyzov, and H. Maître. Sélection non supervisée d'attributs - application à l'indexation d'images satellitaires. In *SFC'07*, Paris, sep 2007.

E. Diday. *Optimisation en classification automatique. Tome 1, 2. (French) [Optimization in automatic classification. Vol. 1, 2].* Institut National de Recherche en Informatique et en Automatique (INRIA), Rocquencourt, 1979.

A.L.N. Fred and A.K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, Jun 2005.

I. Guyon and A. Elisseeff. An introduction to feature and variable selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

R. M. Haralick, K. Shanmugan, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, November 1973.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2): 273–324, 1997. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(97)00043-X.

L.I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley-Interscience, July 2004.

I. Kyrgyzov. *Mining Satellite Image database of landscapes and application to urban zone: clustering, consensus and categorisation.* PhD thesis, TELECOM ParisTech, 2008.

I.O. Kyrgyzov, H. Maître, and M. Campedel. A method of clustering combination applied to satellite image analysis. In *IEEE - International Conference on Image Analysis and Processing ICIAP 2007*, pages 81–86, Modena, Italy, sep 2007.

S.G. Mallat. *A Wavelet tour of signal processing.* Elsevier, second edition, 2003.

F. Marcotorchino and P. Michaud. Agrégation de similarités en classification automatique. *Rev. Stat. Appl.*, 30(2):21–44, 1982.

P. Michaud and F. Marcotorchino. Modeles d'optimisation en analyse des données rélationnelles. *Math. Sci. Hum.*, 67:7–38, 1979.

P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):301–312, 2002. ISSN 0162-8828. doi: http://dx.doi.org/10.1109/34.990133.

J. Ontrup and H. Ritter. Perceptual grouping in a neural model: Reproduced human texture perception. Technical Report SFB360-TR-98/6, Neuroinformatics group University of Bielefeld, Germany, 1998.

J. Weston, A. Elisseeff, G. Bakir, and Fabian Sinz. The spider for matlab - v1.4, 2004. URL `http://www.kyb.tuebingen.mpg.de/bs/people/spider`.