

ESFS: A new embedded feature selection method based on SFS

Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen

▶ **To cite this version:**

Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen. ESFS: A new embedded feature selection method based on SFS. [Research Report] Ecole Centrale Lyon; Université de Lyon; LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/École Centrale de Lyon; Tsinghua University, Bejing, China. 2008. hal-01984705

HAL Id: hal-01984705
https://hal.science/hal-01984705

Submitted on 17 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESFS: A new embedded feature selection method based on SFS

Zhongzhe Xiao[1], Emmanuel Dellandrea[1], Weibei Dou[2], Liming Chen[1]

[1]LIRIS Laboratory – Ecole Centrale de Lyon, 36 avenue Guy de Collongue, 69134 Ecully Cedex, France
{zhongzhe.xiao, emmanuel.dellandrea, liming.chen}@ec-lyon.fr
[2]Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.China
douwb@mail.tsinghua.edu.cn

Abstract. Feature subset selection is an important subject when training classifiers in Machine Learning (ML) problems. Too many input features in a ML problem may lead to the so-called "curse of dimensionality", which describes the fact that the complexity of the classifier parameters adjustment during training increases exponentially with the number of features. Thus, ML algorithms are known to suffer from important decrease of the prediction accuracy when faced with many features that are not necessary. In this paper, we introduce a novel embedded feature selection method, called ESFS, which is inspired from the wrapper method SFS since it relies on the simple principle to add incrementally most relevant features. Its originality concerns the use of mass functions from the evidence theory that allows to merge elegantly the information carried by features, in an embedded way, and so leading to a lower computational cost than original SFS. This approach has successfully been applied to the emergent domain of emotion classification in audio signals.

Keywords: feature selection, emotion classification, evidence theory, audio, speech, music.

1 Introduction

When a classification problem has to be solved, the common approach is to compute a wide variety of features that will carry as much as possible different information to perform the classification of samples. Thus, numerous features are used whereas, generally, only a few of them are relevant for the classification task. Including the other in the feature set used to represent the samples to classify, may lead to a slower execution of the classifier, less understandable results, and much reduced accuracy (Hal, 1997). In this context, the objective of feature selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and gaining a deeper insight into the underlying processes that generated the data.

Thus, a feature selection method aims at finding the most relevant features. There exist considerable works in the literature on the question. Interesting overviews include (Kohavi, 1997; Guyon, 2003). However, the relevance notion is not perfectly

defined and may depend on the feature selection method. One of these definitions (Blum, 1997) is to consider that a feature f is relevant if it is incremental useful to a learning algorithm L with respect to a feature subset S: the accuracy that L produces an hypothesis using the feature set $f \cup S$ is higher than the accuracy achieved only using S. In the case of classification problems, the accuracy can be the correct classification rate.

Feature selection methods can be categorized into three main categories according to the dependence to the classifiers: filter approaches, wrapper approaches and embedded approaches (Kojadinovic, 2000).

Filter methods include Relief method (Arauzo-Azofra, 2004), Focus algorithm (Almuallim, 1991), and normally evaluate the statistical performance of the features over the data without considering the proper classifiers. The irrelevant features are filtered out before the classification process (Hal, 1997). Their main advantage is their low computational complexity which makes them very fast. Their main drawback is that they are not optimized to be used with a particular classifier as they are completely independent of the classification stage.

Wrapper methods on the contrary evaluate feature subsets with the classification algorithm in order to measure their efficiency according to the correct classification rate (Kohavi, 1997). Thus, feature subsets are generated thanks to some search strategy, and the feature subset which leads to the best correct classification rate is kept. Among algorithms widely used, we can mention Genetic Algorithm (GA) and Sequential Forward Selection (SFS) methods. The computational complexity is higher than the one of filter methods but selected subsets are generally more efficient, even if they remain sub-optimal (Spence, 1998).

In embedded feature selection methods, similarly to wrapper methods, feature selection is linked to the classification stage, this link being in this case much stronger as the feature selection in embedded methods is included into the classifier construction. Recursive partitioning methods for decision trees such as ID3, C4.5 and CART are examples of such method. Embedded methods offer the same advantages as wrapper methods concerning the interaction between the feature selection and the classification. Moreover, they present a better computational complexity since the selection of features is directly included in the classifier construction during training process.

In our work, we introduce a new embedded feature selection method we have developed and called ESFS, inspired from the wrapper method SFS since it relies on the simple principle to add incrementally most relevant features, and making use of the term of mass function which is introduced from the evidence theory which allows elegantly to merge feature information in an embedded way, leading to a lower computational cost than original SFS.

This approach has been evaluated on the problem of emotion classification in audio signals. We consider two types of data: speech and music samples. As speech samples present different signal properties than music samples, two different feature sets are considered. The speech feature set includes 226 features, whereas the music feature set includes 188. The high number of features compared to the relatively low number of samples available for training classifiers (about 500 samples) suggests the use of a feature selection method to improve classification accuracy.

The reminder of this paper is organized as follows. In section 2, we introduce the evidence theory on which our feature selection method is based, and detailed in section 3. Experimental results are presented in section 4. Finally, conclusions and perspectives are drawn in section 5.

2 Overview of the evidence theory

In our feature selection scheme, the term "belief mass" from the evidence theory is introduced into the processing of features.

Dempster and Shafer wanted in the 1970's to calculate a general uncertainty level from the Bayesian theory. They developed the concept of "uncertainty mapping" to measure the uncertainty between a lower limit and an upper limit (Dempster, 1968). Similar to the probabilities in the Bayesian theory, they presented a combination rule of the belief masses (or mass function) *m()*.

The evidence theory was completed and presented by Shafer in (Shafer, 1976). It relies on the definition of a set of n hypothesis Ω which have to be exclusive and exhaustive. In this theory, the reasoning concerns the frame of discernment 2^{Ω} which is the set composed of the 2^n subsets of Ω. In order to express the degree of confidence we have in a source of information for an event A of 2^{Ω}, we associate to it an elementary mass of evidence $m(A)$.

The elementary mass function or belief mass which presents the chance of being a true statement is defined as:

$m: 2^{\Omega} \rightarrow [0,1],1]$ which satisfies: $$m(\Phi) = 0 \text{ and } \sum_{A \subseteq 2^{\Omega}} m(A) = 1$$

The belief function is defined if it satisfies $Bel(\Phi)=0$ and $Bel(\Omega)=1$ and for any collection $A_1 \ldots A_n$ of subsets of Ω

$$Bel(A_1 \cup \ldots \cup A_n) \geq \sum_{\substack{I \subseteq \{1..n\} \\ I \neq \Phi}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right)$$

The belief function shows the lower bound on the chances, and it corresponds to the mass function with the following formulaes

$$Bel(A) = \sum_{B \subseteq A} m(B) \;\; \forall\, A \subset \Omega \qquad\qquad m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} Bel(B)$$

where $|X|$ means the number of elements in the subset X.

The doubt function is defined as $Dou(A) = Bel(\bar{A})$

and the upper probability function is defined as $Pl(A) = 1 - Dou(A)$

The true belief in A should be between $Bel(A)$ and $Pl(A)$.

The Dempster's combination rule can combine two or more independent sets of mass assignments by using orthogonal sum. For the case of two mass functions, let m_1 and m_2 be mass functions on the same frame Ω, the orthogonal sum is defined as $m=m1 \oplus m2$, to be $m(\Phi)=0$, and

$$m(A) = K \sum_{X \cap Y = A} m_1(X).m_2(Y) \qquad \text{with} \quad K = \frac{1}{1 - \sum_{X \cap Y = \Phi} m_1(X).m_2(Y)}$$

For the case with more than two mass functions, let $m=m1 \oplus \ldots \oplus m2$. It satisfies $m(\Phi)=0$ and

$$m(A) = K \sum_{\bigcap_{A_i} = A} \prod_{1 \le i \le n} m_i(A_i) \qquad \text{with} \quad K = \frac{1}{1 - \sum_{\bigcap_{A_i} = \Phi} \prod_{1 \le i \le n} m_i(A_i)}$$

This definition of mass functions from the evidence is used in our model in order to represent the source of information given by each feature, and to combine them easily and to consider the as a classifier whose recognition value is given by the mass function.

3 ESFS scheme

Recall that an exhaustive search of the best subset of features, leading to explore a space of 2^n subsets, is impractical, we turn to a heuristic approach for the feature selection. The SFS is selected as the basic of our feature selection. For this classifier dependent sub-optimal selection method, we have provided in this work two innovations. First, the range of subsets to be evaluated in the forward process is extended to multiple subsets for each size, and the feature set is reduced according to a certain threshold before the selection in order to decrease the computational burden caused by the extension of the subsets in the evaluation. Second, since the SFS is a classifier dependent method, the concept of belief masses which comes from the evidence theory is introduced to consider the audio feature as a classifier which leads to an embedded feature selection method.

3.1 Method overview

Heuristic feature selection algorithm can be characterized by its stance on four basic issues that determine the nature of the heuristic search process. First, one must determine the starting point in the space of feature subsets, which influences the direction of search and operators used to generate successor states. Second decision involves the organization of the search. As an exhaustive search in a space of 2^n feature subsets is impractical, one needs to rely on a more realistic approach such as greedy methods to traverse the space. At each point of the search, one considers local changes to the current state of the features, selects one and iterates. The third issue concerns the strategy used to evaluate alternative subsets of features. Finally, one must decide on some criterion for halting the search. In the following, we bring our answers to the previous four questions.

The SFS algorithm begins with an empty subset of features. The new subset S_k with k features is obtained by adding a single new feature to the subset S_{k-1} which performs the best among the subsets with $k-1$ features. The correct classification rate achieved by the selected feature subset is used as the selection criterion. In the

original algorithm of SFS, there are totally $n*(n+1)/2$ subsets which need to be evaluated and the optimal subset may be missing in the searching.

In order to avoid departure too far from the optimal performance, we proposed an improvement of the original SFS method by extending the subsets to be evaluated. In each step of forward selection, instead of keeping only one subset for each size of subsets, a threshold is set according to the compromise between the performance and the computational burden (which is decided from the performance from experiments with a small amount of data in our work) and all the subsets with the performance above the threshold are kept to enter the evaluation in the next step. Since remaining multiple subsets in each step may lead to heavy computational burden, only the features selected in the first step (subsets with single feature), thus having the best abilities to discriminate among classes that occur in the training data, are used in the evaluation in posterior steps.

As the features are added to the potential subsets one by one in the SFS process, the forward process of creating a feature subset with size k can be seen as a combination between two elements: a subset with size $k-1$ and a single feature. Thus, if we consider each subset as a feature itself, the process of creating a new feature subset can be interpreted as generating a new feature from two features.

A wrapper feature selection scheme such as the SFS needs to specify a classifier in order to evaluate improvement of classification accuracy as feature selection criterion. In our case, the classifier used in this feature selection method is simply based on the belief masses of the features which are modeled from the distribution of the features for each class obtained from the training data. The belief masses of samples in the testing sets are calculated with the model of the belief masses. The class with the highest belief mass is then taken as the output of the classification. This classifier is repeated for every subset in evaluation for searching the best feature subset. The procedure is detailed in the next subsection.

3.2 Feature selection procedure

The feature selection procedure is introduced in this section with its four steps.

Step 1: Calculation of the belief masses of the single features.

Before the feature selection starts, all features are normalized into [0, 1]. For each feature,

$$Fea_n = \frac{Fea_{n0} - \min(Fea_{n0})}{\max(Fea_{n0}) - \min(Fea_{n0})}$$

where Fea_{n0} is the set of original value of the n^{th} feature, and Fea_n is the normalized value of the n^{th} feature.

By definition of the belief masses, the mass can be obtained by different ways which can represent the chance for a statement to be true. In this paper, the PDFs (probability density functions) of the features computed from the training data are used to represent the masses of the single features.

The curves of PDFs of the features are obtained by applying polynomial interpolation to the statistics of the distribution of the feature values from the training data.

Taking the case of a 2-class classifier as example, the classes are defined as subset A and subset A^C. First, the probability densities of the features in each of the 2 subsets are estimated from the training samples by the statistics of the values of the features in each class. We define the probability density of the k^{th} feature Fea_k in subset A as $Pr_k(A, f_k)$ and the probability density in subset A^C as $Pr_k(A^C, f_k)$, where the f_k is the value of the feature Fea_k .According to the probability densities, the masses of feature Fea_k on these two subsets can be defined as

$$m_k(A, f_k) = \frac{Pr_k(A, f_k)}{Pr_k(A, f_k) + Pr_k(A^c, f_k)} \qquad m_k(A^c, f_k) = \frac{Pr_k(A^c, f_k)}{Pr_k(A, f_k) + Pr_k(A^c, f_k)}$$

where at any possible value of the k^{th} feature f_k, $m_k(A, f_k)+m_k(A^C, f_k)=1$.

In the case of N classes, the classes are defined as $A_1, A_2,\ldots,A_N$. The masses of feature F_k of the i^{th} class A_i can be obtained as

$$m_k(A_i, f_k) = \frac{Pr_k(A_i, f_k)}{\sum_{n=1}^{N} Pr_k(A_n, f_k)}$$ which satisfies $$\sum_{i=1}^{N} m_k(A_i, f_k) = 1$$

Step 2: Evaluation of the single features and selection of the initial set of potential features.

When the distribution model of the belief masses of the single features for the different classes have been extracted from the training data, the single features are evaluated by passing the distribution model derived from the training data. For each sample, its belief mass value can be extracted from feature mass functions. The samples are assigned to the class which has the highest belief mass and thus performances of correct classification rates can be obtained.

Within this process, the single features can then be ordered according to the correct classification rate given by mass functions and thus the best features can be selected.

The features are ordered in descending order according to the correct classification rates $R_{single}(F_k)$ as $\{F_{s1}, F_{s2},\ldots, F_{sN}\}$, where N means the total number of features in the whole feature set.

In order to reduce the computational burden in the feature selection, an initial feature set FS_{ini} is constructed with the best K features in the re-ordered feature set according to a certain threshold in classification rates as $FS_{ini}=\{F_{s1}, F_{s2},\ldots, F_{sK}\}$.

The threshold of the classification rates is decided according to the best classification rate as:

$R_{single}(F_{s_K}) \geq thres_1 * R_{best_1}$

where $R_{best_1}=R_{single}(F_{s_1})$. In our work on emotion analysis, the threshold value $thres_1$ is set to 0.8 according to a balance between the overall performance and the calculation time by experiments. This threshold may vary with different problems, and around 30 features are kept in our applications above the threshold of 0.8.

Only the features selected in the set FS_{ini} will attend in the latter steps of feature selection process. The elements (features) in FS_{ini} are seen as subsets with size 1 at the same time.

Step 3: Combination of features for the generation of the feature subsets.

For the iterations with subsets with size k ($k\geq 2$), the generation of a subset is converted into the creation of a new feature by using an operator of combination from

two original features, and the subsets are selected according to a threshold similar to the case with single features for each size of subsets.

We note the set of all the feature subsets in the evaluation with size k as FS_k and the set of the selected subsets with size k as FS'_k. Thus, FS_1 equals to the original whole feature set, and $FS'_1=FS_{ini}$. From $k=2$, the set of the feature subsets FS_k is noted as: $FS_k=Combine(FS'_{k-1}, FS_{ini})=\{Fc0_{1_k}, Fc0_{2_k},\ldots, Fc0_{Nk_k}\}$

where the function "*Combine*" means to generate new features by combining features from each of the two sets FS'_{k-1} and FS_{ini} with all the possible combinations except the case in which the element from FS_{ini} appears in the original features during the generation process of the element from FS'_{k-1}; $Fc0_{n_k}$ represents the generated new features; and Nk is the number of elements in the set FS_k.

The creation of a new feature from two features is implemented by combining the contribution of the belief masses of the two features, making use of an operator of combination. The combining process works as follows.

Assume that N classes are considered in the classifier. For the i^{th} class A_i, the pre-processed mass m^* for the new feature $Fc0_{t_k}$, which is generated with Fc_{x_k-1} from FS'_{k-1} and Fs_y from FS_{ini}, $Fc0_{t_k}= Combine\ (Fc_{x_k-1}, Fs_y)$, is calculated as

$$m^*(A_i, fc0_{t_k}) = T(m(A_i, fc_{x_k-1}), m(A_i, fs_y))$$

where f_x is the value of the feature F_x, and $T(x,y)$ is an operator of combination that corresponds to a t-norm operator, being a generalization of the conjunctive 'AND' (Schweizer, 1983). The sum of m^*s may not be 1 according to different operators. In order to meet the definition of belief masses, the m^*s can then be normalized as the masses for the new feature:

$$m(A_i, fc0_{t_k}) = \frac{m^*(A_i, fc0_{t_k})}{\sum_{n=1}^{N} m^*(A_n, fc0_{t_k})}$$

The performance of the combined new feature may be better than both two features in the combination. However, the combined new feature may even performance worse than any of the two original features, which will be eliminated in the selection.

The correct classification rates of the combined new features can be obtained with the belief masses by assigning the class with the highest belief mass to the data samples, and the combined new features can then be ordered in descending order according to the correct classification rates as with the single features:

$FS_k=\{Fc0_{1_k}, Fc0_{2_k},\ldots, Fc0_{Nk_k}\}=\{Fc_{1_k}, Fc_{2_k},\ldots, Fc_{Nk_k}\}$

The best feature with size k is noted as $Fc_{best_k}=Fc_{1_k}$, and the recognition rate of feature Fc_{best_k} is recorded as R_{best_k}.

Similar to the selection of FS_{ini} in the evaluation of the single features, a threshold is set to select a certain number of subsets with size k to take part to the next step of forward selection. The set of the subsets remained is noted as

$FS'_k=\{Fc_{1_k}, Fc_{2_k},\ldots, Fc_{N0k_k}\}$

which satisfies $R(Fc_{N0k_k})\geq thres_k * R_{best_k}$. In order to simplify the selection, the threshold value $thres_k$ is set in our work to the same value as 0.8 in every step without any adaptation to each step.

Step 4: Stop criterion and the selection of the best feature subset.

The stop criterion of ESFS occurs when the best classification rate begins to decrease while increasing the size of the feature subsets. In order to avoid missing the

real peak of the classification performance, the forward selection stops when the classification performance continues to decrease in two steps, $R_{best_k} < \min(R_{best_k-1}, R_{best_k-2})$.

4 Experimental Results

The feature selection method proposed in previous section has been evaluated on the problem of emotion classification in speech and music.

4.1 Datasets

Our experiments are performed on two datasets, presented below.

The Berlin emotional speech database is developed by Professor Sendlmeier and his fellows in Department of Communication Science, Institute for Speech and Communication, Berlin Technical University (Sendlmeier). This database contains 535 speech samples (302 from female voices and 233 from male voices) belonging to 7 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness and neutral.

As there is no public music emotion dataset available, we have built a dataset for music emotion recognition. It contains 603 samples of classical music labeled according to four emotions: exuberance, anxious, contentment and depression.

4.2 Feature extraction

A total number of 226 features have been computed to represent each speech sample from Berlin dataset. The corresponding feature set thus includes harmonic features, frequency features, energy features, MFCC features and Zipf features. As speech signal present different signal properties than music, a second set of features has been computed to represent each music sample. This feature set is composed of 188 features, including rhythmic features, tonality features, timbre features and octave-based features.

4.3 Results

Three groups of experiments are made with different features on Berlin dataset: one with all the features without selection, the second with features selected with fisher filter method (Narendran, 1977), and the third with the best features selected by the ESFS.

Five types of one step global classifiers are tested: Multi-layer Perception (Neural Network, marked as MP in the following text), C4.5, Linear Discriminant Analysis (LDA), K-NN, and Naive Bayes (NB). Each classifier is tested with several parameter configurations, and only the best results are kept. The experiments are carried out on TANAGRA platform (Rakotomalala, 2005) with 10-folds cross-validation. The experimental results are listed in *Table 1*.

The features selected by the embedded method ESFS are actually working in a filter way on the several classifiers in this experiment. The result show that for most of the classifiers tested in this experiment, the features selected by ESFS work better than the features selected by fisher filtering criterion. Especially, the features selected by ESFS fit the LDA very well, and classification rate on the LDA with these features is even better than the result from the ESFS itself on female voice samples. This result shows that the ESFS method is able to select the most discriminative features on the

problem of classification of emotional speech, and the features selected with this method are more suitable to be used in the linear classifier methods than the non-linear ones.

Table 1. Comparison between the result without feature selection and with the features selected by ESFS on Berlin dataset.

	FEMALE			MALE		
	NO SELECTION	FILTER SELECTION	ESFS SELECTION	NO SELECTION	FILTER SELECTION	ESFS SELECTION
MP	**65.73**±2.85	66.38±2.73	71.03±1.39	**61.78**±2.93	65.75±3.19	66.44±2.50
C4.5	55.46±2.7	56.22±2.95	55.73±3.38	55.75±0.66	54.66±2.32	56.51±3.53
LDA	60.92±2.56	**70.16**±3.14	**74.00**±2.08	51.16±3.05	**70.62**±2.37	**71.97**±1.57
K-NN	60.14±2.37	64.16±3.44	67.41±1.42	57.88±2.85	61.51±2.23	66.23±2.31
NB	62.67±1.45	59.78±1.10	67.41±1.46	56.30±1.12	57.60±0.59	62.81±2.79
Best	65.73	70.16	74.00	61.78	70.62	71.97
ESFS		71.75%±3.10%			73.77%±2.33%	

We also made experiments on the problem of classification of music emotion with four classes (*Table 2*). In order to test the ESFS itself without the effects of the structure of the classifiers, global classifiers with one step in the classification of the four classes are applied. The same classifiers on the TANAGRA platform – MP, C4.5, LDA, K-NN, and NB – as used on the Berlin dataset are also tested on the music emotion dataset. The result of ESFS on the problem of classification of music emotion with global classifier is 72.80%, which is 2% higher than that obtained from the experiments on TANAGRA as 70.80% with Naïve Bayes. Although the superiority of the result of the ESFS is not so obvious, the ESFS still shows a better performance than the popular used classification schemes, and with lower computational complexity because the feature selection and the classification processes are implemented simultaneously.

Table 2. Comparison of classification accuracy between ESFS and other classifiers (%).

ESFS	MP	C4.5	LDA	K-NN	NB
72.80	69.14±2.2	57.33±2.6	60.86±3.1	68.99±2.0	70.80±1.7

5 Conclusion and Future Work

In this paper, we have presented a novel feature selection method, ESFS, which relies on the simple principle to add incrementally most relevant features. To this purpose, we represent each feature thanks to mass functions, from the evidence theory, which allows to merge the information carried by features, in an embedded way, and so leading to a lower computational cost than wrapper method. Indeed, our

ESFS scheme allows simultaneously to select most relevant features and to perform classification, with no need of an extra classifier.

Experimental results on the problems of emotion classification in speech and music have shown that selecting relevant features improves the classification accuracy, and for this purpose, ESFS, used as a filter selection method, performs better than the traditional filter method, namely Fisher algorithm. Moreover, ESFS, when used as both feature selector and classifier, allows to obtain a better classification accuracy than representative state of the art classifiers, such as neural networks, or decision trees.

We envisage in our future work to use ESFS as the basis of a hierarchical classifier, which will be represented by a binary classification tree where ESFS will be nodes. The purpose of this hierarchical structure is to allow to better separate classes by first separating classes far away from each other and then concentrating on closer classes. Moreover, thanks to ESFS, each subclassifier could have at its disposal its own feature set.

References

Hall, M. A., Smith, L. A. (1997), Feature Subset Selection: A Correlation Based Filter Approach, *International Conference on Neural Information Processing and Intelligent Information Systems, Springer*, p855-858.

Kohavi, R., John, G. H. (1997), Wrappers for Feature Subset Selection, *Artificial Intelligence, Volume 97, Issue 1-2, Special issue on relevance,* p273 – 324.

Guyon, I., Elisseff, A. (2003), An introduction to variable and feature selection, *Journal of Machine Learning Research 3*, p1157 – 1182.

Blum, A. and Langley, P. (1997) Selection of relevant features and examples in machine learning, *Artificial Intelligence Journal*, 245-271.

Kojadinovic, I., Wottka, T.(2000), Comparison between a filter and a wrapper approach to variable subset selection in regression problems*, ESIT 2000*, September 14-15, Aachen, Germany.

Arauzo-Azofra, A., Benitez, J. M., Castro, J. L. (2004), A feature set measure based on Relief, *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, p 104 – 109.

Almuallim, H., Dietterich, T. G. (1991), Learning with many irrelevant features*, Proceedings of the Ninth National Conference on Artificial Intelligence, p 547 – 552, San Jose, CA: AAAI Press*.

Spence, C., Sajda, P. (1998), The role of feature selection in building pattern recognizers for computer-aided diagnosis, *Proceedings of SPIE -- Volume 3338, Medical Imaging 1998: Image Processing, Kenneth M. Hanson, Editor*, pp. 1434-1441.

Dempster, A.P. (1968), A generalization of Bayesian inference. *J. Royal Statistical Soc. Series B,* vol. 30, 1968.

Shafer, G.(1976), A mathematical theory of evidence, *Princeton University Press*.

Narendra, P.M., Fukunaga, K. (1977), A branch and bound algorithm for feature selection, *IEEE Transactions on Computers, C-26(9):* 917-922.

Rakotomalala, R. (2005), TANAGRA : un logiciel gratuit pour l'enseignement et la recherche, *in Actes de EGC'2005, RNTI-E-3*, vol. 2, pp.697-702.

Schweizer B. and Sklar A. (1983), Probabilistic metric spaces, *North Holland, New York*.

Sendlmeier et al., Berlin emotional speech database, available online at http://www.expressive-speech.net/.