Feature selection method using preferences aggregation

Gaëlle Legrand and Nicolas Nicoloyannis

Laboratoire ERIC
Université Lumière Lyon 2
Bât. L-5, av. Pierre Mendès-France
69676 Bron Cedex – France
glegrand@eric.univ-lyon2.fr ; nicolas.nicoloyannis@univ-lyon2.fr

Abstract. The feature selection allows to choose P features among M $(P<M)$ and thus to reduce the representation space of data. This process is increasingly useful because of the databases size increase. Therefore we propose a method based on preferences aggregation. It is an hybrid method between filter and wrapper approaches.

1. Introduction

Due to increasing size of databases, the improvement of data representation quality becomes a main problem in data mining. One of the major difficulties related to data representation quality is data dimension. This problem is linked with the number of exogenous features characterizing each object. Users who want to cover all existing aspects of an endogenous feature and to obtain comprehensible knowledge define a great number of exogenous features. However, among these features, some will be irrelevant, useless and/or redundant. Indeed, it is often difficult or even impossible to distinguish the relevant features from the irrelevant ones.

The problem of data dimension can be summarized by ”Less is more” from Liu and Motoda [21] which means that if we wish to extract useful and comprehensible information from our data, it is initially appropriate to delete irrelevant parts. Feature selection solves this problem. It chooses an optimal features subset according to a particular criterion and reduces the features space by removing those which are irrelevant. Feature selection eliminates useless and redundant features, the learning process is then accelerated and the accuracy of learning algorithms may be improved. It also permits to reduce noise generated by some features. There are a lot of feature selection methods which are gathered in two approaches: the wrapper approach, [10], which use the learning algorithm to test all existing features subsets, and the filter approach, [12], which corresponds to a data pre-processing step preceding the learning phase. The fundamental difference between these two families lies in the fact that the first is related to the learning algorithm whereas the second is completely independent of it.

1.1. Wrapper methods

These methods [5] take the influence of the selected features subset on the performances of the learning algorithm into account. The learning algorithm is used as an evaluation function to test different features subsets. However, its computational cost is too important in most cases [17] : these methods generate all existing features subset.

1.2. Filter methods

Filter approaches are grouped into 5 categories : complete, heuristic, random, fast sequential selection and step by step.

Complete methods test all possible subsets of P features among M features with M the total number of features and P the number of selected features. We can quote MDLM [32] or FOCUS [1], [2] or PRESET [26]. MDLM performs a comparison of all existing features subsets. PRESET is an algorithm based on the rough sets theory. It selects a features subset, named reduction which involves the same consistency on the learning set as the initial features set. All features not belonging to this reduction are eliminated. FOCUS makes a complete search among all features subsets and selects the minimal subset which allows to determine the class of each object. The complexity of FOCUS is about $O\left(N^M\right)$, with N the number of objects and M the number of features. These 3 algorithms are impossible to apply in most of cases due to their very high computational cost.

Heuristic Methods have many representatives. We present only the principal ones. Relief, [13], is an iterative features weight-based algorithm inspired by instance-based learning algorithms. Relief knew many alternatives. The most interesting one is ReliefF, [15], which deals with multi-classes problems. The complexity of Relief and its alternatives is $O(IMN)$ where I is the number of iterations fixed by the user. The Branch and Bound methods, [27], use a selection criterion characterized by the monotonicity property : all subsets, for which the selection criterion is not higher than a threshold, are eliminated. ABB, [20], uses the same principle with inconsistency rate. Its complexity is $O(N2^M)$. The Khi2 Algorithm, [18] carries out simultaneously the features discretization and the elimination of irrelevant features. It is based on the χ^2 statistics. These methods require several accesses to databases.

Random methods main representative is LVF, [19]. LVF selects the smallest features subset generated randomly and which satisfies an inconsistency criterion. Its complexity is $O(IMN)$, with I the number of subset generation. Because of its probabilistic property, the number of selected features tends towards the half of the initial features number. Its complexity is about $O(IMN)$. Like previous methods, these methods require several accesses to databases.

Fast Sequential Selection Method are iterative feature selection methods with a single access to database. The selection process is thus a stepwise process : the first step selects the feature X_1 that is the more correlated with endogenous feature Y ; the second step selects the feature that is the more partially correlated with Y with fixed values for X_1, and so on... In order to have a single database scan, fast correlation measures must be used (such as Kendall rank correlation coefficient, or Pearson correlation coefficient, or modified Rand coefficient,...). This kind of method is represented by MIFS [3], CFS [8], and the method proposed by Lallich and Rakotomalala [16]. These methods are fastest and quite efficient. They appear like the most interesting.

Step-by-step methods use short-sighted criteria to select features. These methods do not take into account the interaction between features and classify features according to their discriminating capacity. This type of methods is effective and very rapid in particular on problems comprising at the same time many features and objects. Their complexity is $O(N \log N)$.

To sum up, wrapper approach and complete methods are inapplicable because of their computational cost and time complexity. Heuristic methods have difficulties with redundant features and, random methods are skewed towards a subset having a number of features about the half of the initial features number. Moreover, most of these methods require several scans of database which imply a high I/O cost. It consequently appears that fast sequential selection methods and step-by-step methods are the more attractive ones since they propose good results as well as very suitable computing cost.

We propose here a new feature selection algorithm. Our method does not belong to wrapper approach nor to filter approach. It is situated at the intersection of filter and wrapper approaches. It offers a reasonable processing time compared with pure wrapper methods. It uses preferences aggregation in the first stage to determine an ordered list of features subsets. The first stage is the filter part. The second and last stage is the wrapper part. The next section is devoted to the initial ideas. Third section deal with the feature selection method. Experimental evaluation is in section 4.

2. Starting point

We start from the following observation : step by step methods using short-sighted criteria such as Shannon entropy are fast, inexpensive and have good results. There are 4 categories of criteria which measure various features specifications :

- **Information measures:** these measures determine the information gain from a feature. The feature which has the greatest information gain, will be preferred to the other features. We can mention Shannon entropy [31], gain ratio [30], normalized gain [11].
- **Distance measures:** they evaluate the separability of classes. They are also know as separability, divergence, or discrimination measures : Euclidian distance measure, Mantaras distance measure [7], Gini coefficient [6].
- **Dependence measures** are all correlation or association measures. They qualify the ability to predict the value of one feature from the value of another. They can be used to find the correlation between a feature and a class. If the correlation of feature X_1 with a class is higher than the correlation of feature X_2 with the same

class, then feature X_1 is preferred to X_2. We can cite chi-squared, Tschuprow coefficient [9] and [25], and Cramer coefficient.

- **Consistency measures:** they use the Min-Features bias in selecting a features subset. The Min-Features bias prefers hypotheses definable over as few features as possible. Two objects are inconsistent if their modalities are identical and if they belong to two different classes. These measures detect redundant features. We can cite the τ of Zhou [33].

However, the use of a short-sighted method generates two problems:

- The choice of criterion is delicate: Which criterion is the most effective?
- The form of result (a list of sorted features) does not allow us to determine the optimal features subset.

The method we propose solves these two problems in the following way:

- There is no criterion better or more effective than others. Each criterion emphasizes some specific features qualities. It seems to be interesting to obtain a result which takes the opinion of different criteria into account. So to obtain this type of results, we use a method of preferences aggregation and several short-sighted criteria.
- Obtaining a sorted list of features limits the interest of the features selection method. Indeed, the question is : how can we determine the optimal size of a features subset? When we have a sorted list of features, one of the methods which seems to be effective to obtain an optimal subset is to use a wrapper approach which adds or removes iteratively elements of the sorted list. At each iteration, the learning algorithm tests if the addition or the suppression of a feature involves an improvement of error rate. However, this process is too expensive to be applied. For this reason, we parameterise the preferences aggregation method so that it doesn't provide an ordering on the features but a preordering. Also, we will not add features one by one but features subset by features subset.

3 Presentation of our method

Our feature selection method is at the intersection of filter and wrapper approaches, [35]. It is a Forward Selection method which makes feature classification possible with the use of short-sighted criteria. The result is a sorted list of disjoint features subsets. This method has 3 steps:

- Calculus and discretization of different criteria for each feature (filter approach),
- Application of preferences aggregation method on results obtained at the previous stage (filter approach),
- Research of the optimal features subset (wrapper approach).

3.1 Calculus and discretization of criteria

We let users choose the short-sighed criteria set. The only condition is : criteria must belong to each categories. For experiments, we select a set of 10 short-sighted criteria: Shannon entropy, gain ratio, normalized gain, Mantaras distance measure, Gini coefficient, chi-squared, Tschuprow coefficient, Cramer coefficient, τ of Zhou. Each criterion for all features are calculated in parallel. The result obtained is a set constituted of 10 ordered lists in the order descending of feature relevance.

A feature is as relevant as another one even if the two features do not bring the same information. Therefore, we introduce the concept of features equivalence. In order to define this concept, we consider a setoff objects $O=\{o_1,...,o_j,...,o_n\}$ described by a features set $X=\{x_1,...,x_i,...,x_p\}$ named initial features set. Given $CR=\{cr_1,...,cr_k,...,cr_{10}\}$ the set of 10 short-sighted criteria with $cr_k=\{cr_{k1},...,cr_{ki},...,cr_{kp}\}$, the set of the criterion k values for each feature of X. The cr_{ki} values of each criterion are normalized with the following transformation: for a feature $x_i \in X$ and a criterion $cr_k \in CR$, the normalized value of criterion is:

$$cr_{ki,N}=\frac{cr_{ki}-Min(\{cr_k\})}{Max(\{cr_k\})-Min(\{cr_k\})} \qquad \textbf{(1)}$$

After their normalization, these values are discretized in deciles. The discretization assigns to each feature $x_i \in X$ a rank for each criterion $cr_k \in CR$ as follows :

– For criteria which must be minimized :

If $cr_{ki,N} \in [0;0.1[$ then $R_{ki} = 1$; If $cr_{ki,N} \in [0.1;0.2[$ then $R_{ki} = 2$; ... ; If $cr_{ki,N} \in [0.9;1]$ then $R_{ki} = 10$

– For criteria which must be maximized :

If $cr_{ki,N} \in [0;0.1]$ then $R_{ki} = 10$; If $cr_{ki,N} \in [0.1;0.2[$ then $R_{ki} = 9$; ... ; If $cr_{ki,N} \in [0.9;1]$ then $R_{ki} = 1$

R_{ki} is the rank assigned to feature $x_i \in X$ for criterion $cr_k \in CR$. The most relevant feature has the smallest rank. Thus the equivalence concept is defined as follows : two features x_i and x_j are equivalents according to a criterion k if and only if for this criterion, they have the same rank :

$$\left(x_i \Leftrightarrow x_j\right) \Leftrightarrow R_{ki} = R_{kj} . \quad \textbf{(2)}$$

3.2 Aggregation of criteria results

For all preferences aggregation methods [23], it is appropriate to define a set of judges and a set of objects. In our case, the objects are initial features and the judges are criteria. We use the preferences aggregation method developed in [28] and [29] and based on [22] and [24]. We don't describe in detail this method but we present its subjacent principle.

For each objects pair $\left(x_i, x_j\right)$, each judge states its opinion $A_k\left(i, j\right)$. A_k, the opinion of judge k is an application of $X \times X$ in $\left\{\Pr ef, N \Pr ef, EQ\right\}$.

Thus,

$A_k\left(i, j\right) = \Pr ef \Leftrightarrow$ judge k prefers x_i to $x_j \Leftrightarrow R_{ki} < R_{kj}$,

$A_k\left(i, j\right) = N \Pr ef \Leftrightarrow$ judge k prefers x_i to $x_j \Leftrightarrow R_{ki} > R_{kj}$,

$A_k\left(i, j\right) = EQ \Leftrightarrow$ judge k considers x_i and x_j like equivalents $\Leftrightarrow R_{ki} = R_{kj}$.

The result we wish to obtain is an opinion OP called opinion of broad preferences and which generates a preordering relation on X. OP is an application of $X \times X$ in $\left\{\Pr ef, N \Pr ef, EQ\right\}$.

Definition 1 : The degree of agreement $\rho_{ij}\left(OP, A_k\right)$ between the advices $OP\left(i, j\right)$ and $A_k\left(i, j\right)$ is defined in table 1.

Table 1. Degree of agreement ρ_{ij}

OP / A_k	$\Pr ef$	$N \Pr ef$	EQ
$\Pr ef$	1	0	1/2
$N \Pr ef$	0	1	1/2
EQ	1/2	1/2	1

Definition 2 : The degree of agreement $DA\left(OP, A_k\right)$ between the opinions OP and A_k is $DA\left(OP, A_k\right) = \sum_{\left(x_i, x_j\right) \in X} \rho_{ij}\left(OP, A_k\right)$.

Definition 3 : The degree of agreement between the opinion OP and the opinion of all judges is $DA\left(OP\right) = \sum_{k=1}^{10} DA\left(OP, A_k\right)$.

Our problem consists in building an opinion OP which generates a preordering on X and which maximizes $DA\left(OP\right)$. The corresponding optimization problem is NP-hard, hence the use of a meta-heuristic. Simulated annealing method [14] is used for maximization. We choose simulated annealing because it's a rapid and easy to use method. The parameters are : the decay rate equal 0.98, the halting condition is a number of iterations which equal $10*|X|$. The neighbourhood of the current solution is defined as follow : a preordering $L'=\{l_1',...,l_h',...,l_H'\}$ is neighbour of a preordering L, $L'\subset V(L)$, if and only if L' derive from L by the movement of only one object. After the application of this aggregation method, we obtain an ordered list of disjoint features subsets $L=\{l_1,...,l_h,...,l_H\}$.

3.3 Optimal features subset

Until now, our method has a filter approach. At this stage, our method has a wrapper approach. The advantage of using a wrapper approach is the use of the influence of the features subset on learning algorithm performances. Detection of the optimal subset is carried out as follows : within the h^{th} iteration, the features subset $l_h \in L$ is added to the optimal features subset. The optimal features subset is the one having the smallest error rate on the learning set.

4 Experimentations.

In our experiments we used 14 databases from the UCI collection [4]. The quantitative features are discretized with Fusinter method, [14]. The features selection is carried out on 30% of the initial set of objects while keeping the initial distribution of classes. Experimentations with MIFS and ReliefF are also carried out on these same 30%. The 70% remainder are used for the learning stage. For that, we choose a 10-fold-cross-validation and learning algorithms are ID3, Sipina and Naïve Bayesian (NB). Tests before selection are also carried out on these same 70%. Tables 2, 3 and 4 show error rate and the associated Standard deviation (Sd) obtained before and after features selection respectively with ID3, Naïve Bayesian and Sipina by using our method. The results obtained with ID3 and BN are interesting. Except for some bases, we can see an error rate reduction and/or a stabilization of the results (Sd reduction). For Sipina, the results before and after selection are practically identical and sometimes there is an error rate degradation. For Cleve, Heart and German with Sipina, we can observe an important increase of the error rate. Tables 5, 6 and 7 indicate the number of selected features respectively with ID3, Sipina and Naive Bayesian. Our results are between those of ReliefF and those of MIFS. Tables 8, 9 and 10 allow us compare our method with ReliefF and MIFS. Results are sometimes equivalent. Our method obtain better results in most of case. Table 11 shows the number of iterations carried out by our method. The maximum number of iterations is about 9 (for Vehicle). The number of learning algorithm call in our method is then smaller than in pure wrapper methods.

Table 2. Tests with ID3

Bases	Without selection		With selection	
	Error rate	Sd	Error rate	Sd
Austra	16.60	4.57	15.29	3.48
Breast	5.95	1.95	4.27	2.8
Cleve	18.53	8.68	21.9	8.67
CRX	14.73	5.68	15.7	3.1
German	31.86	7.53	26.14	4.87
Heart	27.05	10.29	26.32	11.04
Iono	21.37	8.39	11.73	5.59
Iris	3.73	4.57	4.73	4.74
Monks-1	25.22	8.3	25.18	7.56
Monks-2	34.91	6.79	34.89	6.71
Monks-3	1.28	1.28	3.88	2.69
Pima	26.11	5.43	24.5	5.15
Tic Tac Toe	33.43	5	25.16	6.31
Vehicle	34.24	4.96	28.75	5.44

Table 3. Tests with BN

Bases	Without selection		With selection	
	Error rate	Sd	Error rate	Sd
Austra	14.26	4.58	15.27	3.61
Breast	2.65	1.31	2.65	2.05
Cleve	21	6.63	17.77	6.14
CRX	14.67	3.14	15.69	3.99
German	23.71	6.58	23.43	4.62
Heart	17.37	7.46	17.89	7.14
Iono	6.83	5.06	7.25	5.88
Iris	6.45	7.14	2.82	4.31
Monks-1	25.22	6	25.19	4.68
Monks-2	38.94	4.14	34.92	5.11
Monks-3	3.88	2.9	3.85	3.67
Pima	21.14	5.42	22.83	5.73
Tic Tac Toe	29.61	5.15	27.83	3.92
Vehicle	34.27	5.52	33.95	4.18

Table 4. Tests with Sipina

Bases	Without selection		With selection	
	Error rate	Sd	Error rate	Sd
Austra	16.73	3.95	15.28	6.02
Breast	7.13	2.29	6.73	4.84
Cleve	21.47	8.57	31.67	10.87
CRX	16.3	6.22	17.13	6.05
German	28.14	5.5	31.71	4.51
Heart	23.16	10.04	27.89	7.82
Iono	7.73	6.95	6.88	2.58
Iris	4.64	6.17	4.64	6.17
Monks-1	20.11	4.89	25.18	3.72
Monks-2	38.24	7	34.89	8.79
Monks-3	1.79	2.58	3.87	3.09
Pima	24.3	4.46	25.05	4.36
Tic Tac Toe	20.67	3.77	26.06	7.5
Vehicle	47.26	6.24	50.58	5.63

Table 5. Number of selected features with ID3

Bases	Without selection	Our method	ReliefF	MIFS
Austra	14	1	2	13
Breast	9	3	6	9
Cleve	13	7	6	8
CRX	15	3	2	7
German	20	5	14	3
Heart	13	2	2	13
Iono	34	2	25	8
Iris	4	3	4	3
Monks-1	6	1	2	1
Monks-2	6	1	2	2
Monks-3	6	2	2	3
Pima	8	2	7	4
Tic Tac Toe	9	7	5	3
Vehicle	18	14	18	6

Table 6. Number of selected features with Naïve Bayesian

Bases	Without selection	Our method	ReliefF	MIFS
Austra	14	2	2	13
Breast	9	7	6	9
Cleve	13	5	6	8
CRX	15	5	2	7
German	20	9	14	3
Heart	13	8	2	13
Iono	34	26	25	8
Iris	4	2	4	3
Monks-1	6	1	2	1
Monks-2	6	1	2	2
Monks-3	6	2	2	3
Pima	8	5	7	4
Tic Tac Toe	9	7	5	3
Vehicle	18	12	18	6

Table 7. Number of selected features with Sipina

Bases	Without selection	Our method	ReliefF	MIFS
Austra	14	1	2	13
Breast	9	4	6	9
Cleve	13	1	6	8
CRX	15	3	2	7
German	20	1	14	3
Heart	13	2	2	13
Iono	34	26	25	8
Iris	4	2	4	3
Monks-1	6	1	2	1
Monks-2	6	1	2	2
Monks-3	6	2	2	3
Pima	8	1	7	4
Tic Tac Toe	9	3	5	3
Vehicle	18	10	18	6

Table 8. Tests with ReliefF and MIFS (ID3)

Bases	Our method		MIFS		ReliefF	
	Error rate	Sd	Error rate	Sd	Error rate	Sd
Austra	15.29	3.48	17.17	4.12	15.31	5.23
Breast	4.27	2.8	5.9	2.64	5.29	3.16
Cleve	21.9	8.67	24.68	10.27	40.54	7.77
CRX	15.7	3.1	16.12	6.7	17.54	5.88
German	26.14	4.87	27.43	5.06	30.14	6.01
Heart	26.32	11.04	28.42	9.76	27.38	9.06
Iono	11.73	5.59	15.75	8.71	11.78	3.94
Iris	4.73	4.74	4.82	6.58	3.73	4.57
Monks-1	25.18	7.56	25.20	7.71	55.52	3.34
Monks-2	34.89	6.71	34.91	6.7	34.9	8.63
Monks-3	3.88	2.69	3.86	2.86	3.88	3.34
Pima	24.5	5.15	24.87	4.83	25.05	7.69
Tic Tac Toe	25.16	6.31	30.81	7.11	30.51	5.9
Vehicle	28.75	5.44	40.62	7.39	42.25	6.52

Table 9. Tests with ReliefF and MIFS (Naïve Bayesian)

Bases	Our method		MIFS		ReliefF	
	Error rate	Sd	Error rate	Sd	Error rate	Sd
Austra	15.27	3.61	14.28	3.08	15.28	5.15
Breast	2.65	2.05	2.86	1.87	3.45	2.56
Cleve	17.77	6.14	20.52	11.34	40.67	4.33
CRX	15.69	3.99	14.66	5.7	16.53	2.8
German	23.43	4.62	26.29	3.63	30.71	4.96
Heart	17.89	7.14	17.89	10.04	21.05	10.53
Iono	7.25	5.88	5.22	4.4	9.32	6.22
Iris	2.82	4.31	4.64	6.17	6.45	7.14
Monks-1	25.19	4.68	25.20	7.18	51.9	8.2
Monks-2	34.92	5.11	34.92	6.24	34.92	6.65
Monks-3	3.85	3.67	3.86	2.87	3.85	3.85
Pima	22.83	5.73	21.33	4.3	25.04	3.41
Tic Tac Toe	27.83	3.92	28.87	5.42	27.97	4.19
Vehicle	33.95	4.18	39.85	8.01	45.82	8.78

Table 10. Tests with ReliefF and MIFS (Sipina)

Bases	Our method		MIFS		ReliefF	
	Error rate	Sd	Error rate	Sd	Error rate	Sd
Austra	15.28	6.02	16.35	6.65	15.28	5.25
Breast	6.73	4.84	7.13	2.29	5.9	3.8
Cleve	31.67	10.87	30.41	10.7	40.56	10.4
CRX	17.13	6.05	17.95	5.23	16.12	4.72
German	31.71	4.51	26.29	4.53	31	4.61
Heart	27.89	7.82	23.16	6.74	22.11	6.57
Iono	6.88	2.58	7.70	6.22	19.4	6.85
Iris	4.64	6.17	4.55	9.32	4.64	6.17
Monks-1	25.18	3.72	25.19	6.35	17.48	8.4
Monks-2	34.89	8.79	34.91	4.86	34.93	8.83
Monks-3	3.87	3.09	4.63	2.99	3.86	4.02
Pima	25.05	4.36	22.07	4.84	25.07	6.43
Tic Tac Toe	26.06	7.5	27.40	6.06	28.27	5.16
Vehicle	50.58	5.63	63.17	6.75	49.07	5.07

Table 11. Number of learning algorithm call with our method

Bases	Number of iterations	Number of iterations with BN	Number of iterations
Austra	2	3	2
Breast	3	5	5
Cleve	5	4	2
CRX	2	3	2
German	5	7	2
Heart	2	4	2
Iono	3	6	6
Iris	3	3	3
Monks-1	2	2	2
Monks-2	2	2	2
Monks-3	2	2	2
Pima	3	4	2
Tic Tac Toe	4	4	3
Vehicle	9	7	6

5 Conclusion.

In this article, we present a feature selection method based on preferences aggregation. It is a hybrid method between filter and wrapper approaches having the advantages of each approach and reducing theirs disadvantages :

- The influence of the selected features on the learning algorithm is taken into account. Thus, the selected features are different according to the used algorithm.
- The computational cost is largely lower than the computational cost of pure wrapper methods due to the use of a preordering.

Because of users can choose the short-sighted criteria set and the learning algorithm for wrapper stage, our method can be qualified "meta-method".

Concerning the number of selected features, ours results are comparable and even better with those obtained by ReliefF and MIFS. Concerning the accuracy, we can observe an error rate reduction after selection.

We plan to improve our method according to two aspects. The discretization method used for the criteria values must be more suitable. We would like, also, that the result of the method of preferences aggregation is not a list of features subsets, but the optimal features subset.

References

1. Almuallim H., and Dietterich T. G. : Learning with many irrelevant features. In Proceedings of the Ninth National Conference on Artificial Intelligence, 547- 552, Menlo Park, CA: AAAI Press, 1991.
2. Almuallim, H., and Dietterich, T. G. : Efficient algorithms for identifying relevant features. In Proc. of the 9th Canadian Conference on Artificial Intelligence, Vancouver, BC (1992) 38-45.
3. Battiti, R. : Using mutual information for selecting features in supervised neural net learning, IEEE Trans. on Neural Networks (1994) vol. 5, 537–550.
4. Blake, C.L. and Merz, C.J. : UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/, (1998).
5. Blum, A. L. and Langley, P. : Selection of relevant features and examples in machine learning. Artificial Intelligence, 245-271, 1997.
6. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. : Classification and Regression trees. The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA (1984).
7. De Mantaras, R.L. : A distance-based attribute selection measure for decision tree induction. Machine Learning, (1991) 6:81-92.
8. Hall, M. : Correlation-based feature selection of discrete and numeric class machine learning. In Proceedings of the International Conference on Machine Learning, pages 359-366, San Francisco, CA (2000). Morgan Kaufmann Publishers.
9. Hart, A. : Experience in the use of an inductive system in knowledge eng. In M. Bramer, editor, Research and Development in Expert Systems. Cambridge Univ. Press, Cambridge, MA, (1984).
10. John, G.H., Kohavi, R., Pfleger, K. : Irrelevant Features and the Subset Selection Problem, Proc. of the 11th International Conference on Machine Learning ICML94, (1994) 121-129.
11. Jun, B.H., Kim, C.S., Song, H.Y. and Kim, J. : A New Criterion in Selection and Discretization of Attributes for the Generation of Decision Trees , IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 19, n12, (1997) 1371--1375.
12. Kira, K., and Rendell, L. A. : The feature selection problem: Traditional methods and a new algorithm. In Tenth National Conference on Artificial Intelligence, (1992) 129-134. MIT Press.
13. Kira, K. and Rendell, L. : A practical approach to feature selection. In Proceedings of the Tenth International Conference on Machine Learning, Amherst, Massachusetts, (1992). Morgan Kaufmann.
14. Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. : Optimization by simulated annealing. Science, (1983) 220:671-680.
15. Kononenko, I. : Estimating attributes: analysis and extensions of Relief. In L. De Raedt and F. Bergadano, editors, Machine Learning: ECML94, 171-182. Springer Verlag, 1994.
16. Lallich, S., Rakotomalala, R. : Fast feature selection using partial correlation for multivalued attributes. Proceedings of the 4th European Conference on Knowledge Discovery in Databases, PKDD 2000, (2000) 221–231.
17. Langley, P., and Sage, S. : Oblivious decision trees and abstract cases. In Working Notes of the AAAI94 Workshop on Case-Based Reasoning, 1994. In press.
18. Liu, H. and Setiono, R. : Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE Int'l Conference on Tools with Artificial Intelligence, 1995.
19. Liu, H. and Setiono, R. : A probabilistic approach to feature selection-a filter solution. In Proceedingof Internationnal Conference on Machine Learning, (1996) 319-327.

20. Liu, H., Motoda H., and Dash M. : A monotonic measure for optimal feature selection. In Proceedings of European Conference on Machine Learning, (1998) 101-106.

21. Liu, H. and Motoda, H. : Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, 1998.

22. Marcotorchino, J.-F. and Michaud, P. : Heuristic approach to the similarity aggregation problem. Methods of Operations Research, (1981) 43 : 395-404.

23. Marcotorchino, J.-F. : Agrégation de similarités en classification automatique, thèse de Doctorat d'Etat, Université Paris 6, (1981).

24. Michaud, P. : Agrégation à la majorité 1 : Hommage à Condorcet, Rapport du Centre Scientifique IBM-France, N°F-051, (1982).

25. Mingers, J. : Expert systems --- rule induction with statistical data. Journal of the Operational Research Society, (1987) 38:39-47.

26. Modrzejewski, M. : Feature Selection Using Rough Sets Theory. (1993) 213-226 of: Proceedings of the European Conference on Machine Learning. Springer.

27. Narendra, P.M. and Fukunaga, K. : A Branch and Bound algorithm for feature subset selection. IEEE Transactions Computers, C-26:917, September 1977.

28. Nicoloyannis, N., Terrenoire, M. and Tounissoux, D. : An optimisation model for aggregationg preferences : A simulated annealing approach. Health and System Science, (1998) 2(1-2) :33-44.

29. Nicoloyannis, N., Terrenoire, M. and Tounissoux, D. : Pertinence d'une classification. Revue Electronique sur l'Apprentissage par les Données, (1999) 3(1) :39-49.

30. Quinlan, J. : Introduction of Decision Trees, Machine Learning, vol. 1, (1986) 81-106.

31. Shannon C.E.. A mathematical theory of communication. Bell System Technical Journal, 27:379--423,623--656, 1948.

32. Sheinvald, J., Dom, B. and Niblack, W. : A modelling approach to feature selection. In: Proceedings of Tenth International Conference on Pattern Recognition, (1990) 1:535--539.

33. Zhou, X. and Dillon, T.S. : A Statistical--Heuristic Feature Selection Criterion for Decision Tree Induction. IEEE Transactions on Pattern Analysis and Machine Intelligence, (1991) 13, 8:834-841.

34. Zighed, D. A., Rakotomalala, R., Rabaséda, S. : A discretization method of continous attributes in induction graphs, in Proc. Of the 13th European Meetings on Cybernetics and System Research, (1996) 997-1002.

35. Walid Erray, "WF : Une méthode de sélection de variables combinant une méthode filtre rapide et une approche enveloppe", 11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux, Septembre 2004.