Feature selection and preferences aggregation

Gaelle Legrand and Nicolas Nicoloyannis

Laboratoire ERIC
Université Lumière Lyon 2
Bat. L ; 5, av. Pierre Mendès-France
69676 Bron Cedex - France
(e-mail: 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. This process gets more and more useful because of the databases size increases. Therefore we propose a method based on preferences aggregation. It is an hybrid method that lies filter and wrapper approaches.
Keywords: Feature selection, wrapper approach, filter approach, preferences aggregation.

1 Introduction

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. The Knowledge Discovery in Databases (KDD) process can extract useful knowledge and patterns from the rapidly growing volumes of data to improve the performance of various classifiers and to reduce the running time. Feature selection is an essential step of the KDD process: it eliminates irrelevant, noisy and redundant features, it selects the most relevant features and it reduces the effective number of features under consideration, the data mining step is then accelerated and the calculative cost may be reduced (see [Sémani *et al.*, 2005]).
This paper addresses feature selection for supervised learning. We propose a new feature selection algorithm which is situated at the intersection of filter and wrapper approaches. It uses preferences aggregation to determine an ordered list of features subsets. The next section reviews existing feature selection methods. The third section presents our starting point. Section 4 presents our feature selection method. Experimental evaluations are presented in section 5.

2 Existing feature selection methods

Feature selection methods are gathered in two approaches: wrapper approach, [John *et al.*, 1994], and filter approach, [Kira and Rendell, 1992a]. Wrapper approach takes the influence of 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.

Filter approaches are grouped into 5 categories:
Complete methods test all possible features subsets. Their computational cost is very high: MDLM [Sheinvald *et al.*, 1990]...
Heuristic methods have many representatives like Relief, an iterative feature weight-based algorithm inspired by instance-based learning algorithms, (see [Kira and Rendell, 1992b]). These methods require several accesses to databases.
Random methods main representative is LVF, [Liu and Setiono, 1996]. Because of their probabilistic property, the number of selected features tends towards the half of the initial features number. Like previous methods, these methods require several accesses to databases.
Fast sequential selection method principle is an iterative feature selection with a single access to databases. In order to have a single data scan, fast correlation measures must be used such as Kendall rank correlation coefficient. This kind of methods is represented by MIFS [Battiti, 1994], or the method proposed by Lallich and Rakotomalala (see [Lallich and Rakotomalala, 2000]). These methods are the fastest and quite efficient.
Step-by-step methods use short-sighted criteria to select features. This type of methods is effective and very rapid particularly for problems with many features and objects.
Each approach is characterized by a search procedure to generate the next candidate subset (see [Langley, 1994]) and an evaluation criterion to evaluate the subset under consideration. There are 4 categories of criteria which measure various feature specifications: **Information measures**: these measures determine the information gain: Shannon entropy [Shannon, 1948], gain ratio [Quinlan, 1986],...; **Distance measures**: they evaluate the separability of classes: Gini coefficient [Breiman *et al.*, 1984], Mantaras distance measure [De Mantaras, 1991]...; **Dependence measures** are the whole correlation or association measures: Tschuprow coefficient [Hart, 1984]...; **Consistency measures**: These measures detect redundant features: τ of Zhou [Zhou and Dillon, 1991].

3 Starting point

We start from the following observation: step by step methods using short-sighted criteria are fast and have good results. However, the use of a step-by-step method generates two problems: The choice of criterion is delicate, which criterion is the most effective? and the form of result (a list of sorted features) doesn't provide us with the optimal features subset.

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

- There is no criterion better or more effective than others. Each criterion emphasizes some specific feature qualities. It seems to be interesting to

obtain a result which takes the opinion of different criteria into consideration. So to obtain this type of results, we use a set of criteria and a preferences aggregation method.
- Obtaining a sorted list of features limits the interest of feature selection : we parameterize the aggregation method so that it doesn't provide with an ordering on the features but a preordering. Also, we don't add features one by one but features subset by features subset.

4 Presentation of our method

Our feature selection method is at the intersection of filter and wrapper approaches which makes the features classification possible with the use of short-sighted criteria. This method has 3 steps:

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

4.1 Calculus and discretization of criteria

We let users choose the short sighted criteria set. The only condition is that there must be a representative of each criteria categories. For experiments and tests, we choose a set of 10 short-sighted criteria: Shannon entropy, gain ratio, normalized gain, Mantaras distance measure, Gini coefficient, chi-squared, Tschuprow coefficient, Cramer coefficient, and τ of Zhou.

Each criterion for all features are calculated parallely. The result is a set of 10 ordered lists (order descending) of feature relevance.

A feature can be as relevant as another one even if the two features don't bring the same information type. So, we introduce the concept of features equivalence.

In order to define this concept, we consider a set of objects $O = \{o_1, ..., o_n\}$ described by the initial features set $X = \{x_1, ..., x_i, ..., x_p\}$, and a set of K short-sighted criteria $CR = \{cr_1, ..., cr_k, ..., cr_K\}$ with $cr_k = \{cr_{k1}, ..., cr_{kp}\}$, the set of criterion values for each feature.

Values for each criterion are normalized with the following transformation: $cr_{ki,N} = (cr_{ki} - Min(cr_k)) \setminus (Max(cr_k) - Min(cr_k))$ for a feature x_i and a criterion cr_k.

After their normalization, these values are discretized in deciles. The discretization assigns to each feature a rank R_{ki} for each criterion. This rank is such that the most relevant feature has the smallest rank (For a criterion which must be minimized: If $cr_{ki,N} \in [0; 0.1]$ then $R_{ki} = 1$... If $cr_{ki,N} \in [0.9; 1]$ then $R_{ki} = 10$; For a criterion which must be maximized: If $cr_{ki,N} \in [0; 0.1]$ then $R_{ki} = 10$... If $cr_{ki,N} \in [0.9; 1]$ then $R_{ki} = 1$).

Thus the equivalence concept is defined as follows: two features are equivalent according to a criterion if and only if they have the same rank for this criterion. We tested various combination of normalization and discretization methods. The combination described here gave us the most interesting and general results on the tested datasets. It could be interesting to modify this combination according to data structure.

4.2 Aggregation of the results of criteria

For all aggregation methods (see [Vincke, 1982], [Tanguiane, 1991]), the set of judges and the set of objects must be defined. In our case, the objects correspond to features and the judges correspond to criteria.
We use the aggregation method developed in [Nicoloyannis *et al.*, 1998] and [Nicoloyannis *et al.*, 1999] based on pairwise comparison concept developed in [Marcotochino, 1984a] and [Marcotochino, 1984b]. We don't describe in details this method but we present its subjacent principle.

For each features pair (x_i, x_j), each judge (criterion) states its opinion $A_k(i,j)$. A_k, the opinion of a judge k is an application of $X \times X$ in $\{Pref, NPref, EQ\}$. Thus,
$A_k(i,j) = Pref$: the judge k prefers x_i to x_j, $R_{ki} < R_{kj}$
$A_k(i,j) = NPref$: the judge k prefers x_j to x_i, $R_{ki} > R_{kj}$
$A_k(i,j) = EQ$: the judge k considers x_j and x_i as equivalent, $R_{ki} = R_{kj}$.
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 $\{Pref, NPref, EQ\}$.
Definition 1: The degree of agreement $\rho_{ij}(OP, A_k)$ between $OP(i,j)$ and $A_k(i,j)$ is defined in Table 1.

OP/A_k	$Pref$	$NPref$	EQ
$Pref$	1	0	1/2
$NPref$	0	1	1/2
EQ	1/2	1/2	1

Table 1. Degree of agreement

Definition 2: The degree of agreement $DA(OP, A_k)$ is $DA(OP, A_k) = \sum \rho_{i,j}(OP, A_k)$.

Definition 3: The degree of agreement between the opinion OP and the opinion of all judges is $DA(OP) = \sum DA(OP, A_k)$.

Their problem consists in building an opinion OP which generates a preordering on X and which maximizes $DA(OP)$. The corresponding optimization problem is NP-hard, hence requires the use of a meta-heuristic. The simulated annealing method [Kirkpatrick *et al.*, 1983] is used for maximization. The simulated annealing method is used because it's a rapid and easy

to use method by [Nicoloyannis *et al.*, 1998]. But , they can use another methods. The parameters are : the decay rate is set to 0.98, the halting condition is a number of iterations which is set to $10 \times |X|$. The neighbourhood of the current solution is defined as follows: a preordering $\acute{L}$ belongs to the neighbourhood of a preordering $L = \{l_,, ..., l_m, ..., l_M\}$, $(\acute{L} \in V(L))$, if and only if $\acute{L}$ derives from L by the movement of only one object $x_i \in l_m, l_m \subset L$: x_i is flipped into l_{m+1}, $(m < M)$ or into l_{m-1}, $(m = M)$; Or x_i constitutes a group by itself.

After the application of this aggregation method, we obtain an ordered list of disjoint features subsets $L = \{l_1, ..., l_h, ..., l_H\}$.

4.3 Optimal features subset

Until this step, our method belong to filter approach. From this step, our method belong to wrapper approach. The advantage of using a wrapper approach is to take into consideration the influence of the features subset on the learning algorithm performances. The detection of the optimal subset is carried out as follows: within the h^{th} iteration, the features subset l_h is added to the optimal features subset. The optimal features subset is the one having the smallest error rate on the learning set.

5 Experimentations

For our experiments we used 11 databases from the UCI repository (see [Merz and Murphy, 1996]). Quantitative features are discretized with Fusinter method developed in [Zighed *et al.*, 1996]. The feature selection is carried out on 30% of the initial set of objects keeping initial classes distribution. The 70% remaining are used for the learning phase. For that, we use a 10-fold-cross-validation and the learning algorithms are ID3 and Naive Bayesian. The tests without selection are also carried out on these same 70% of studied base. After the application of our selection method, we can see some improvements in error rate with ID3 and the Naive Bayesian (Tables 2 and 3). Our method is comparable with MIFS and ReliefF and sometimes better. Tables 2 and 3 show the number of iteration carried out by our method. The maximum number of iterations is about 9 (for Vehicle). The number of learning algorithm runs in our method is then smaller than in pure wrapper methods. For our method, the number of selected features depends on the learning algorithm (Table 4). This number is often smaller than the number of features selected by MIFS et ReliefF.

6 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 their disadvantages:

Bases	Our method Error (Sd)	MIFS Error (Sd)	ReliefF Error (Sd)	Without selection Error (Sd)	Number of iterations with our method
Austra	15,29 (3,48)	17,17 (4,12)	15,31 (5,23)	16,6 (4,57)	2
Breast	4,27 (2,8)	5,9 (2,64)	5,29 (3,16)	5,95 (1,95)	3
Cleve	21,9 (8,67)	24,68 (10,27)	40,54 (7,77)	18,53 (8,68)	5
CRX	15,7 (3,1)	16,12 (6,7)	17,54 (5,88)	14,73 (5,68)	2
German	26,14 (4,87)	27,43 (5,06)	30,14 (6,01)	31,86 (7,53)	5
Heart	26,32 (11,04)	28,42 (9,76)	27,38 (9,06)	27,05 (10,29)	2
Iono	11,73 (5,59)	15,75 (8,71)	11,78 (3,94)	21,37 (8,39)	3
Iris	4,73 (4,74)	4,82 (6,58)	3,73 (4,57)	3,73 (4,57)	3
Monks-1	25,18 (7,56)	25,2 (7,71)	55,52 (3,34)	25,22 (8,3)	2
Monks-2	34,89 (6,71)	34,91 (6,7)	34,9 (8,63)	34,91 (6,79)	2
Monks-3	3,88 (2,69)	3,86 (2,86)	3,88 (3,34)	1,28 (1,28)	2
Pima	24,5 (5,15)	24,87 (4,83)	25,05 (7,69)	26,11 (5,43)	3
Tic Tac Toe	25,16 (6,31)	30,81 (7,11)	30,51 (5,9)	33,43 (5)	4
Vehicle	28,75 (5,44)	40,62 (7,39)	42,25 (6,52)	34,24 (4,96)	9

Table 2. Test with ID3

Bases	Our method Error (Sd)	MIFS Error (Sd)	ReliefF Error (Sd)	Without Selection Error (Sd)	Number of iterations with our method
Austra	15,27 (3,61)	14,28 (3,08)	15,28 (5,15)	16,6 (4,57)	3
Breast	2,65 (2,05)	2,86 (1,87)	3,45 (2,56)	5,95 (1,95)	5
Cleve	17,77 (6,14)	20,52 (11,34)	40,67 (4,33)	18,53 (8,68)	4
CRX	15,69 (3,99)	14,66 (5,7)	16,53 (2,8)	14,73 (5,68)	3
German	23,43 (4,62)	26,29 (3,63)	30,71 (4,96)	31,86 (7,53)	7
Heart	17,89 (7,14)	17,89 (10,04)	21,05 (10,53)	27,05 (10,29)	4
Iono	7,25 (5,88)	5,22 (4,4)	9,32 (6,22)	21,37 (8,39)	6
Iris	2,82 (4,31)	4,64 (6,17)	6,45 (7,14)	3,73 (4,57)	3
Monks-1	25,19 (4,68)	25,2 (7,18)	51,9 (8,2)	25,22 (8,3)	2
Monks-2	34,92 (5,11)	34,92 (6,24)	34,92 (6,65)	34,91 (6,79)	2
Monks-3	3,85 (3,67)	3,86 (2,87)	3,85 (3,85)	1,28 (1,28)	2
Pima	22,83 (5,73)	21,33 (4,3)	25,04 (3,41)	26,11 (5,43)	4
Tic Tac Toe	27,83 (3,92)	28,87 (5,42)	27,97 (4,19)	33,43 (5)	4
Vehicle	33,95 (4,18)	39,85 (8,01)	45,82 (8,78)	34,24 (4,96)	7

Table 3. Test with Naive Bayesian

- 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.

We plan to improve our method according to two aspects. The discretization method used for the criteria values must be better. Also we would like the result of the preferences aggregation method to be the optimal features subset.

Bases	Without selection	Our method with ID3	Our method with BN	ReliefF	MIFS
Austra	14	1	2	2	13
Breast	9	3	7	6	9
Cleve	13	7	5	6	8
CRX	15	3	5	2	7
German	20	5	9	14	3
Heart	13	2	8	2	13
Iono	34	2	26	25	8
Iris	4	3	2	4	3
Monks-1	6	1	1	2	1
Monks-2	6	1	1	2	2
Monks-3	6	2	2	2	3
Pima	8	2	5	7	4
Tic Tac Toe	9	7	7	5	3
Vehicle	18	14	12	18	6

Table 4. Number of selected features

References

[Battiti, 1994]R. Battiti. Using mutual information for selecting features in supervised neural net learning. 5:537–550, July 1994.

[Breiman *et al.*, 1984]L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression trees, The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA*. 1984.

[De Mantaras, 1991]R.L. De Mantaras. A distance-based attribute selection measure for decision tree induction. In *Machine Learning*, volume 6, pages 81–92, 6-9 1991.

[Hart, 1984]A. Hart. Experience in the use of an inductive system in knoowledge eng. In M. Bramer, editor, *Research and Development in Expert Systems*. Cambridge Univ. Press, Cambridge, MA, 1984.

[John *et al.*, 1994]George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Int'l Conf. on Machine Learning*, pages 121–129, 1994.

[Kira and Rendell, 1992a]K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In MIT Press, editor, *Tenth Nat. Conf. on Artificial Intelligence*, pages 129–134, 1992.

[Kira and Rendell, 1992b]K. Kira and L.A. Rendell. A practical approach to feature selection. In *Proc. of the Tenth Int'l Conf. on Machine Learning*, pages 500–512, 1992.

[Kirkpatrick *et al.*, 1983]S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[Lallich and Rakotomalala, 2000]S. Lallich and R. Rakotomalala. Fast feature selection using partial correlation for multi-valued attributes. In *Proc. of the 4th European Conf. on Knowledge Discovery in Databases, PKDD 2000*, pages 221–231, 2000.

[Langley, 1994]P. Langley. Selection of relevant features in machine learning. In *Proc. of the AAAI Fall Symposium on Relevance*, pages 1 – 5, 1994.

[Liu and Setiono, 1996]Huan Liu and Rudy Setiono. A probabilistic approach to feature selection - a filter solution. In *Int. Conf. on Machine Learning*, pages 319–327, 1996.

[Marcotochino, 1984a]F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, Étude n°f-071. Technical report, Centre Scientifique IBM-France, Février 1984.

[Marcotochino, 1984b]F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, partie ii Étude n°f-071. Technical report, Centre Scientifique IBM-France, Mai 1984.

[Merz and Murphy, 1996]C. Merz and P. Murphy. Uci repository of machine learning databases. *http://www.ics.uci.edu/ mlearn/MLRepository.html*, 1996.

[Nicoloyannis *et al.*, 1998]N. Nicoloyannis, M. Terrenoire, and D. Tounissoux. An optimisation model for aggregating preferences: A simulated annealing approach. *Health and System Science*, 2(1-2):33–44, 1998.

[Nicoloyannis *et al.*, 1999]N. Nicoloyannis, M. Terrenoire, and D. Tounissoux. Pertinence d'une classification. *READ*, 3(1):39–49, 1999.

[Quinlan, 1986]J.R. Quinlan. Introduction of decision trees. In *Machine Learning*, volume 1, pages 81–106, 1986.

[Sémani *et al.*, 2005]Dahbia Sémani, Carl Frélicot, and Pierre Courtellemont. Un critère d'évaluation pour la sélection de variables. In *EGC*, pages 91–102, 2005.

[Shannon, 1948]C.E. Shannon. A mathematical theory of communication. In *Bell System Technical Journal*, 1948.

[Sheinvald *et al.*, 1990]Sheinvald, Dom, Niblack, and Rendell. A modeling approach to feature selection. In *Tenth Int. Conf. on Pattern Recognition*, 1990.

[Tanguiane, 1991]A. S. Tanguiane. *Agregation and representation of preferences: Introduction to Mathematical Theory of Democracy, Springer Verlag.* 1991.

[Vincke, 1982]Ph. Vincke. Aggregation of preferences: a review. *European Journal of Operational Research*, 9:17–22, 1982.

[Zhou and Dillon, 1991]X. Zhou and T.S. Dillon. A statistical–heuristic feature selection criterion for decision tree induction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 13, pages 834–841, 1991.

[Zighed *et al.*, 1996]D. A. Zighed, R. Rakotomalala, and S. Rabasééda. A discretization method of continous attributes in induction graphs. *Proc. Of the 13th European Meetings on Cybernetics and System Research*, 3(1):997–1002, 1996.