A Hybrid Approach to Feature Selection

Moussa Boussouf

IRIN, Université de Nantes, 2 rue de la Houssiniére,
BP 92208 - 44322, Nantes Cedex 03, France.
boussouf@irin.univ-nantes.fr

Abstract. The irrelevant and redundant features may degrade the learner speed (due to the high dimensionality) and reduce both the accuracy and comprehensibility of induced model. To cope with these problems, many methods have been proposed to select a subset of pertinent features. In order to evaluate these subsets, two main approaches are generally distinguished: (1) filter approach: which considers only data i.e. algorithm-independent; (2) wrapper approach: which takes into account both data and a given learning algorithm i.e. algorithm-dependent.
In this paper, we address the problem of subset selection using α–RST (a generalized rough sets theory). We propose an algorithm to find a set of α–reducts which are non deterministic reducts. To select the best one among them, we also propose a *Hybrid Approach* by putting filter and wrapper together to overcome the disadvantages of each approach. Our study shows that generally the highest-accuracy-subset is not the best one as regards to the filter criteria. The highest accuracy subset is found by the new approach with minimum cost.

1 Introduction

In supervised machine learning, an induction algorithm deals with a set of training instances where each instance is described by a vector of feature values and a class label. The induction task consists creating a model of data (training set). This model can be used to predict the label of new instances. The irrelevant and redundant features may reduce predictive accuracy, degrade the learner speed (due to the high dimensionality) and reduce the comprehensibility of the induced model. So, doing away with these features or selecting relevant ones became necessary.

The difficulties which were faced during the feature selection prodecss can be explained as follows: for N features, there are 2^N possible subsets, evaluating them is an impracticable process. The optimal selection can only be done by testing all possible sets of M features chosen from N, i.e. by applying the criterion $\binom{N}{M} = \frac{N!}{M!+(N-M)!}$ times. If there are M relevant features, the total number of times is $\sum_{i=0}^{M}\binom{N}{i} = O(N^M)$. This is prohibitive when N and/or M is large.

To deal with the problem of feature selection, many methods have been proposed. In general, they can be classified into two categories: (1) the filter approach, which serves as a filter to sieve the irrelevant and/or redundant features without taking into account the induction algorithm [1][5][10]; and (2) the wrapper approach , which uses the induction algorithm itself as a black box in the phase of attributes selection to select a good features subset which improve the performance, i.e. the accuracy of the induction algorithm [4][8][11][16]. Although the wrapper approach has significantly improved the accuracy of well known algorithms, like C4.5 and Naive-Bayes, its generalization is limited for many reasons i.e.: (1) the former's computational cost, which results from calling the induction algorithm for each feature subset considered; (2) dealing with large datasets being impracticable. Using the wrapper approach to evaluate random generated subsets LVW, Liu and Setenio in [11] concluded that it is not commendable to use LVW in the applications where time is a critical factor. On the other hand, the main critics of filter approach are: (1) it totally ignores the effects of selected feature subset on the performance of the induction algorithm [4]; (2) various heuristics tend to overestimate the multi-valued attributes [9]. In order to cope with the above problems, the two main approaches were extended considering probability foundations: a probabilistic filter approach [10] and a probabilistic wrapper approach [11] [6].

In the rough sets theory [13][14], many reducts can be found. Their number depends on the indiscernibility of examples and cannot be known beforehand. Consequently, the evaluation and selection of the best reduct is still a serious problem, especially when the number of reducts is large.

In this paper, we use the generalized rough sets theory, called α–RST, which is proposed by Quafafou in [15] as the theoretical framework. We propose a new attribute selection process, called *Hybrid Approach.* This new approach combines the two above approaches. It inherits the advantages and eliminates the disadvantages of both filter and wrapper approach.

The remainder of this paper is organized as follows: section 2 presents the feature selection in the rough sets framework. It introduces the new definitions of the generalized concepts, which are necessary for our investigation. Section 3 presents our reducts algorithm. In order to evaluate the generated reducts, section 4 is an examination of two filters. Section 5 presents our experimental results with filter approach, wrapper approach and with the new hybrid approach. We conclude in section 6.

2 Feature subsets selection in rough sets theory

In the rough sets theory [13][14], the most important and fundamental notions are the need to discover *redundancy*, *dependency* between features, *reduction* of features and definition of the *core*, which is a set of attributes which contains all indispensable features. The feature subset selection can be viewed as finding reducts. In this context, different works have been developed to deal with the problem of subset selection. Moderzejewski in [12] proposes a heuristic feature selector algorithm, called PRESET. It consists ordering attributes to obtain an optimal preset decision tree (according to absolute significance of attribute measure). Kohavi and Frasca in [7] have shown that, in some situations, the *useful* subset does not necessarily contain all the features in the *core* and may be different from a reduct. Using α–RST, we have proposed in [16] an algorithm based on wrapper approach to solve this problem. We have shown that we can obtain lower size reducts with higher accuracy than those obtained by classic rough sets concepts.

Before presenting the algorithm which finds α–reducts, we will present an overview of the necessary concepts of α–RST.

Generalized information system In rough sets theory, an information system has a data table form. Formally, an information system S is a 4-tuple $S = (U, Q, V, f)$, *where* U *: is a finite set of objects.* Q *:* is *a finite set of attributes.* $V = \cup V_q$ *, where* V_q *is a domain of attribute* q*.* f *is an information function assigning a value of attribute for every object and every attribute, i.e.* $f : U \times Q \mapsto V$*, such that for every* $x \in U$ *and for every* $q \in Q$ $f(x, q) \in V_q$.

To deal with quantitative attributes, a preprocessing discretization phase is necessary to transform continuous attributes into qualitative terms (nominal value). This process may influence the results of systems using rough sets theory [17]. In order to take into account uncertainty inherent to both data and preprocessing, Quafafou in [15] has proposed a new information system. The information function f being defined as follows:
$f : U \times Q \rightarrow V \times [0, 1]$ *such that for every* $x \in U$ *and for every* $q \in Q$ $f(x, q) \in (V_q, [0, 1])$.

Each attribute q of each object x is described by a pair of a nominal value and a cardinal value. The cardinal value represents the degree of possibility, $\pi \in [0, 1]$, that the attribute q may have the value V_q for the object x (Table 1).

Definition 1. **α–Indiscernibility relation :** *Let R be a subset of Q, α a given similarity threshold. The α–indiscernibility relation denoted by $IND(R, \alpha)$ is defined as*

$$IND(R, \alpha) = \{(x, y) \in U \times U \mid f(x, q) = f(y, q) \ \forall q \in R, \ and \ \Im(\pi_x, \pi_y) \geq \alpha\}$$

If $(x, y) \in IND(R, \alpha)$, then x and y are α–indiscernible with respect to R, that means that they have the same value of attributes and that their similarity degree, computed by the function $\Im$, is greater than α. Consequently, the equivalence class of an obejct x, denoted $[x]_{IND(R,\alpha)}$, is defined by the set $\{y \in U | (x, y) \in IND(R, \alpha)\}$. The opposite of this relation is called α–*discernibility* relation, denoted $DIS(R, \alpha)$.

Definition 2. **α–Dependency:** *Let P and R be two subsets of attributes, such that $R \subseteq P$ and $\alpha \in [0, 1]$. R α–depends on P if and only if $\exists \beta \in [0, 1]$ such that:* $P \xrightarrow{\beta} R \Leftrightarrow \forall B \in U/IND(P, \alpha) \ \exists B\prime \in U/IND(R, \alpha) \ deg(B \subseteq B\prime) \geq \beta$

The α–dependency can be seen as partial dependency between values of R and P, i.e. R partially explains P.

Definition 3. **α–reducts :** *Let P and R be two subsets of attributes, such that $R \subseteq P$. R is an α–reduct of P if and only if $\exists \beta \in [0,1]$ such that (i) $P \xrightarrow{\beta} R$ and $R \xrightarrow{\beta} P$ and (ii) R is minimal. R is minimal means there is no subset T of R which is an α–reduct of P.*

3 α–Reducts Algorithm

In this section we show how to find α–reducts using α–discernibility relation. Firstly, we calculate the α–discernibility–list, denoted αDL, of all minimal α–discernible subsets which contain the class label. Each element of this list contains a subset of features which discern a pair of examples. To deal with redundancy, all non minimal subsets are deleted. Secondly, we construct the minimal α–reducts–list, denoted αRL, from minimal α–discernible–list.

α–Reducts Algorithm

```
Input    GIS : Generalized Information System of N examples;
         Q : All Features includes Class;
         C : Class;
         $\alpha$ : similarity threshold;
Output   $\alpha$RL := { {} } : $\alpha$–Reducts–List;
         $\alpha$DL =: {} : $\alpha$–Discernibility–List;
for i := 1 to N+1
   for j := i+1 to N
      TempSubset := $\alpha$–DIS(Q,$\alpha$,GIS[i],GIS[j]);
      /* the set of features which discern GIS[i] and GIS[j] */
      if $C \subset$ TempSubset
         if $\not\exists$ E $\in \alpha$ DL such that $E \subseteq$ TempSubset
           add(TempSubset - { C }, $\alpha$DL);
         endif
      endif
   endfor
endfor /*end construction of minimal $\alpha$–Discernibility–List */
if card($\alpha$DL) $\leq 1$ $\alpha$RL := $\alpha$DL
else for each element EDL of $\alpha$DL
         for each element ERL of $\alpha$RL
            for each feature $F$ of EDL
               ERL := ERL $\cup F$;
            endfor
         endfor
         Delete all not minimal elements of $\alpha$RL;
   endfor
endif
```

The example bellow (Table 1) illustrates the presented concepts. For each example, the first value expresses the nominal value of an attribute, the second represents the degree of possibility that this attribute may have this value.
$U = \{e1, e2, ..., e8\}$,
$Q = \{W, X, Y, Z, C\}$
$\alpha\text{-}DIS(Q, 0, e1, e2) = \{W, X, Y, C\}$;
$\alpha\text{-}DIS(Q, 0, e1, e3) = \{W, X, Y, Z, C\}$.

	W	X	Y	Z	C
e1	5 1.0	1 0.6	3 0.6	2 0.7	2
e2	4 1.0	4 0.8	4 1.0	2 1.0	0
e3	3 0.9	2 1.0	2 1.0	0 1.0	1
e4	5 0.7	3 0.6	4 0.7	1 0.6	1
e5	3 0.6	4 1.0	3 0.8	1 0.8	2
e6	4 0.6	1 1.0	3 1.0	0 1.0	0
e7	3 1.0	3 1.0	4 0.8	1 1.0	1
e8	5 0.7	3 0.7	4 0.9	2 0.7	2

Table 1. Example of 8 elements

Where α=0, the reader can verify that:
α–discernibility–list: $\alpha DL = \{\{X, Y\}, \{W, X\}, \{Z\}\}$.
α–reducts–list : $\alpha RL = \{\{X, Z\}, \{W, Y, Z\}\}$.

The strong characteristics of α–Reducts Algorithm are: (1) The parameter α influences and controls the granularity and the number of reducts, i.e.: when *alpha* increases, the size of reducts decreases and the number of reducts generally increases; (2) we obtain the classical reducts (corresponding to classical framework of rough sets) when α=0; (3) the best reduct with highest accuracy is obtained when $\alpha > 0$ i.e. using α–RST concepts [16]; (4) theoretically, the algorithm can find $\binom{N}{\lfloor N/2 \rfloor}$ reducts.

4 Evaluating α–Reducts

As described above, α–reducts algorithm can generate many reducts. Selecting the best one is still a serious problem, especially when the number of α–reducts is high. Consequently the application of the wrapper method is impracticable. We focus our study on examining two filter algorithms which evaluate a given subset entirely. Other heuristic measures are summarized in [9].

Almuallim and Dietterich MIG: In order to improve FOCUS algorithm, Almuallim and Dietterich in [1] proposed tree heuristics for the MIN–FEATURES bias. The Mutual–Information–Greedy algorithm use the entropy measure to evaluate a subset entirely. This algorithm searches the minimum-size attribute subset sufficient to maintain consistency on the training data. Generally, databases are not consistent even when all attributes are present, i.e. a subset cannot be more consistent than its superset. Caruana and Freitag in [2] solved this problem to apply FOCUS in CAP by returning the smallest subset of attributes that contains the same inconsistency compared with all attributes.
Liu and Setiono inconsistency measure: Liu and Setiono in [10] proposed a measure to evaluate selected features. The inconsistency rate of data described by a selected subset is checked against a prespecified threshold. The one which has the lowest value is chosen for further tests using a learning algorithm.

To apply the filter measures described above, the subset which has minimum value must be returned.

5 Experiment Results

In order to evaluate candidate strong feature subsets generated by α–reducts algorithm, we ran experiments on 7 real-world datasets taken from the UCI Irvine repository, their characteristics are summarized in Table 2. Original data are transformed using Fayad and Irani method [3].

Datasets	Exam.	Att.	Cla.	%Num.
Iris	150	4	3	100
Pima	768	8	2	100
Australian	690	14	2	43
Glass	214	9	6	100
Heart	270	13	2	46
Vehicle	746	18	4	100
Wave	600	21	3	100

Table 2. Datasets considered: the number of examples and attributes, class cardinality and the percentage of numeric features

To estimate the accuracy for feature subsets we used 5-fold cross validation. The algorithm ALPHA is used as an inducer algorithm.

We applied the above filter on generated reduct. The best subset, the one which has the lowest value is chosen for further tests using the learning algorithm. We also applied the wrapper model around all reducts.

5.1 Filter *or* wrapper

In our experiments we found that the best subset selected by Liu&Setanio filter is generally different from the subset selected by Almualim&Dietrich filter. Table 3 shows that the reduct chosen by the two filters improves slightly the accuracy of ALPHA (with all features) for Pima, Australian and Heart. The accuracy falls for Vehicle, Wave and significantly for Glass. On the other hand, the reduct selected by wrapper method improves the accuracy of ALPHA in all databases except for Glass. The improvements are significant for Heart, Vehicle and Wave.

Datasets	α	N. Red.	ALPHA	Filter/ALPHA		Wrapper/	Size of
				L&S	A&D	ALPHA	best Red
Iris	0.2	1	96.00	96.00	96.00	96.00	3
Pima	1.0	3	75.39	75.78	75.78	75.78	7
Australian	0.8	6	81.44	82.90	82.32	**83.04**	12
Glass	1.0	**10**	**61.68**	56.34	56.34	**59.91**	5
Heart	0.9	**51**	83.33	84.81	84.81	**85.56**	7
Vehicle	0.9	**288**	68.44	67.73	68.91	**70.33**	10
Wave	0.5	**4349**	74.00	69.67	72.67	**77.50***	11
Mean			77.18	76.18	76.69	78.44	

Table 3. Similarity threshold, number of α–reducts; the accuracy of ALPHA, the accuracies of ALPHA using Liu and Setiono filter, Almualim and Dietrich filter, wrapper model and size of best reduct. *: Due to the former's computational cost, we ran wrapper method around 1600 reducts.

The accuracy of the reducts chosen by the wrapper method is superior to that of the filter method, especially when the number of reducts is higher than one. In the Wave database the difference with L&S filter equals 4.83% and with A&D filter equals 7.83% .

In order to capture the best reduct which improves significantly the accuracy, and which is not obtained by the filter method, we have considered filtering a few reducts and execute the induction algorithm around them, which is the topic of the next section.

5.2 The Hybrid Approach: filter *and* wrapper

As described above, the main disadvantage of the wrapper approach is the former's computational cost. For example running inducer algorithm on 4349 subsets of Wave database takes a lot of time. However, the various filters do not obtain relevant results. We propose a new *Hybrid Approach*, it combines the two above approaches, which were based on tree phases: (1) α–reducts computation, (2) filtration of α–reducts to focus our attention on only a few reducts, (3) use of the wrapper method to select the best one among the filtered ones. This new approach inherits the advantages and eliminates the disadvantages of both filter and wrapper approach. We define a threshold of allowable filter rate γ. The learning algorithm is applied only on each reduct in which the filter criterion rate is lower than γ.

Hybrid algorithm

```
Input    GIS : Generalized Information System (training data);
         $\alpha$ : similarity threshold;
         $\gamma$ : allowable filter criterion rate;
Output   Hsfs : Hybrid strong feature subset;
MaxAcc=0;
CandidateHsfs := $\alpha$Reducts(GIS,$\alpha$) /*Calling $\alpha$–Reducts algorithm*/
while NotEmpty(CandidateHsfs) do
      Current := select(CandidateHsfs);
      CandidateHsfs := CandidateHsfs - Current;
      if filter(Current) $\leq \gamma$
         Acc=IAlgo(GIS,$\alpha$); /*Calling Inducer Algorithm*/
         if Acc > MaxAcc
            Hsfs := current;
            MaxAcc := Acc;
         endif
      endif
endwhile
```

We apply the hybrid approach on databases which have many reducts. Experimentally, the allowable filter rate value depends on the original datasets inconsistency. The inconsistency of Glass, Heart, Vehicle and Wave, according to L&S

measure criterion are: 11.21%, 0.74%, 3.31% and 0.0% respectively. In addition to the advantages of α–reducts algorithm mentioned in the 3rd section, all generated reducts when α=0 have the same inconsistency value, and they are equal to inconsistency value of original datasets (with all attributs).

For the remainder of the experiments, we put the allowable filter rate equal to: 10%, 2%, 6% and 0.5% for the above databases respectively (we put other adjusted value for A&D filter).

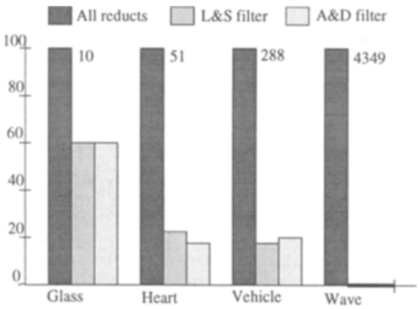

Figure 1. Pourcentage of filtred reducts using L&S and A&D filter.

Datasets	Hybryd/ALPHA
Glass	59.91
Heart	85.56
Vehicle	70.33
Wave	77.50

Table 4. The accuracy of reducts selected by the hybrid approach

Figure 1 shows the percentage of filtered reducts. The induction algorithm used around 6 reducts among 10 for Glass. It is only used to evaluate 22% and 17% of subsets for Heart and Vehicle using L&S inconsistency measure. The best subset found by the wrapper approach is captured in the mentioned interval. It is found with minimum cost (very early) comparing with the wrapper approach. The very interesting result is obtained in Wave database. Among 4349 reducts, we only evaluated 33 reducts (0.76%). It is the 5th best reduct according to L&S filter. It improves the accuracy of ALPHA by 3.5%. In addition, its accuracy is greater than the L&S filter by 7.83%.

6 Conclusion

In this paper we have studied the problem of feature selection using the generalized rough sets theory, α–RST. We have proposed an algorithm to find non deterministic reducts. We have developed the wrapper approach and some filter measures which evaluate a subset entirely. Assuming that (α–)rough sets theory can generate many (α–)reducts, our experiments show that the accuracy of the best reduct using a filter method is lower than the best one using the wrapper approach. The hybrid approach proposed in this paper overcomes the bias of filter approach, by proposing many probable best reducts and speeds up the wrapper approach, by wrapping only around the filtered ones. It inherits the performance of the wrapper approach by guaranteeing the selection of the best reduct with respect to an induction algorithm.

References

1. Almuallim, H., Dietterich, T.G.: Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, **69(1-2)** (November 1994) 279–305
2. Caruana, R., Freitag, D.: Greedy attribute selection. Proceedings of the Eleventh International Conference. M. Kaufman Publ. in Cohen&Hirsh eds, Machine learning Inc. (1994) 28-36
3. Fayyad U.M., Irani K.B.: Multi-interval Discretization of Continuous-attributes for classification learning IJCAI'93, (1993) 1022–1027
4. John, G. H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection Problem. In Proceedings of the Eleventh International Conference on Machine Learning, (1994) 121–129
5. Kira, K., Rendell, L.A.: The feature selection problem: traditional methods and a new algorithm. In Proceedings of the 9th National Conference on Artificial Intelligence, (1992) 129–134
6. Kohavi, R.: Feature subset selection as search probabilistic estimates. AAAI Fall Symposium on Relevance, (1994) 122–126
7. Kohavi, R., Frasca, B.: Useful feature subset and rough sets reducts. In Proceedings of the Third International Workshop on Rough Sets and Soft Computing, (1994) 310–317
8. Kohavi, R., Sommerfield, D.: Feature subset selection using the wrapper method: over-fitting and dynamic search space topology. In proceeding of the First International Conference on Knowledge Discovery and Data Mining, (1994) 192–197
9. Kononenko, I.: On biases of multi-valued attributes. In proceeding oh the 14th international joint conference on artificial intelligence, in C.S.mellish ed. (1995) 1034–1040.
10. Liu, H., Setiono, R.: A probabilistic approach for feature selection: A filter solution. In 13th International Conference on Machine Learning (ICML'96), (1996) 319–327
11. Liu, H., Setiono, R.: Feature selection and classification - a probabilistic wrapper approach. In Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES, (1996) 419–425
12. Modrzejewski, M.: Feature selection using rough sets theory. In Proceedings of the European Conference on Machine Learning, (1993) 213–226
13. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht, The Netherlands (1991)
14. Pawlak, Z.: Rough Sets: present state and the future. Foundations of Computing and Decision Sciences, **18(3-4)**, (1993) 157–166.
15. Quafafou, M.: α-RST: A generalization of rough sets theory. In Proceedings in the Fifth International Workshop on Rough Sets and Soft Computing, RSSC'97. (1997)
16. Quafafou, M., Boussouf, M.: Induction of Strong Feature Subsets. In 1st European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97, (1997) 384–392
17. Slowinsky K., Slowinsky R.: Sensitivity analysis of rough classification. International journal of Man-Machine studies, **32** (1990) 693–705