Induction of Strong Feature Subsets

Mohamed Quafafou and Moussa Boussouf

IRIN, University of Nantes, 2 rue de la Houssiniere, BP 92208 - 44322, Nantes Cedex 03, France.
quafafou@irin.univ-nantes.fr

Abstract

The problem of features subset selection can be defined as the selection of a relevant subset of features which allows a learning algorithm to induce small high-accuracy concepts. To achieve this goal, two main approaches have been developed, the first one is algorithm-independent (filter approach) which considers only the data, when the second approach takes into account both the data and a given learning algorithm (wrapper approach). Recent work were developed to study the interest of rough set theory and more particularly its notions of reducts and core to deal with the problem of features subset selection. Different methods were proposed to select features using both the core and the reduct concepts, whereas other researches show that useful features subset does not necessarily contain all features in cores. In this paper, we underline the fact that rough set theory is concerned with deterministic analysis of attribute dependencies which are at basis of the two notions of reduct and core. We extend the notion of dependency which allows us to find both deterministic and non-deterministic dependencies. A new notion of strong reducts is then introduced and leads to the definition of strong feature subsets (SFS). The interest of SFS is illustrated by the improvement of the accuracy of our learning system, called Alpha, on real-world datasets.

1 Introduction

The problem of features subset selection has been studied by researchers working on different fields, for instance, statistics, pattern recognition, machine learning and knowledge discovery [11]. It consists of the selection of a subset of relevant attributes which will be used by a learning algorithm. The selected subset can be evaluated according to: (1) the complexity of the description induced by the learning algorithm using the selected subset, and (2) the improvement of the accuracy of the learning algorithm using the induced concepts. An exhaustive search of subsets in the features space, which tests all possible subsets, is prohibitive because it inquires exponential computation. In order to overcome this problem, heuristic methods are generally used to search the relevant features upon which a learning algorithm will be focused. Different definitions were proposed to formalize the notion of relevance making assumption on the nature of features to be considered, i.e., boolean, continuous, noisy, etc. [5]. A lot of methods have been developed to find an optimal subset of relevant features [6], [1]. Two main approaches have been developed, the first one is algorithm-independent, i.e., the filter approach [1], considering only the data, whereas the

second, i.e., the wrapper approach [5] takes into account both the data and a learning algorithm.

Recently, other types of works have been developed to deal with the problem of subset selection in the context of rough set framework [13], [14]. An heuristic feature selector algorithm [12], called Preset, was developed based on the rough set theory. On the other hand, experiments reported in [9] have shown that, in some situations, useful subsets do not necessarily contains all features in cores and may be different from reducts. Considering this observed difference, the latter authors suggest to start the research from the core to identify useful subsets. This paper gives an explanation of the results of the two previous results and proposes a new method to deal with the problem of subset selection in the rough sets framework. It offers an alternative view of the problem considering that useful subset does not result only from modifications of the core using some heuristics.

Our purpose results from a deep study on both the theoretical foundations of rough sets and its main concepts which influence the subset selection problem. In this paper we underline that rough set is only concerned with the analysis of deterministic attributes dependencies, which plays a major role in subset selection processes based on rough set framework. We show also that the granularity of knowledge is primary of importance in rough set theory, hence, it plays a key role in the selection problem. We extend the dependency notions to allow the analysis of both deterministic and non-deterministic attribute dependencies. On the other hand, we generalize the notion of indiscernabilty to control the knowledge granularity. Next we propose a method to discover Strong Feature Subsets (SFS) and we discuss the interest of this new concept through our experiments.

Section 2 introduces formally three key notions in rough set theory, which are dependency, reducts and core. Generalizations of these notions are given in section 3. The problem of the control of knowledge is introduced and the SFS algorithm, which research strong feature subset, is sketched in Section 4. Section 5 is dedicated to the discussion of our experiments. We conclude in section 6.

2 Dependency and reducts in rough set theory

The basic operations in rough set theory are approximations which are defined according to the indiscernability notion [15], [3]. In fact, objects of the universe can be grouped according to values of a given set of attributes R. Each group contains objects with the same values of attributes in R. This means that the information given by R is not sufficient to distinguish objects in the same group: objects of the same group are said indiscernable. These groups, called R-elementary sets, form basic granules of knowledge about the universe and are used to compute approximations of concepts. The indiscernability relation, denoted IND(R), is an equivalence relation defined on the equality of attributes values. The quotient set U/IND(R) contains equivalence classes, sets of indiscernable objects, which are granules of knowledge representation. In rough set

theory, we say that a set of attributes R depends on a set of attributes P, denoted $P \rightarrow R$, if and only if all P-elementary sets are subsets of some R-elementary sets.

Let P be a set of attributes which may be reduced such that the reduced set R provides the same quality of classification as the original set of attributes. This means that elementary sets generated by the reduced set R are identical to those generated by the original set of attributes P, i.e., IND(P) = IND(R). The smallest reduced set of P is called a reduct. In the case where an information system has more than one reduct, the intersection of these reducts is computed. The resulted intersection is called the core and represents the most significant attributes. The following proposition states the relation between the dependency and reduction notions.

Proposition 1: Let P and R be two subsets of attributes, $R \subseteq P$. R is a reduct of P if and only if $P \rightarrow R$ and $R \rightarrow P$ and R is minimal.

The proposition 1 shows that rough set is only concerned with the analysis of deterministic attribute dependencies. Previous works [12], [9] developed to deal with features subset selection using rough set theory inherit its deterministic view and start from the notion of reducts and core which are both based on the dependency notion. In contrast, we have developed a formalism [17], i.e., an extension of rough set theory, to recognize non-deterministic relationships which will be at the basis of the identification of strong feature subsets.

3 Alpha-reducts are generalized reducts

The notion of Alpha-reducts results from our extension which substantially enhance the generalization and the application of rough set theory to real world problems [17], [18]. Different extensions of the standard rough set theory have been suggested in [2], [4], [16]. We consider that values of attributes may be vague, for instance, high, low, medium, etc., so, the domain of each attributes is a set of linguistic terms represented by membership functions to specify how to evaluate the truth of the attribute value on the learning examples. A membership function maps a numerical domain of an attribute to the interval [0, 1], i.e., the abscissa represents the values of the attribute, whereas the ordinate represents the truth value. Each value is then associated with a linguistic term and a truth value; if a numerical value is associated with several membership functions, only the one with the maximum grade is kept. Consequently, a transformation process replaces the original description vector of an example into two vectors containing qualitative descriptions, i.e., linguistic terms, and quantitative description, i.e., truth values. A parametrized indiscernability relation, denoted IND(R,α), is then defined based on both qualitative and quantitative descriptions:

$$\forall x, y \in U, xIND(R, \alpha)y \iff xR_0y \text{ and } \Im(x, y) \geq \alpha$$

We consider that two elements x and y are indiscernable if and only if they have the same values for all attributes, i.e., xR_0y , and if their similarity degree

is greater than a given similarity threshold α given by a teacher. The quotient set U/IND(R,α) contains equivalence classes which are granules of knowledge representation.

In this paper we focus only on the extension of notions of dependency, reduction and core. Thus, the dependency relation is not black or white relation, we say that a set of attributes R depends on a set P if and only if each P-elementary set X have a non empty intersection with at least an R-elementary set Y, and the inclusion degree of X in Y is greater than a dependency parameter, noted β. We call this property Alpha dependency of attributes.

Definition 1: Let P and R be two subsets of attributes,$R \subseteq P$ and $\alpha \in$ [0,1]. R Alpha-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 previous definition introduces the notion of α-dependency which can be seen as a partial dependency between attributes. Consequently, the values of attributes in R are partially determined by values of attributes in P. We say that R partially explains P and there is only a partial functional dependency between values of R and P. Now, we introduce the notion of α-reduct.

Definition 2: let P and R be two subsets of attributes, such that $R \subseteq P$. R is an alpha-reduct of P if and only if $\exists B \in [0,1]$ such that (i) $P \xrightarrow{\beta} R$ and $R \xrightarrow{\beta} P$, (ii) β is maximal and (iii) R is minimal.

R is minimal means that there is no subset T of R which is an Alpha-reduct of P. Besides, β is maximal means that $\nexists \beta\prime \in [0,1]\ \beta\prime > \beta\ P \xrightarrow{\beta\prime} R$ and $R \xrightarrow{\beta\prime} P$. As a set of attributes may have more than one Alpha-reduct, we generalize the notion of the core of an attribute P by introducing the notion of Alpha-core.

Definition 3: Let $\alpha \in [0,1]$, called the core threshold, and θ-reduct(P) a family of all $\theta - reduct$ of P, with θ greater than α. The Alpha-core of P is the intersection of all these θ-reducts: $\alpha - core(P) = \bigcap_{\theta \geq \alpha} \theta - reduct(P)$.

In this section, we have generalized three key concepts of the rough set theory, which are dependency, reduction and core. These concepts are generally used in the rough set analysis to construct minimal subsets of attributes (reducts) which have the same indiscernability power as the whole set of attributes. This analysis leads to the construction of deterministic reductions and a deterministic core, i.e., the decision depends on reduction and the core is an essential part of the whole set of attributes. Alpha-rough set allows the user to explore the information using different thresholds related to the reduction or the core. This analysis may lead to the construction of strong reductions and cores, which are only consistent with a part of the dataset and we may regard the remaining inconsistency information as noisy data.

4 Knowledge granularity control-based subset selection

The method presented in this paper to search SFS is based on a wrapper model, indeed, the induction learning algorithm Alpha is used during the subset selection.The SFS is defined here as the "best alpha-reduct", which gives the greatest accuracy. The dependency of two subsets of attributes R and P depends on the quotient sets U/IND(R, α) and U/IND(P, α), which contain equivalence classes, i.e., knowledge blocks, resulting from the universe partition. The control of the knowledge granularity means that the control of the size of elements included in the quotient spaces. To achieve this goal, we vary the similarity threshold α in the range [0, 1], thus, the coarsest partitioning results when α is equal to 0 and the finest partitioning is obtained with α equal to 1. Note that the classical rough set theory considers only the coarsest partitioning of the universe, i.e., α =0. Consequently, works presented previously using rough set theory to deal with feature subset selection have used only coarse granules of knowledge. We present the SFS algorithm to determine the SFS according to the parameter α, which determines the finest of granularity to consider during the search of alpha-reducts.

Consequently, the resulted SFS changes according to the finest of granules of knowledge. The research of strong subset consists of two phases: (1) research of candidate strong features subsets, (2) the selection of one subset among them. As we mentioned earlier, the candidate SFS are alpha-reducts which depend on both the dataset, i.e., DB, and the similarity threshold, i.e., α. The function Alpha-Reducts produces the set Candidate_sfs of candidate SFS. It is an adaptation of a classical algorithm generally used in rough set framework to compute reducts. Next, a learning algorithm, i.e., IAlog, is applied and its accuracy is estimated for each candidate subset.

```
SFS algorithm
input DB: dataset
       α : similarity threshold
output sfs: strong feature subset
begin
MaxAcc = 0
universe partition
Candidate_sfs = Alpha-Reducts(DB, α)
while not_empty(Candidate_sfs ) do
    Current = select(Candidate_sfs)
    Candidate_sfs = Candidate_sfs - Current
    Acc = IAlgo(DB, Current, α)
    if (Acc > MaxAcc) then
      begin
      sfs = Current
      MaxAcc = Acc
      end
end
end
```

The SFS is the Alpha-reduct which have the highest accuracy. In conclusion, the resulted SFS may change according to the dataset DB and the similarity threshold α which determines the granularity of knowledge to be used during the research. Note that the use of the classical rough set framework leads to a research of reducts in the particular situation where α is none.

5 Experimental results

We report here some results of our experiments to show the interest of SFS, which improve significantly the accuracy of Alpha on real-world data. We consider two real-world datasets, heart and pima, taken from the UCI Irvine repository. We vary the parameter α from 0 to 1 by step 0.1, all candidate SFS, i.e., Alpha-reducts, are determined for each value of α . Next, the accuracy of each "candidate SFS" is estimated and the SFS is the candidate one with the highest accuracy.

The original data is transformed using five membership functions for each continued attribute. Indeed, the range of a continuous attribute is split by hand into five intervals with approximately the same length. A membership function is then defined on each subinterval. Even if the discretization is not optimal, Alpha's performances are increased using only a SFS instead of all the original set of attributes. The accuracy of Alpha was estimated using 5-fold cross-validation.

Table 1 shows the evolution of the SFS according to the granularity of knowledge considered during the research of Alpha-reducts. For the heart dataset, we have found 34 Alpha-reducts using the coarsest granules of knowledge, i.e., α= 0. The SFS is {0,1,2,3,4,8,10,11,12}; its accuracy is of 81.41%, whereas the accuracy of Alpha is equal to 73.61%. The number of the candidate SFS grows when α increases from 0 to 1, the maximum number of reducts is equal to 155 for $\alpha = 0.9$. The most important SFS appear when $\alpha = 0.9$ and improves significantly the accuracy of Alpha from 71.00% to 84.01%. Of course, this optimal SFS will not be found if we use only the classical framework of rough set. The accuracy of Alpha was also improved on the pima dataset using SFS. In this case, only one reduct was found with the coarsest granules of knowledge; it is equal to the set of all original attributes. In contrast with the heart dataset, the use of classical rough set framework does not lead to any improvement of the accuracy of Alpha. The SFS with the highest accuracy, i.e., {1,3,5,6}, is obtained with the finest granules of knowledge, i.e., α=1. Other experiments not reported here confirm that SFS improves generally the accuracy of Alpha and that the optimal SFS does not appear when we use the coarsest granules of knowledge,i.e., α=0.

Figure 1 shows the evolution of the size of the candidate SFS set according the granularity of knowledge. This evolution depends also on the datasets, for instance the maximum number of SFS obtained for pima equal to 25, is less than the minimum number of SFS for heart, which is equal to 34. The evolution of the number of SFS is not linear even if it generally increases with fine granules of knowledge. The size of the candidate SFS set influences the time performances of the SFS algorithm to find the optimal SFS. However, the following heuristic has resulted from our experimentation: "the useful SFS appears in the biggest candidate SFS set". This heuristic may reduce the complexity of our algorithm especially when we consider large datasets. Figure 2 shows the average size of candidate SFS sets: the bigger the granule of knowledge, the smaller the candidate SFS.

	α	Alpha	SFS	Alpha-SFS
heart	0.0	**73.61**	{0,1,2,3,4,8,10,11,12}	**81.41**
	0.4	71.00	{1,2,4,5,6,7,8,11,12}	81.04
	0.7	**77.32**	{0,2,3,4,10,11,12}	81.41
	0.9	71.00	**{2,3,4,6,10,11,12}**	**84.01**
	1.0	72.11	{2,4,10,11,12}	81.30
pima	0.0	**65.36**	{0,1,2,3,4,5,6,7}	**65.36**
	0.5	66.93	{0,1,2,3,4,6,7}	69.92
	0.8	67.45	{0,1,2,4,5,6}	71.48
	0.9	67.45	{0,1,2,6,7}	72.27
	1.0	67.45	**{1,3,5,6}**	**74.09**

Table 1: Improvement of Alpha's accuracy using SFS

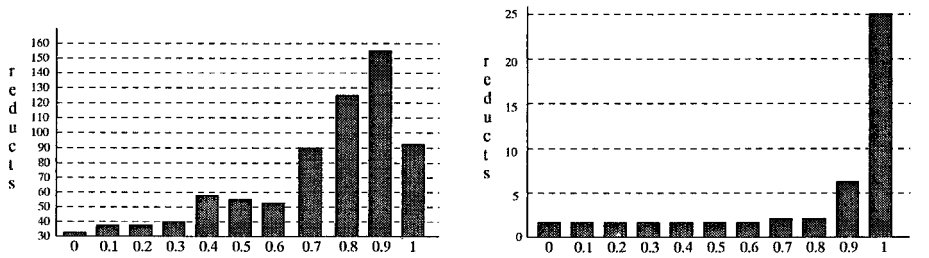

Figure 1: Evolution of the number of the candidate SFS according to the granularity of knowledge

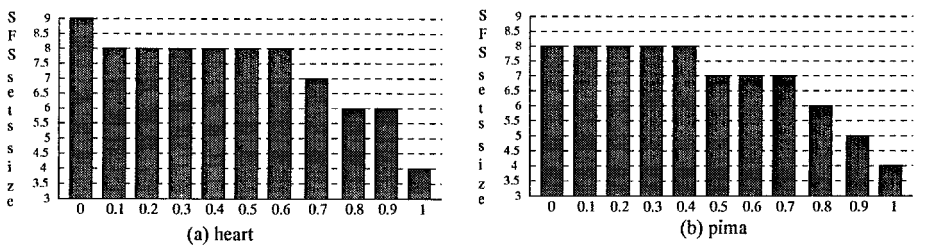

Figure 2: Average size candidate SFS set according to the granularity of knowledge

6 Conclusion

This paper is focused on the problem of features subset selection using rough set theory. Related work have been developed to solve this problem concerned with the two main problems: (1) development of methods to solve the selection problem using rough set theory, (2) evaluation of the interest of the core and reduct notions for the same problem. A heuristic feature selector algorithm [12] was developed, on the other hand, experiments reported in [9] have shown

that, in some situations, useful subset does not necessarily contain all features in the core and may be different from a reduct. However, the authors of these experiments believe that the core may help the identification of useful subsets if we take it as the input of research process. Until now, no method is proposed to achieve this goal to make clear this problem. Our analysis is different, we consider that it is difficult to find the appropriate heuristics, which transform the core and/or reducts into useful subsets. Our proposition starts from the following two remarks : (1) rough set theory is only concerned with the analysis of deterministic attributes dependencies, (2) granularity of knowledge is of primary importance in rough set theory and influences the research of reducts and core. Consequently, we propose a generalization of the dependency notion which are the basis of the definitions of both reduct and core. We generalize the notion of indiscernability to allow a control of the partitioning of the universe, and hence the granularity of knowledge. We show then that the rough set theory uses only the coarsest granules of knowledge, which reduces, in some situations, their interest, in contrast with useful subsets. Finally, we have defined a notion of strong features subset (SFS) and we have shown its interest in the context of a wrapper approach.

References

[1] Almuallim, H., Dietterich, T.G. (1994): Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69 (1-2), pp.279–305

[2] Deogun J.S., Raghavan, Sever H. (1995): Exploiting upper approximation in the rough set methodology. In International Conference on Knozledge Discovery and Data Mining, pp.69-74.

[3] Grzymala-Busse J.W. (1995): Rough Sets. Advances in imaging and physics, Vol. 94, pp.151-195

[4] Hadjimichael M., Wong S.K.K. (1995): Fuzzy representation in rough set approximation. Advances in imaging and physics, Vol. 94, pp.151-195

[5] John, G. H., Kohavi, R., Pfleger, K. (1994): Irrelevant features and the subset selection Problem. In Proceedings of the Eleventh International Conference on Machine Learning, pp.121-129.

[6] Kira, K., Rendell, L.A. (1992): The feature selection problem: traditional methods and a new algorithm In Proceedings of the 9th National Conference on Artificial Intelligence, pp.129-134.

[7] Kittler, J. (1986): Feature selection and extraction In Young, T.Y. and Fu, K.S. (eds.), Handbook of Pattern Recognition and Image Processing. Academic Press, New York.

[8] Kohavi, R. (1994): Feature subset selection as search probabilistic estimates AAAI Fall Symposium on Relevance, pp.122-126.

[9] Kohavi, R., Frasca, B. (1994): Useful feature subset and rough set reducts In Proceedings of the Third International Workshop on Rough Sets and Soft Computing, pp.310-317.

[10] Kohavi, R., Sommerfield, D. (1995): Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In proceeding of the First International Conference on Knowledge Discovery and Data Mining, pp.192-197.

[11] Langley, P. (1994): Selection of relevant features in machine learning In Proceedibgs of the AAAI Fall Symposium on Relevance. AAAI Press.

[12] Modrzejewski, M. (1993): Feature selection using rough sets theory. In Proceedings of the European Conference on Machine Learning, pp.213-226.

[13] Pawlak Z. (1991): Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[14] Pawlak Z. (1993): Rough Sets: present state and the future. Foundations of Computing and Decision Sciences, 18 (3-4), pp.157-166

[15] Pawlak Z., Grzymala-Busse J.W., Slowinski R., Ziarko W. (1995): Rough Sets. Communications of the ACM, Vol. 38, No 11, pp.89-95

[16] Pawlak Z., Wong S.K.M., Ziarko W. (1990): Rough Sets: Probabilistic versus Deterministic In B.R. Gaines and J.H. Boose, editors, Machine Learning and Uncertain Reasoning, pp.227-241, Academic Press.

[17] Quafafou, M. (1997): α-RST: A generalization of rough set theory. In proceedings of the Fifth International Workshop on Rough Sets and Soft Computing, North Carolina, USA.

[18] Quafafou, M. (1997): On the Roughness of Fuzzy Sets In proceedings of the European Symposium on Intelligent Techniques, Bari, Italy.

[19] Vafai, H., De Jong, K. (1993): Robust Feature Selection Algorithms In Proceeding of the Fifth International Conference on Tools with Artificial Intelligence, IEEE Computer Society Press, pp.356-363.

[20] Ziarko, W. (1991): The discovery, analysis, and representation of data dependencies in databases. In Piatetsky-Shapiro, G., and Frawley, W., eds., Knowledge Discovery in Databases, MIT Press.