Characterization of Relevance and Irrelevance in Empirical Learning Methods based on Rough Sets and Matroid Theory

Shusaku Tsumoto and Hiroshi Tanaka
Department of Information Medicine, Medical Research Institute,
Tokyo Medical and Dental University
TEL: +81-3-3813-6111 (6159), FAX: +81-3-5684-3618
email: {tsumoto, tanaka}@tmd.ac.jp

Abstract

One of the most important characteristics of empirical learning methods, such as AQ, ID3(CART), C4.5 and CN2, is that they find variables which are relevant to classification. In this paper, we define relevance in empirical classifier as relevance of each given attribute to apparent or predictive classification, and describe this type of relevance in terms of rough sets and matroid theory. The results show that these algorithms can be viewed as the greedy algorithms searching for apparent classification and that their weight functions may play an important role in predictive classification.

1. Introduction

1.1 Motivations

One of the most important characteristics of empirical learning methods, such as AQ (Michalski 1983; Michalski 1985) ID3 (CART)(Breiman et al. 1984; Quinlan 1986), C4.5(Quinlan 1986), and CN2(Clark & Niblett 1989), is that they find some variables which are relevant to classification.

These four classical methods consist of two main procedures, splitting in ID3, which corresponds to INDUCE method in AQ, and pruning in ID3, which corresponds to truncation in AQ. First, these methods calculate combination of attribute-value pairs, which is the best for classification of training samples. In other words, the procedures derive solutions relevant to classification of original samples, which we call *apparent relevance*, or *allocation relevance.* However, the induced results are only optimal to these given data, and may not be optimal to the future cases. That is, they are overfitting to the training samples. So, in the second step, they remove some variables in order to resolve this undesirable nature, which we call *overfitting irrelevance.* These kinds of relevance and irrelevance are discussed in the Machine Learning literature (Breiman et al. 1984; Clark & Niblett 1989; Michalski et al. 1986; Quinlan & Rivest 1989).

However, there is also other kind of irrelevance. While these empirical learning methods are powerful in the case of training samples whose sample size is small and which have a lot of attributes, their performance will degrade when training samples have many attributes. In the case of AQ, a large number of rules consistent with training samples are derived, and in the case of ID3, too little rules are obtained. We call this irrelevance *irrelevant rule generation.*

In this paper, we focus on the former two types of relevancies and irrelevancies in empirical learning methods, and studying formal characterization of these three irrelevancies. So our sense of relevance in this paper is that "**Relevance in empirical learning methods (empirical classifiers) is defined as relevance of each given attribute to apparent or predictive classification.**".

For the characterization, we introduce matroid theory(Welsh 1976) and rough sets(Pawlak 1991) to construct a common framework for empirical machine learning methods which induce knowledge from attribute-value pattern database. Combination of the concepts of rough sets and matroid theory gives us an excellent framework and enables us to understand these relevancies and irrelevancies and the differences of these methods clearly.

Using this framework, we obtain four interesting conclusions from our approach. First, AQ method is equivalent to the greedy algorithm for finding bases of Matroid from space spanned by attribute-value pairs (ID3 method calculates ordered greedoids, which are defined by weaker axioms than matroids.) Second, according to the computational complexity of the greedy algorithm, the efficiency of both methods depends on the total number of attributes, especially, dependent variables. Third, when we give a suitable weight function, the greedy algorithm calculates combination of attribute-value pairs which is optimal to the weight function. Fourth, the induced results are optimal to the training samples if and only if the conditions on independence are hold. So if adding some new examples make independent attributes change their nature into dependent ones, the condition of deriving optimal solution is violated.

The paper is organized as follows: in Section 2, the elementary concepts of matroid theory are introduced,

and several characteristics are discussed. Section 3 presents AQ method as the Greedy algorithm for AQ matroid. Section 4 gives formulation of weight functions. In Section 5, we consider about overfitting irrevalence as to weight functions.

1.2 Notation and Some Assumptions

In this paper, due to the limitation of the space, we only focus on inducing method of stars in AQ, reduction method in Rough Set Theory, and splitting method in ID3, and we do not consider about generalization (Michalski 1983), truncation (Michalski et al. 1986) and pruning(Breiman et al. 1984; Quinlan 1986). These methods are also formalized by matroid theory if we strengthen our original matroid model, defined as below, by providing some additional concepts. And, moreover, we also have to omit the proofs of the theorems because of the space limitation. For further information, readers could refer to (Pawlak 1991; Tsumoto 1994; Welsh 1976).

Below in this subsection, we mention about the following four notations used in this paper. First, for simplicity, we deal with classification of two classes, one of which are supported by a set of positive examples, denoted by D_+ and the other of which are by a set of negative examples, $D_- = U - D_+$, where U is the total training samples. And the former class is assumed to be composed of some small clusters, denoted by D_j, that is, $D_+ = \cup_j D_j$.

Second, we regard an attribute-value pair as an **elementary equivalence relation** as defined in rough sets (Pawlak 1991). We denote the combination of attribute-value pairs, which is called *the complex of selectors* in terms of AQ theory, by an equivalence relation, R. A set of elements which supports R, which is called a *partial star* in AQ, is referred to as an **indiscernible set**, denoted by $[x]_R$. For example, let $\{1,2,3\}$ be a set of samples which supports an equivalence relation R. Then, we say that a partial star of R is equal to $\{1,2,3\}$ in terms of AQ. This notion can be represented as $[x]_R = \{1,2,3\}$ in terms of rough sets.

Third, when we describe a conjunctive formula, we use the ordinary logical notation. Furthermore, when an equivalence relation is described as attribute-value pair, we denote this pair by $[attribute = value]$. For example, if an equivalence relation R means "a=1 and b=0", then we write it as $R = [a = 1] \bigwedge [b = 0]$.

Finally, we define partial order of equivalence relations as follows:

Definition 1 (Partial Order of Relations) *Let $A(R_i)$ denote the set whose elements are the attribute-value pairs included in R_i. If $A(R_i) \subseteq A(R_j)$, then we represent this relation as:*

$$R_i \preceq R_j.$$

For example, let R_i represent a conjunctive formula, such as $a \wedge b \wedge c$, where a, b, c are elementary equivalence relations. Then $A(R_i)$ is equal to $\{a, b, c\}$. If we use the notation of Michalski's APC(Annotated Predicate Calculus) (Michalski 1983), R_i can be represented as, say $[a = 1] \& [b = 1] \& [c = 1]$, then $A(R_i)$ is equal to a set of selectors, $\{[a = 1], [b = 1], [c = 1]\}$.

2. Matroid Theory

2.1 Definition of a Matroid

Matroid theory abstracts the important characteristics of matrix theory and graph theory, firstly developed by Whitney (Whitney 1935) in the thirties of this century. The advantages of introducing matroid theory are the following: 1)Since matroid theory abstracts graphical structure, this shows the characteristics of formal structure in graph clearly. 2)Since a matroid is defined by the axioms of independent sets, it makes the definition of independent structure clear. 3)The greedy algorithm is one of the algorithms for acquiring an optimal base of a matroid. This algorithm is studied in detail, so we can use well-established results in our problem.

Although there are many interesting and attractive characteristics of matroid theory, for the limitation of space, we only discuss about duality, and the greedy algorithm. For further information, readers might refer to (Welsh 1976).

First, we begin with definition of a matroid. A matroid is defined as an independent space which satisfies the following axioms:

Definition 2 (Definition of a Matroid) *The pair $M(E, \mathcal{J})$ is called a matroid (or an independence space),if*

1) E is a finite set,
2)$\emptyset \in \mathcal{J} \subset 2^E$,
3)$X \in \mathcal{J}, Y \subset X \Rightarrow Y \in \mathcal{J}$,
4) $X, Y \in \mathcal{J}, card(X) = card(Y) + 1 \Rightarrow (\exists a \in X - Y)(Y \cup \{a\}) \in \mathcal{J}$.

If $X \in \mathcal{J}$, it is called **independent**, *otherwise X is called* **dependent**. □

One of the most important characteristic of matroid theory is that this theory refers to the notion of independence using the set-theoretical scheme. As shown in (Pawlak 1991), we also consider the independence of the attributes in terms of rough sets, which uses the set-theoretical framework. Therefore our definition of independence can be also partially discussed using matroid theory, which is discussed later.

2.2 the Greedy Algorithm

Since it is important to calculate a base of a matroid in practice, several methods are proposed. In these methods, we focus on the greedy algorithm. This algorithm can be formulated as follows:

Definition 3 (the Greedy Algorithm) *Let B be a variable to store the calculated base of a matroid, and E denote the whole set of attributes. We define the*

Greedy Algorithm to calculate a base of a matroid as follows:

1. $B \leftarrow \phi$
2.Calculate "priority queue" Q using weight function of E.
3.If B is a base of $M(E, \mathcal{J})$ *then stop. Else go to 4.*
4.$e \leftarrow first(Q)$*, which has a minimum weight in Q.*
5. If $B \cup \{e\} \in \mathcal{J}$ *then* $B \leftarrow B \cup \{e\}$*. goto 2.* □

This algorithm searches one solution which is optimal in terms of one weight function. Note that a matroid may have many bases. The base derived by the greedy algorithm is optimal to some **predefined** weight function. So, for example, when we describe a weight function as a monotonic function of an apparent error rate, the solution is optimal to an apparent rate, that is, in the language of statistics, the algorithm calculates the best class allocation of training samples. Hence if we cannot derive a suitable weight function we cannot get such an optimal base. In the following, we assume that we can define a good weight function for the greedy algorithm, and we discuss about weight functions in Section 4 and Section 5.

Under this assumption, this algorithm has the following characteristics:

Theorem 1 (Computational Complexity) *The complexity of the greedy algorithm is*

$$\mathcal{O}(mf(\rho(M)) + m \log m)$$

where $\rho(M)$ *is equal to a rank of matroid M,m is equal to the number of the elements in the matroid,* $|E|$*, f represents a function of computational complexity of an independent test, which is the procedure to test whether the obtained set is independent, and is called independent test oracle.* □

Theorem 2 (the Optimal Solution) *The optimal solution is derived by this algorithm if and only if a subset of the attributes satisfies the axioms of the matroid.* □

(For the limitation of the space, the proofs of these theorems are not given in this paper. Readers might refer to (Welsh 1976).) This theorem is very important when we discuss about optimal solution of learning algorithms. This point is discussed in section 5.

3. AQ as the Greedy Algorithm

Here we show that our "rough sets" reformulation of AQ algorithm is equivalent to the greedy algorithm for calculating bases of a matroid. Under the above assumption we can constitute a matroid of AQ method, which we call *AQ matroid* as follows:

Theorem 3 (AQ matroid) *Let B denote the base of a matroid such that* $[x]_B = D_k$*. If we define an independent set* $\mathcal{J}(D_k)$ *as* $\{A(R_j)\}$ *which satisfies the following conditions:*

1) $R_j \preceq B$,
2) $[x]_B \subseteq [x]_{R_j}$,
3) $\forall R_i$ *s.t.* $R_i \prec R_j \preceq B$, $D_j = [x]_B \subseteq [x]_{R_j} \subset [x]_{R_i}$,

where the equality holds only if $R_j = B$*. then this set satisfies the definition of a matroid. We call this type of matroid,* $M(E, \mathcal{J}(D_k))$*,AQ matroid.* □

The first condition means that a base is a maximal independent set and each relation forms a subset of this base. And the second condition is the characteristic which satisfies all of these equivalence relations. Finally, the third condition denotes the relationship between the equivalence relations: Any relation R_i which forms a subset of $A(R_j)$ must satisfy $[x]_{R_j} \subset [x]_{R_i}$. Note that these conditions reflect the conditional part of AQ algorithm. For example, let a and b elementary equivalence relations, and let $[x]_a$ and $[x]_b$ be equal to {1,2,3} and {2,3,5}. If the set which supports a target concept is $D_+ = \{2\}$, then $D_+ \subset [x]_{a \wedge b} (= \{2,3\}) \subset [x]_a (= \{1,2,3\})$. Hence $\{a\}$, $\{b\}$ and $\{a,b\}$ belong to the independent sets for the target concept. It is also notable that each D_k has exactly one independent set $J(D_k)$. Therefore the whole AQ algorithm is equivalent to the greedy algorithm for acquiring a set of bases of AQ matroid, denoted by $\{\mathcal{J}(D_k)\}$. Furthermore, since the independent test depends on the calculus of indiscernible sets, is less than $\mathcal{O}(\rho(M) * n^2)$ where n denotes a sample size, the computational complexity is given as follows:

Theorem 4 (Complexity of AQ) *Assume that we do not use constructive generalization. Then the complexity of AQ algorithm is less than*

$$\mathcal{O}(mn^2\rho(M)) + m \log m)$$

where $\rho(M)$ *is equal to a rank of matroid M, m is equal to the number of the elements in the matroid,* $|E|$*.* □

Hence the computational complexity of AQ depends mainly on the number of the elements of a matroid, since it increases exponentially as the number of the attribute-value pairs grows large.

4. Heuristics as Weight Functions

Other rule induction methods, such as C4.5 and CN2, and induction of decision trees, such as ID3, can be described in the framework. The main difference among these methods is what kinds of weight functions are used. Actually, these weight functions are described as functionals $f(\alpha_R(D_+))$ of the accuracy measure, $\alpha_R(D_+)$ which is defined as:

$$\alpha_{R_i}(D_+) = \frac{card\ [x]_{R_i} \bigcap D_+}{card\ [x]_{R_i}}.$$

For example, the information-theoretic entropy measure, which is used in ID3 and CN2, can be rewritten as:

$$\sum_{j=\{+,-\}} -\alpha_{R_i}(D_j) \log_2 \alpha_{R_i}(D_j)$$

Also, the *significant measure*, which is defined in CN2 and can be viewed as a variant of the Kullback-Leibler measure, is also rewritten as:

$$\sum_{j=\{+,-\}} \alpha_{R_i}(D_j) \log_2 \frac{\alpha_{R_i}(D_j)}{e_{R_i}(D_j)}$$

where $e_{R_i}(D_j) = card\ D_j / card\ U$. As mentioned above, the greedy algorithm searches an optimal solution which is exactly optimal to a weight function. Therefore classificatory power, or predictive accuracy strongly depends on this weight function, which seems to be dependent on applied data. We discuss about this issue in the next section.

5. Optimal Solution

As discussed in the above section, when we adopt a weight function which is described as a monotonic function of apparent error rate, we obtain an optimal solution which is the best for apparent error rate. So, in this case, Theorem 3 tells us that an optimal solution is obtained only when relations between training samples and attributes-value pairs satisfy the conditions of AQ matroid.

However, this assumption is very strict, since apparent error rate depends on only given training samples. In practice, it is often violated by new additional training samples. For example, when in the old training samples, $R_i \prec R_j$ implies $[x]_{R_j} \subset [x]_{R_i}$, additional samples cause the latter relation to be $[x]_{R_j} = [x]_{R_i}$. In other words, additional samples cause independent variables to be dependent. In this case, the former derived solution is no longer optimal to this weight function. This problem is also discussed from the viewpoint of predictive error rate $\hat{\alpha}_{R_i}(D)$ defined in the following equation:

$$\begin{aligned} \hat{\alpha}_{R_i}(D_+) &= \frac{card\ \{([x]_{R_i} \bigcap D_+) \bigcup ([x]^c_{R_i} \bigcap D^c_+)\}}{card\ \{[x]_{R_i} \bigcup [x]^c_{R_i}\}} \\ &= \varepsilon_{R_i} \alpha_{R_i}(D_+) + (1 - \varepsilon_{R_i}) \alpha^c_{R_i}(D^c_+) \end{aligned}$$

where ε_{R_i} denotes the ratio of training samples to total population, $\alpha_{R_i}(D_+)$ denotes an apparent accuracy, and $\alpha^c_{R_i}(D_+)$ denotes the accuracy of classification for unobserved cases, $[x]^c_{R_i}$ and D^c_+.

Therefore the value of ε_{R_i} determines whether $\alpha_{R_i}(D)$ is suitable to predictive classification or not. On one hand, if ε_{R_i} is near to 0, then $\hat{\alpha}_{R_i}(D_+)$ may be quite different from $\alpha_{R_i}(D_+)$. So, in this case, an optimal solution based on apparent accuracy is less reliable. On the other hand, if ε_{R_i} is near to 1, then $\hat{\alpha}_{R_i}(D_+)$ may be equal to $\alpha_{R_i}(D_+)$. So, in this case, an optimal solution based on apparent accuracy is much reliable. As shown in the above formula, since ε_{R_i} is dependent on sampling from total population, predictivity depends on sampling from total population. Hence it is a very important factor whether sampling is good or not.

The above formula also suggests that, if we have a weight function which is a monotonic function of predictive error rate, then we derive a base optimal to it. Unfortunately, it is impossible to derive such function, since we can only estimate predictive error rate. Some approaches discuss about these functions, which are known as MDL function (Quinlan & Rivest 1989), and their usefulness is ensured in their papers.

Due to the limitation of the space, we cannot fully discuss about the relation between the solutions by MDL function and matroid theory. Recent research shows that weight function must be satisfied with some important constraints derived by matroid theory and greedoid theory, and that the new concepts of *matroid rigidity* are closely related with these relations. In (Tsumoto 1994), some results on the above relations are partially discussed. To apply fully these facts to our formalization of learning methods will be our future work.

References

Breiman, L., et al. 1984. *Classification And Regression Trees.* Belmont, CA: Wadsworth International Group.

Clark, P., Niblett, T. 1989. The CN2 Induction Algorithm. *Machine Learning*, **3**,261-283.

Michalski, R.S. 1983. A Theory and Methodology of Machine Learning. Michalski, R.S., Carbonell, J.G. and Mitchell, T.M., *Machine Learning - An Artificial Intelligence Approach*, 83-134, Morgan Kaufmann, CA.

Michalski, R.S., et al. 1986. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In:*Proceedings of AAAI-86*, 1041-1045, Morgan Kaufmann, CA.

Pawlak, Z. 1991. *Rough Sets*, Kluwer Academic Publishers, Dordrecht.

Quinlan, J.R. 1986. Induction of decision trees, *Machine Learning*, **1**, 81-106.

Quinlan, J.R. and Rivest, R.L. 1989. Inferring Decision Trees Using the Minimum Description Length Principle, *Information and Computation*, **80**, 227-248.

Quinlan, J.R. 1993. C4.5 - Programs for Machine Learning, Morgan Kaufmann, CA.

Tsumoto, S. and Tanaka, H. 1994. Algebraic Specification of Empirical Inductive Learning Methods based on Rough Sets and Matroid Theory. In:*Proceedings of the second Conference on Artificial Intelligence and Symbolic Mathematical Computing.*

Welsh, D.J.A. 1976. *Matroid Theory*, Academic Press, London.

Whitney, H. 1935. On the abstract properties of linear dependence, *Am. J. Math.*, **57**, 509-533.