Machine Learning, 41, 175–195, 2000

A Formalism for Relevance and Its Application in Feature Subset Selection

DAVID A. BELL da.bell@ulst.ac.uk
HUI WANG h.wang@ulst.ac.uk
School of Information and Software Engineering, Faculty of Informatics, University of Ulster at Jordanstown, Shore Road, Newtownabbey, BT37 0QB, Northern Ireland

Editor: William W. Cohen

Abstract. The notion of relevance is used in many technical fields. In the areas of machine learning and data mining, for example, relevance is frequently used as a measure in feature subset selection (FSS). In previous studies, the interpretation of relevance has varied and its connection to FSS has been loose. In this paper a rigorous mathematical formalism is proposed for relevance, which is quantitative and normalized. To apply the formalism in FSS, a characterization is proposed for FSS: preservation of learning information and minimization of joint entropy. Based on the characterization, a tight connection between relevance and FSS is established: maximizing the relevance of features to the decision attribute, and the relevance of the decision attribute to the features. This connection is then used to design an algorithm for FSS. The algorithm is linear in the number of instances and quadratic in the number of features. The algorithm is evaluated using 23 public datasets, resulting in an improvement in prediction accuracy on 16 datasets, and a loss in accuracy on only 1 dataset. This provides evidence that both the formalism and its connection to FSS are sound.

Keywords: machine learning, knowledge discovery, data mining, relevance, feature subset selection

1. Introduction

There has been an interest in explicitly studying and using *relevance* in a wide range of areas; in particular, in the area of machine learning for feature subset selection (FSS) (Greiner & Subramanian, 1994). But what do we mean by the term *relevance*? Relevance has been studied for over fifty years (Keynes, 1921), and it has an agreed, commonsense meaning to do with the relationships between objects. We believe that it can be given a more rigorous definition, and that this definition can be useful in FSS.

Broadly speaking, the purpose of FSS is to select a subset of features from the feature space which is *good* enough regarding its ability to describe the training dataset and to predict for future cases. There is a wealth of algorithms for FSS (Littlestone, 1988; Almuallim & Dietterich, 1991; Kira & Rendell, 1992; Aha & Bankert, 1994; Caruana & Freitag, 1994; Kononenko, 1994; John, Kohavi, & Pfleger, 1994; Skalak, 1994; Kohavi & Sommerfield, 1995; Kononenko, Simec, & Robnik-Sikonja, 1997; Liu & Setiono, 1997). With regard to how to evaluate the goodness (quality) of a subset of features, the FSS methods fall into two broad categories: the *filter approach* and the *wrapper approach*. In the filter approach, a good feature set is selected as a result of pre-processing based on properties of the data itself and independent of the learning algorithm. In the wrapper approach, FSS is done with

the help of learning algorithms. The FSS algorithm conducts a search for a good feature set using the learning algorithm itself as part of the evaluation function. Typically, the feature subset which performs best for the learning algorithm will be selected.

FSS has a traditional close link with the notion of relevance. For example, FOCUS (Almuallim & Dietterich, 1991), RELIEFF (Kira & Rendell, 1992) and Schlimmer's model (Schlimmer, 1993) use "relevance" to estimate the goodness of the feature subset in one way or another. Although the wrapper approach does not use a relevance measure directly, it is shown by Kohavi and Sommerfield (1995) that the "optimal" feature subset obtained this way must be from the relevant feature set (strongly relevant and weakly relevant features). This can be seen from the following overview.

1.1. The use of relevance in FSS

Although the notion of relevance is used by many FSS algorithms, the measure of relevance varies and the relationship between relevance and FSS is of an intuitive nature. In this section we describe the use of relevance in FSS briefly. In the filter category we look at RELIEFF (Kira & Rendell, 1992), FOCUS (Almuallim & Dietterich, 1991; Almuallim & Dietterich, 1994) and Schlimmer's approach (Schlimmer, 1993), and in the wrapper category we look at the work by John, Kohavi, and Pfleger (1994).

In RELIEFF, a good subset of features is one where each feature has a "relevance level" greater than a given threshold. The notion of relevance[1] has not been rigorously justified against the agreed common understanding of this notion. The algorithm works by associating with each feature a weight indicating the relevance level of that feature to the decision attribute and returns a set of features whose weights exceed a threshold.

In RELIEFF, the selected feature subset is highly dependent on the user-specified threshold. It is clear that a threshold value which works well for one dataset doesn't necessarily work well in another.

In FOCUS a good feature subset is a minimal subset which is consistent with the training dataset. Minimality is in the sense of set cardinality, and consistency is in the sense that the selected feature subset preserves the dependence in the original dataset. Such a feature subset is then regarded as relevant.

In Schlimmer's approach, a good subset is one of the minimal determinations[2] which is then regarded as relevant, but nothing is mentioned as to which one is *optimal*. The algorithm carries out a systematic search through the space of feature subsets for all minimal determinations (not just one) which are consistent with the training dataset. The feature subset selected by FOCUS is a determination, while Schlimmer's approach aims to find all determinations.

John, Kohavi, and Pfleger (1994) were the first to present the wrapper idea as a general framework for FSS. The generic wrapper technique uses some measure to select among alternative features. One natural scheme involves running the learning algorithm over the training dataset using a given set of features, then measuring the accuracy of the learned structure on the testing dataset.

The wrapper approach does not use a relevance measure directly; rather, it uses the accuracy obtained by applying a learning algorithm as the measure for the goodness of

feature subsets. However, Kohavi and Sommerfield (1995) show that the optimal feature subset obtained this way must be from the relevant feature subset (strongly relevant and weakly relevant features).

1.2. Our approach

From the above discussion we can see that the notion of relevance has been used extensively in FSS literature, but its interpretation varies in different cases. We start off from a common understanding of this notion, develop a rigorous and quantified formalism for it, use the formalism to establish a tight connection between relevance and FSS, and finally develop an algorithm for FSS making use of our relevance concept. Our study on FSS in this paper is of the filter type.

In the rest of the paper, we first present a brief review of relevance; then we introduce our information theoretic formalism of relevance, which is justified with regard to the axiomatization of relevance. To demonstrate the usefulness of this relevance formalism, we apply it to the problem of FSS in the areas of machine learning and data mining. We characterize FSS with two requirements, and based on the characterization we formally establish the quantitative relationship between FSS and relevance. This is then used to develop an algorithm for FSS based on the relevance formalism. The algorithm is evaluated using public datasets.

2. A brief review of relevance

There are basically two current lines of research on relevance in the context of AI: the formalization of the commonsense notion of relevance, and the problem oriented characterization of relevance. This review is presented along these two lines.

2.1. Formalization of the commonsense notion of relevance

The notion of relevance has been formally investigated in the philosophy literature (Keynes, 1921; Carnap, 1962; Gärdenfors, 1978). The focus of the discussion was on formalizing a concept of relevance that would fit the commonsense notion of the word. There is a widely agreed traditional understanding in broad terms of this concept, expressed by Gärdenfors as follows:

Definition 2.1 (Gärdenfors, 1978). On the basis of prior evidence e, a hypothesis h is considered, and the change in the likelihood of h due to additional evidence i is examined. If the likelihood of h is changed by the addition of i to e, i is said to be relevant to h on the evidence e; otherwise it is irrelevant. In particular, if the likelihood of h is increased due to the addition of i to e, i is said to be positively relevant to h; if the likelihood is decreased, i is said to be negatively relevant.

However investigators of matters pertaining to the concept differ in the formal technical details, and in how to define the *relevance function* in attempts to quantify the concept.

These differences have led to arguments as to which can best capture the commonsense notion of relevance from first principles.

Keynes and Carnap each have their own definitions of relevance function. Keynes uses the relevance quotient (Keynes, 1921), and Carnap uses the probability difference (Carnap, 1962). Following the work of Keynes and Carnap, Gärdenfors (1978) proposes a set of six axioms (or in Gärdenfors' terms, logical conditions), which he argues should be observed by an appropriate definition of relevance with regard to the commonsense meaning of the concept. Gärdenfors has shown that both Keynes' and Carnap's relevance functions satisfy the first five axioms, but not the sixth. He then replaces the traditional definition with a stronger one that yields a relevance relation which, in his opinion, agrees well with the commonsense term (Gärdenfors, 1978). He further shows that this definition satisfies all six axioms above.

2.2. Conditional independence

Conditional independence among variables finds its application in belief networks (Pearl, 1988), where irrelevance is identified with conditional independence, and relevance is identified with the negation of irrelevance (Lakemeyer, 1995). Irrelevance is in fact the basis of constructing belief networks.

The study of conditional independence is summarized by Pearl (1988), where a number of properties are identified. Here we re-state these properties as a set of axioms.

Axiom 2.1 (Pearl, 1988). Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be three disjoint sets of variables. If $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y})$ stands for the relation "$\mathcal{X}$ is independent of $\mathcal{Y}$ given $\mathcal{Z}$" in some probability distribution p, then "$\Im$" must satisfy the following four independent conditions:

- Symmetry: $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y}) \Longleftrightarrow \Im(\mathcal{Y}, \mathcal{Z}, \mathcal{X})$;
- Decomposition: $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y} \cup \mathcal{W}) \Rightarrow \Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y})$ and $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{W})$;
- Weak union: $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y} \cup \mathcal{W}) \Rightarrow \Im(\mathcal{X}, \mathcal{Z} \cup \mathcal{W}, \mathcal{Y})$;
- Contraction: $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y})$ and $\Im(\mathcal{X}, \mathcal{Z} \cup \mathcal{Y}, \mathcal{W}) \Rightarrow \Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y} \cup \mathcal{W})$;

2.3. Machine learning

There are many contributions in this domain (Blum, 1994; Almuallim & Dietterich, 1991; Gennari, Langley, & Fisher, 1989; John, Kohavi, & Pfleger, 1994). The definitions identify a feature as either relevant or irrelevant to a concept or task. John, Kohavi, and Pfleger (1994) and Kohavi (1994) show that these definitions give unexpected results, and that the dichotomy of relevance vs irrelevance is not enough. An alternative definition of relevance is then proposed which distinguishes between *strong relevance* and *weak relevance*.

Definition 2.2 (John, Kohavi, & Pfleger, 1994; Kohavi, 1994). Let $\mathcal{S}_i$ be the set of all features except X_i, i.e., $\mathcal{S}_i = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m\}$, and let Y be a decision attribute not in $\mathcal{S}_i$. Denote by s_i a value-assignment to all features in $\mathcal{S}_i$. Then X_i is *strongly*

relevant iff there exist some x_i, y, and s_i with $P(X_i = x_i, \mathcal{S}_i = s_i) > 0$ such that

$$P(Y = y \mid \mathcal{S}_i = s_i, X_i = x_i) \neq P(Y = y \mid \mathcal{S}_i = s_i).$$

A feature X_i is *weakly relevant iff* it is not strongly relevant, and there exists a subset of features $\mathcal{S}'_i$ of $\mathcal{S}_i$ for which there exists some x_i, y, and s'_i with $P(X_i = x_i, \mathcal{S}'_i = s'_i) > 0$ such that

$$P(Y = y \mid X_i = x_i, \mathcal{S}'_i = s'_i) \neq P(Y = y \mid \mathcal{S}'_i = s'_i)$$

Under this definition, X_i is strongly relevant if the probability of the outcome (given all features) can change when we eliminate knowledge about the value of X_i. Strong relevance implies that a given feature is *strongly relevant* if it is indispensable in the sense that it cannot be removed without loss of prediction accuracy. A feature is *weakly relevant* if it can sometimes contribute to prediction accuracy.

These definitions are all qualitative by nature and are concerned only with relevance among features (or attributes, variables). The distinction of strong relevance and weak relevance (John, Kohavi, & Pfleger, 1994; Kohavi, 1994) has the advantage of flexibility: using this distinction, we can select either strongly relevant features or weakly relevant features to satisfy different learning requirements (Kohavi & Sommerfield, 1995). Taking this one step further, it is then reasonable to expect a finer distinction of relevance, and even further, a quantitative concept of relevance.

2.4. Our objective

We have briefly reviewed some work on relevance, and in particular, we have identified two axiomatic characterizations of relevance (Gärdenfors, 1978; Pearl, 1988) and the underlying definitions. Clearly there could be other treatments of relevance and other characterizations.[3] These studies are rooted in different subject areas and they capture the broad meaning of the commonsense notion of relevance (in Definition 2.1) in their respective contexts.

Keynes', Carnap's and Gärdenfors' measures for relevance do not apply directly in FSS, but they serve as a basis for further study. Conditional independence and strong/weak relevance are applicable in FSS, but they are largely of a qualitative nature. Our objective is to establish a quantitative relationship between relevance and FSS in the sense that, ideally, a certain degree of relevance corresponds to a certain degree of improvement in learning performance as a result of FSS. For this, we want to find a quantitative formalism which complies with the commonsense meaning of relevance and is usable for the purpose of FSS and possibly in other domains.

3. Information theoretic formalism of relevance

Given a set of discrete variables and their joint probability distribution, we can examine the relevance relationship between variables, as well as between instances of variables. The

former is referred to as *variable relevance*, and the latter as *instance relevance*. We stick to the commonsense meaning of relevance in Definition 2.1.

We will be using the following notational convention throughout this paper. Capital letters $X, Y, Z, \ldots$ are used to represent variables, lower case letters to represent instances of variables, and calligraphic letters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$ to represent sets. A single variable is regarded as a singleton set of variables. Exceptions will be noted explicitly.

3.1. *Variable relevance*

The relevance of one variable to another (target) variable is here understood in information theoretic terms, as the *mutual information between the two variables relative to the entropy of the target variable*, or in other words, the relative reduction of entropy (uncertainty) of one variable due to the knowledge of another. The bigger the reduction, the higher the relevance. Formally we have:

Definition 3.1. Given three sets of variables $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ with a joint probability distribution p, let $I(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$ be the mutual information between $\mathcal{X}$ and $\mathcal{Y}$ given $\mathcal{Z}$, and let $H(\mathcal{X} \mid \mathcal{Y})$ be the entropy of $\mathcal{X}$ given $\mathcal{Y}$.[4] If $H(\mathcal{Y} \mid \mathcal{Z}) \neq 0$, then the *variable relevance* of $\mathcal{X}$ to $\mathcal{Y}$ given $\mathcal{Z}$, denoted $r_p(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$, is defined as

$$r_p(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = \frac{I(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})}{H(\mathcal{Y} \mid \mathcal{Z})} = \frac{H(\mathcal{Y} \mid \mathcal{Z}) - H(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z})}{H(\mathcal{Y} \mid \mathcal{Z})}$$

If $H(\mathcal{Y} \mid \mathcal{Z}) = 0$, then $r_p(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$.

Where there is no ambiguity, p will be dropped for brevity.

This definition says that the relevance of $\mathcal{X}$ to $\mathcal{Y}$ given $\mathcal{Z}$ is indicated by the relative reduction of uncertainty of $\mathcal{Y}$ when $\mathcal{X}$ and $\mathcal{Z}$ are known. With this notion we can express a degree of relevance by stating that $\mathcal{X}$ is relevant to $\mathcal{Y}$ given $\mathcal{Z}$ with degree $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$. This is the *conditional case* in the sense the relevance between $\mathcal{X}$ and $\mathcal{Y}$ is conditioned by $\mathcal{Z}$, and $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$ is therefore called *conditional relevance*. When $\mathcal{Z}$ is dropped, the relevance between two variables is not conditioned by any other variable, therefore this is the *unconditional case*, and $r(\mathcal{X}; \mathcal{Y})$ is called *unconditional relevance*.

Example 3.1. This example is taken from Cover & Thomas (1991). Let $(\mathcal{X}, \mathcal{Y})$ have the following joint distribution:

$\mathcal{Y} \setminus \mathcal{X}$	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

By definition we have $r(\mathcal{X}; \mathcal{Y}) = 0.187500$, $r(\mathcal{Y}; \mathcal{X}) = 0.214286$.

Theorem 3.1. *The following properties for variable relevance follow Definition* 3.1 *and the properties of mutual information and entropy* (*Cover & Thomas*, *1991*).

- *Uniformity*: $0 \leq r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) \leq 1$. *That is, the relevance value lies between two fixed extremes.*
- *Reflexiveness*: $r(\mathcal{X}; \mathcal{X} \mid \mathcal{Z}) = 1$. *If* $\mathcal{Y} \subseteq \mathcal{X}$, *then* $r(\mathcal{X}; \mathcal{Y}) = 1$.
- *Weak symmetry*: $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) > 0 \iff r(\mathcal{Y}; \mathcal{X} \mid \mathcal{Z}) > 0$. *But in general*, $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) \neq r(\mathcal{Y}; \mathcal{X} \mid \mathcal{Z})$.
- *Monotonicity*: *If* $\Sigma \subseteq \Omega$, *then* $r(\Omega; \mathcal{Y}) \geq r(\Sigma; \mathcal{Y})$.
- *Intransitivity*: *In general* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) > 0$ *and* $r(\mathcal{Y}; \mathcal{W} \mid \mathcal{Z}) > 0 \not\Rightarrow r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z}) > 0$, *and* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$ *and* $r(\mathcal{Y}; \mathcal{W} \mid \mathcal{Z}) = 0 \not\Rightarrow r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z}) = 0$.
- *Saturizability*: *If* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 1$, *then* $r(\mathcal{W}; \mathcal{Y} \mid \mathcal{Z}, \mathcal{X}) = 0$, *where* $\mathcal{W}$ *is any set of variables.*

Theorem 3.2 (*Dependence vs independence*)**.** *Suppose* $\mathcal{X}$, $\mathcal{Y}$, *and* $\mathcal{Z}$ *are three sets of variables with a joint distribution* p. *Then* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 1 \iff \mathcal{Y}$ *is conditionally fully dependent on* $\mathcal{X}$ *given* $\mathcal{Z}$; $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0 \iff \mathcal{Y}$ *is conditionally independent of* $\mathcal{X}$ *given* $\mathcal{Z}$.

The proof is in the Appendix.

This theorem shows that the definition of variable relevance agrees with two extreme cases of probabilistic dependence: full dependence and full independence. In other words, two extreme cases can be identified by our variable relevance measure: 0 for extreme irrelevance (conditional independence) and 1 for extreme relevance (full dependence).

Theorem 3.3 (*Chain rule of variable relevance*)**.** *Given three sets of variables* $\mathcal{X}$, $\mathcal{Y}$ *and* $\mathcal{Z}$,

$$\begin{aligned} r(\mathcal{X}, \mathcal{Y}; \mathcal{Z}) &= r(\mathcal{X}; \mathcal{Z}) + r(\mathcal{Y}; \mathcal{Z} \mid \mathcal{X}) - r(\mathcal{X}; \mathcal{Z}) \times r(\mathcal{Y}; \mathcal{Z} \mid \mathcal{X}) \\ &= r(\mathcal{Y}; \mathcal{Z}) + r(\mathcal{X}; \mathcal{Z} \mid \mathcal{Y}) - r(\mathcal{Y}; \mathcal{Z}) \times r(\mathcal{X}; \mathcal{Z} \mid \mathcal{Y}). \end{aligned}$$

The proof is in the Appendix. This theorem establishes the relationship between the relevance of a set of variables and the relevance of its subsets.

The following theorem shows that variable relevance satisfies Pearl's axiomatic characterization (Axiom 2.1 above).

Theorem 3.4. *Identifying Pearl's notation* $\Im(\mathcal{X}, \mathcal{Z}, \mathcal{Y})$ *as* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$, *we can reproduce Pearl's conclusions in Axiom* 2.1. *In other words, the extreme irrelevance in the above definition, i.e.,* $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$ *satisfies Pearl's axioms as in Axiom* 2.1. *Specifically, let* $\mathcal{X}$, $\mathcal{Y}$, *and* $\mathcal{Z}$ *be three disjoint sets of variables. Then we have*

- *Symmetry*: $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0 \iff r(\mathcal{Y}; \mathcal{X} \mid \mathcal{Z}) = 0$;
- *Decomposition*: $r(\mathcal{X}; \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}) = 0 \Rightarrow r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$ *and* $r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z}) = 0$;

– *Weak union*: $r(\mathcal{X}; \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}) = 0 \Rightarrow r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z} \cup \mathcal{W}) = 0$.
– *Contraction*: $r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0$ *and* $r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z} \cup \mathcal{Y}) = 0 \Rightarrow r(\mathcal{X}; \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}) = 0$.

The proof is in the Appendix.

3.2. *Instance relevance and event relevance*

Given two (sets of) variables and the joint distribution, instance relevance concerns the relationship between instances of one variable and instances of the other: for any two instances x and y of the two variables X and Y respectively, the relevance of x to y concerns the relative change of likelihood of y when x is known. Instance relevance is defined similarly to variable relevance using mutual information and entropy. It has been shown (Wang, 1996) that variable relevance is the averaged instance relevance across all possible pairs of instances, and that instance relevance satisfies Subramanian's axiomatic characterization of irrelevance (Subramanian & Genesereth, 1987).

Event relevance is another type of relevance definable in terms of instance relevance. Given a set of objects, we can consider the relevance between events, which are subsets of objects. We have two reasons for this. Firstly, it has logical implications. When A is an event, we can talk about the event happening (A) as well as not happening ($\bar{A}$). Secondly, some existing and important work (Keynes, 1921; Carnap, 1962; Gärdenfors, 1978) falls into this category. It has been shown that, similar to Keynes' and Carnap's relevance formalisms, event relevance satisfies 5 of Gärdenfors' 6 axioms (Gärdenfors, 1978).[5]

3.3. *Discussion*

In this section we have discussed variable relevance in detail and have briefly introduced instance relevance and event relevance, since we believe that variable relevance is more useful in machine learning than the other two types. All these types of relevance are defined in a uniform way: mutual information between variables (instances, events) relative to entropy of target variable (instance, event). We have shown that variable relevance satisfies Pearl's axiomatic characterization of dependence among variables (Theorem 3.4). It is also shown by Wang (1996) that instance relevance satisfies Subramanian's axiomatic characterization of irrelevance (Subramanian & Genesereth, 1987), and that event relevance satisfies Gärdenfors' relevance axioms (except one) (Wang, 1996). This implies that our uniform approach to relevance is sound.

Strong and weak relevance as in Definition 2.2 can be characterized in terms of our relevance formalism as follows. From Theorem 3.2, we know that with the inequalities in the definition, the variable relevance is greater than zero. Then Definition 2.2 can be re-stated in terms of variable relevance as follows: X_i *is strongly relevant if* $r(X_i; Y \mid \mathcal{S}_i) > 0 \Rightarrow r(X_i; Y) > 0$;[6] *and* X_i *is weakly relevant if* $r(X_i; Y \mid \mathcal{S}'_i) > 0$ *for a subset* $\mathcal{S}'_i$ *of* $\mathcal{S}_i$. Hence a feature X_i is irrelevant to Y if $r(X_i; Y \mid \mathcal{S}'_i) = 0$ for any subset $\mathcal{S}'_i$ of $\mathcal{S}_i$. Therefore $r(\Pi; Y) > 0$ where Π is the feature subset selected by the wrapper approach due to the monotonicity of variable relevance, as well as the fact that all features are in the relevant set (Kohavi & Sommerfield, 1995).

4. Application of relevance in FSS

In this section we attempt to characterize FSS in order to establish a tight connection between relevance and FSS. Based on this we devise an algorithm for FSS, which is evaluated using real world data.

4.1. Characterization of FSS

We first of all define the problem of FSS in the context of machine learning. Let $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ be a set of features, and let Y be a decision attribute. Let $d(X_1), d(X_2), \ldots, d(X_n)$ and $d(Y)$ be their respective domains. For $Q \subseteq \mathcal{X} \cup \{Y\}$, let $\mathcal{U}(Q) = \prod_{x \in Q} d(x)$, and $D(Q)$ denote a subset of $\mathcal{U}(Q)$. $D(Q)$ can be understood as a database relation (Ullman, 1989) except that no key is required.

The input to a supervised learning algorithm is a dataset $D(\mathcal{X}, Y)$ where $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$. The task of learning is to induce a structure (e.g., a decision tree, a neural network) such that, given $t \in \mathcal{U}(\mathcal{X})$, it is possible to accurately predict a label $y \in d(Y)$ for t.

The problem of FSS is then to search for a subset Π of $\mathcal{X}$ which not only performs well on the training dataset, but also predicts well on unseen new cases—this is what we mean when we say a subset of features is good enough. Our objective in this section is to characterize an optimal feature subset, from first principles and some known principles.

4.1.1. The preservation of learning information. Given a dataset $D(\mathcal{X}, Y)$, the learning task is to characterize the hidden relationship between $\mathcal{X}$ and Y so that this relationship can be used to predict on future cases (either one in the dataset or a new case). We call this relationship *learning information*. A natural measure of this relationship is the mutual information (Cover & Thomas, 1991). Specifically, given a dataset $D(\mathcal{X}, Y)$, the learning information is the mutual information $I(\mathcal{X}; Y)$.

A feature subset may lose learning information. Therefore any good feature subset should preserve the learning information in the dataset. For $\Pi \subseteq \mathcal{X}$, if $I(\Pi; Y) = I(\mathcal{X}; Y)$, then Π is said to have preserved the learning information. This leads to the following definition.

Definition 4.1. $\Pi \subseteq \mathcal{X}$ is a *sufficient feature subset*, or *SFS* for short, if and only if $I(\Pi; Y) = I(\mathcal{X}; Y)$.

Therefore a basic requirement for FSS is that any selected feature subset should be an SFS.

From information theory (Cover & Thomas, 1991) we know that for any $\Pi \subseteq \mathcal{X}$, $I(\Pi; Y) \leq I(\mathcal{X}; Y)$, and equality holds if Π is an SFS. From this and the additivity of mutual information we know that given an SFS, Π, removing all of the other features, called collectively Σ ($\Sigma = \mathcal{X} \backslash \Pi$), will not lose learning information contained in the original dataset. In other words, Y is conditionally independent of Σ given Π, when Π is an SFS. Having such a Π, any superset, Σ, of Π is also an SFS. This property helps in determining SFSs without having to calculate the mutual information. This property is exploited in the design of an FSS algorithm later.

4.1.2. Minimization of joint entropy: Occam's razor. Given a dataset, there may be a number of SFSs. However they may not all be equally good for prediction. An optimal feature subset should perform best in prediction. However it is not easy to determine which of two subsets of features predicts better without seeing how they actually perform, since the future is rather misty. What we can do is to focus on the training dataset itself and then apply some empirical principles. There are a number of empirical principles: Occam's razor (Wolpert, 1990), the maximum entropy principle (Shore & Johnson, 1980), the minimum description length (Rissanen, 1986; Quinlan & Rivest, 1989), the minimum message length (Wallace & Freeman, 1987) and the relative least general generalization principle (Muggleton, 1992). Here we use the Occam's razor principle.

Occam's razor, also known as *the principle of parsimony*, is a tool that has application in many areas of science, and it has been incorporated into the methodology of experimental science. It is also becoming influential in machine learning, where it can be formulated as: given two hypotheses that are both consistent with a training set of examples of a given task, the simpler one should perform better on future examples of this task (Blumer et al., 1987; Wolpert, 1990). It has been shown (Blumer et al., 1987) that, under very general assumptions, Occam's razor produces hypotheses that with high probability will be predictive of future cases.

One basic question concerns the meaning of "simplicity", usually called *Occam simplicity*. Typically Occam simplicity is associated with the difficulty of implementing a given task, viz. *complexity of implementation*. Examples of complexity measures are: the number of hidden neurons in neural networks (Amirikian and Nishimura, 1994); the number of leaf nodes of a decision tree (Fayyad & Irani, 1990; Fayyad & Irani, 1992); the minimum description length (MDL) (Rissanen, 1986; Quinlan and Rivest, 1989); and the encoding length (Schweitzer, 1995). The first 3 are all model-dependent while the fourth is model independent. Since we are looking at FSS independently of any learning model, we choose to use encoding length. In particular we use Shannon's entropy as the encoding length measure.

Using entropy as the Occam simplicity measure in our context, we then have: given a dataset $D(\mathcal{X}, Y)$, Occam's razor dictates the selection of an SFS, Π, which minimizes $H(\Pi, Y)$, where H is Shannon's entropy function. This leads to the following definition.

Definition 4.2. $\Pi \subseteq \mathcal{X}$ is an *occam-optimal feature subset* if and only if Π is an SFS and there is no Σ such that $H(\Sigma, Y) < H(\Pi, Y)$. In other words, Π is an SFS that minimizes the joint entropy of the features and the decision attribute.

The following lemma shows that the feature subset Π which minimizes the joint entropy minimizes the marginal entropy.

Lemma 4.1. *Given a dataset* $D(\mathcal{X}, Y)$, *consider two SFSs* $\Pi, \Sigma \subseteq \mathcal{X}$. $H(\Pi, Y) \leq H(\Sigma, Y) \Longleftrightarrow H(\Pi) \leq H(\Sigma)$.

Proof: Since both Π and Σ are SFSs, by definition we have $I(\Pi; Y) = I(\Sigma; Y) = I(\mathcal{X}; Y)$. So $H(Y) - H(Y \mid \Pi) = H(Y) - H(Y \mid \Sigma) \Longleftrightarrow H(Y \mid \Pi) = H(Y \mid \Sigma)$. Furthermore we have $H(\Pi) \leq H(\Sigma) \Longleftrightarrow H(\Pi) + H(Y \mid \Pi) \leq H(\Sigma) + H(Y \mid \Sigma) \Longleftrightarrow H(\Pi, Y) \leq H(\Sigma, Y)$. □

According to this lemma, an occam-optimal feature subset would be the sufficient one which has the least marginal entropy.

4.1.3. Occam-optimal feature subset maximizes relevance. In the previous two sections we have derived two characterizations of FSS: preservation of mutual information, and minimization of joint entropy. In this section we are going to show the above two characterizations can be re-stated in terms of relevance in an even more concise form.

From the definition of relevance in Section 3, we have $I(\Pi; Y) = I(\mathcal{X}; Y) \Longleftrightarrow r(\Pi; Y) = r(\mathcal{X}; Y)$ for any $\Pi \subseteq \mathcal{X}$. So preserving learning information amounts to preserving the relevance relationship. Since $r(\Pi; Y) \leq r(\mathcal{X}; Y)$ in general (due to the fact that $I(\Pi; Y) \leq I(\mathcal{X}; Y)$), any Π which preserves learning information in fact also maximizes the relevance $r(\mathcal{X}; Y)$.

Let Π and Σ be SFSs. Since, by definition, $I(\Pi; Y) = I(\Sigma; Y) = I(\mathcal{X}; Y)$, we have $H(\Pi, Y) \leq H(\Sigma, Y) \iff H(\Pi) \leq H(\Sigma) \iff I(\Pi; Y)/H(\Pi) \geq I(\Sigma; Y)/H(\Sigma) \iff r(Y; \Pi) \geq r(Y; \Sigma)$. Therefore, in conjunction with the previous requirement, an occam-optimal feature subset, Π, would be the SFS which maximizes the relevance $r(Y; \mathcal{X})$.

Summarizing the above discussion we have the following theorem:

Theorem 4.1. *Given a dataset $D(\mathcal{X}, Y)$, $\Pi \subseteq \mathcal{X}$ is an occam-optimal feature subset if and only if Π maximizes both $r(\Pi; Y)$ and $r(Y; \Pi)$.*

This theorem formalizes the intuitive connection between relevance and FSS.

4.2. A relevance-based algorithm for FSS

From the previous section we know that an occam-optimal feature subset should be such that it is a sufficient subset of features and it has the highest $r(Y; \mathcal{X})$ relevance value. A straightforward algorithm based on this will systematically examine all feature subsets and find one which satisfies Theorem 4.1. Unfortunately, as shown by Davies and Russell (1994), the class of problems, which examines all subsets of features to search for the one which satisfies some optimal conditions, turns out to be NP-hard. So we should be satisfied with a heuristic algorithm.

In this section we present a simple and usable heuristic FSS algorithm which is based on the characterization in the previous section. Instead of evaluating each possible subset of features, our approach endeavors to build one incrementally: initially, we have an empty feature subset; then we gradually add in features one by one; this process continues until the relevance of this feature subset to the decision attribute saturates (usually with value 1).

The question is, how to select the feature to be added at each stage? Our objective is to find an SFS, Π, such that $r(Y; \Pi)$ is maximal among all possible SFSs. This is the globally occam-optimal solution, which is NP-hard. Then we turn to find locally occam-optimal solutions at each stage, which as a whole leads to a quasi-optimal solution at last. Specifically, let Π_i be the current feature subset, and Σ_i be the current set of features that are not in Π_i. We calculate $r(Y; \Pi_i \cup \{X\})$ for every $X \in \Sigma_i$ and let Π_{i+1} be $\Pi_i \cup \{X_0\}$

where X_0 is such that $r(Y; \Pi_i \cup \{X_0\})$ is largest. This process is repeated until $r(\Pi_i; Y)$ saturates.

Based on the above discussion, we design the following FSS algorithm.

Algorithm 4.1 (RELFSS**:** FSS based on relevance)**.** *Given a dataset* $D(\mathcal{X}, Y)$, *where* $|\mathcal{X}| = N$,

```
Π = {}, and Σ = 𝒳;
Repeat:
    Let X ∈ Σ be such that r(Y; Π ∪ {X}) is largest for
                                        all elements in Σ.
    Let Π = Π ∪ {X} and Σ = Σ\{X}.
Until r(Π; Y) = r(𝒳; Y)
Return Π.
```

From the monotonicity of relevance, $r(\Pi; Y) = r(\mathcal{X}; Y)$ is certain to happen and Π is then an SFS. After this, no other features can improve the relevance strength.

Now we look at the time complexity of the algorithm. The algorithm iterates from $k = N$ down to $k = k_0$, where k is the number of features in Σ. Suppose there are m instances in the training dataset. At iteration k, we need to compute the relevance $r(Y; \Pi \cup \{X\})$ for all k features, hence we have a step complexity of $O(mk)$. To find the feature with largest relevance value, we need $k-1$ comparisons, hence a complexity of $O(k-1)$. In the worst case we need to iterate from $k = N$ to $k = 1$, hence the complexity is $\sum_{k=N}^{1}(mk+k-1) = O(mN^2)$. Therefore the overall complexity for the above algorithm is $O(mN^2)$.

The Π selected by RELFSS is guaranteed to be an SFS. It is also guaranteed to be locally occam-optimal at each stage in the sense that $r(Y; \Pi)$ is locally maximal, but it is not guaranteed to be globally occam-optimal across all stages.

In our implementation, we assume a uniform distribution on the tuples. The distribution on individual features are projections of the uniform distribution. In the case of discrete features, we use $H(X) = -\sum_{x \in d(X)} p(x) \log p(x)$. In the case of continuous features, we use $h(X) = H(X^{\triangle}) - \log m$, where $X^{\triangle}$ is the quantized X in such a way that X is simply treated as discrete, and m is the number of distinct values of $X^{\triangle}$.

Modifying the stopping condition of RELFSS as "until Σ is empty", we can get an ordering of the features. The selected feature subset is in fact the collection of the first k features in the ordering. By adding or removing one or two features in this ordering and evaluating the new feature subset by a learning algorithm, we may sometimes get a better feature subset. This is a kind of wrapper/filter hybrid approach.

4.3. An example

In this Section we are going to illustrate the RELFSS algorithm using the Heart dataset. This dataset has a mixture of discrete and continuous features: 4 discrete and 9 continuous.

There are 14 features in this dataset, $X_1, X_2, \ldots, X_{14}$[7] where $Y \stackrel{\text{def}}{=} X_{14}$ is decision attribute. At the first level, Π is empty and we calculate $r(Y; X_i)$ $(i = 1, 2, \ldots, 13)$ resulting in

the following:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
0.03	0.07	0.11	0.03	0.09	0.00	0.02	0.06	0.15	0.06	0.09	0.11	0.18

Since $r(Y; X_{13})$ is maximal, X_{13} is selected and added into Π. At the second level, we calculate $r(Y; \Pi \cup \{X_i\})$ resulting in the following:

X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
0.08	0.12	0.13	0.08	0.11	0.12	0.11	0.09	0.14	0.09	0.11	0.13

X_9 is then selected and added into Π. We then calculate $r(Y; \Pi \cup \{X_i\})$ resulting in

X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}	X_{13}
X_9	X_9	X_9	X_9	X_9	X_9	X_9	X_9	X_9	X_9	X_9
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_{10}	X_{11}	X_{12}
0.10	0.11	0.11	0.09	0.12	0.12	0.10	0.11	0.10	0.10	0.12

Feature X_{12} is then selected. Similarly X_3, X_5 are subsequently selected and Π is $\{X_{13}, X_9, X_{12}, X_3, X_5\}$, where X_9 is discrete and the rest are continuous. Since $r(\Pi; Y) = 1$, we stop the selection process and Π is the selected subset of features.

4.4. Experiment and evaluation

In order to evaluate the RELFSS algorithm, we ran RELFSS on some public datasets available from the UCI repository, from which the appropriate references of origin can be obtained. Most of the datasets are frequently used in literature. Some general information about these datasets is given in Table 1.

Because most of the datasets contain missing values, a preprocessing step was necessary to apply the RELFSS algorithm to these datasets. Missing values were replaced by the mean value in case of ordinal features, and by the most frequent value (i.e., the mode) otherwise.

We ran the RELFSS algorithm for each dataset, fed the selected feature subsets to C4.5, and then cross-validated (5-fold cross-validation was used) the datasets with selected features. Table 2 shows the C4.5 results with and without RELFSS. The C4.5 module we used is the one in the Clementine package.[8] To validate our relevance measure we varied the RELFSS algorithm by replacing the relevance measure $r(Y; \Pi \cup \{x\})$ with the mutual information measure, $I(Y; \Pi \cup \{x\})$. We refer to the varied RELFSS algorithm as RELFSS'. We repeated the same experiment with RELFSS' and recorded the prediction accuracy, which is also

Table 1. General information about the datasets.

Datasets	Features	Train size	Test size	Classes
Annealing (Ann.)	38	798	CV-5	6
Australian (Aus.)	14	690	CV-5	2
Auto	25	205	CV-5	6
Bands (Ban.)	39	512	CV-5	2
Breast (Bre.)	9	286	CV-5	2
Breast-W (Bre.W)	10	683	CV-5	2
Crx	15	690	CV-5	2
Horse-Colic (Col.)	22	368	CV-5	2
Diabetes (Dia.)	8	768	CV-5	2
Heart (Hea.)	13	270	CV-5	2
German (Ger.)	20	1000	CV-5	2
Glass (Gla.)	9	214	CV-5	6
Hepatitis (Hep.)	19	155	CV-5	2
Iris	4	150	CV-5	3
Lenses (Len.)	4	24	CV-5	3
Monk-1 (M-1)	6	124	432	2
Monk-2 (M-2)	6	169	432	2
Monk-3 (M-3)	6	122	432	2
Mushroom (Mus.)	22	8124	CV-5	2
Sonar (Son.)	60	208	CV-5	2
Tic-Tac-Toe (TTT)	9	958	CV-5	2
Vehicle (Veh.)	18	846	CV-5	4
Vote	18	232	CV-5	2

shown in Table 2. Besides, we also cite the results about feature subset selection from the wrapper literature (Kohavi, 1994), which provide complementary information.

From Table 2 we see that applying the RELFSS algorithm improves prediction accuracy consistently for almost all of the datasets while applying the RELFSS' algorithm does not. Note that the average improvement is 2.93%, clearly not a dramatic improvement. A probable explanation is that most UCI datasets do not have a lot of irrelevant attributes. Another probable explanation is that RELFSS is a filter type FSS algorithm which only searches for a subspace in the original data space which can improve learning performance. A general FSS method for consistently improving learning performance dramatically has yet to be discovered, but the feature transformation type of FSS has shown some promise (Liu & Setiono, 1998).

Parity problems are well known to be difficult for many machine learning algorithms, as well as for feature subset selection algorithms. We evaluated the RELFSS algorithm using a well known parity dataset—Monks-2.[9] We added some random (irrelevant) features into

Table 2. Prediction accuracy on decision trees generated by C4.5 without and with RELFSS or RELFSS', together with the selected feature subsets. Note that RELFSS' is a variation of RELFSS in that its relevance measure is replaced by the mutual information measure. Results from the wrapper literature with respect to feature subset selection are also cited. Here, the postfix *W* means wrapper.

		C4.5-RELFSS		C4.5-RELFSS'		
Data	C4.5 Acc.	Selected features	Acc.	Selected features	Acc.	C4.5-W Acc.
Ann.	91.8	1,3,4,9,11–14,16	93.7	1,3,5,6,8,10,12,20,25,33,35	93.0	–
Aus.	85.2	2,3,8–10,13,14	85.7	2,3,8,14	84.2	–
Auto	72.2	2,6,9–13,17	76.1	11,14,26	61.5	–
Ban.	68.8	1,2,6,9,12,33	69.0	1,2	61.1	–
Bre.	74.7	2,3,5–7,10	74.7	2–10	74.7	74.7
Bre.W	93.8	3–7,9	95.3	1,3	91.8	–
Crx	85.9	2,3,9–11,14,15	86.4	2,3,9,15	85.1	–
Col.	80.9	1,10,17,22	85.9	1,4,7,8,10–12,16–18,20,22	84.5	85.3
Dia.	72.9	2,5–8	74.2	2,6,7	73.3	–
Hea.	77.1	3,5,9,12,13	82.2	5,8,10	69.6	79.2
Ger.	70.5	1,2,3,5,6,15,20	74.1	1,5	68.8	–
Gla.	63.9	2–4,6,8	72.3	1–9	63.9	62.5
Hep.	80.7	12–14	84.4	12,13,15,16,18,19	79.4	84.6
Iris	94.0	3,4	94.0	1,3,4	94.0	92.0
Len.	83.3	4,5	86.7	1	63.5	–
M-1	74.31	1,2,4,5	88.89	1,4,5	70.0	–
M-2	65.05	1–6	65.05	1–6	65.05	–
M-3	97.22	1,2,4–6	97.22	1,2,5,6	97.22	–
Mus.	100.0	6,9,13,20,21	100.0	6,10,21	99.4	–
Son.	69.4	9–13	73.1	9–13,49	66.9	–
TTT	86.2	1–9	86.2	1–7,9	85.5	–
Veh.	69.9	3,6–9,11,12	66.0	7,11,12	55.0	–
Vote	96.1	1,2,4–17	97	1,2,4–17	97	95.2
Average	80.60	N/A	82.96	N/A	77.59	N/A

the original dataset to see whether RELFSS can select the original (relevant) features. The result is reported in Table 3. From this table we can see that RELFSS can select a large proportion of relevant features from parity datasets when irrelevant features abound. But as the number of irrelevant features increases beyond a certain point (in this experiment, this point is 7 times the number of relevant features), the proportion of relevant features selected starts to decrease. The reason why RELFSS can select a large proportion of relevant features in parity problem is the fact that RELFSS is incremental: the selection of individual features is dependent on the previous selection.[10] So if one a_i is selected, then it is very likely that another a_j will stand out with higher relevance value and hence be selected.

Table 3. Feature subsets selected by RELFSS for Monks-2 data when extra irrelevant features are artificially added. The original features are labeled by a_i ($i = 1, 2, \ldots, 6$) and the added features are labeled by n_j ($j = 0, 1, \ldots$). Note that NIF is short for *Number of Irrelevant Features* and NOFS is short for *Number of Original Features Selected*.

Dataset	NIF	Features selected	NOFS
Monk-2	0	a4,a5,a6,a1,a2,a3	6
Monks-2-1	1	a5, a3, a1, n0, a4, a6, a2	6
Monks-2-5	5	a5, a3, a1, n0, n3, a2, n1, a6	5
Monks-2-10	10	a5, a3, n1, n8, n0, a6, n2, n4	3
Monks-2-15	15	a5, a3, n1, n8, n0, a6, n4, n12	3
Monks-2-20	20	a5, a3, n1, n8, n0, a6, n16, a2	4
Monks-2-25	25	a5, a3, n1, n8, n0, a6, n16, a2	4
Monks-2-30	30	a5, a3, n1, n8, n0, a6, n16, a2	4
Monks-2-35	35	a5, a3, n1, n8, n0, a6, n16, a2	4
Monks-2-40	40	a5, a3, n1, n8, n0, a6, n16, a2	4
Monks-2-45	45	n42, n23, a5, n11, n24, n7, a2, n32	2
Monks-2-50	50	n42, n23, a5, n48, n29, n8, a1, a6	3
Monks-2-55	55	n42, n23, a5, n48, n29, n8, a1, a6	3
Monks-2-60	60	n42, n23, a5, n48, n22, n58, a1, n34	2

The program, datasets, and other related information are available at `http://www.infj.ulst.ac.uk/~cbcj23/relfss.html`.

5. Summary and conclusion

Relevance is a familiar notion in daily life. It has been used frequently in the study of FSS. However, a quantitative and rigorous formalism usable for FSS has not been available to date. In this paper we have proposed a quantitative formalism for relevance based on mutual information and entropy. The formalism was rigorously validated against the common understanding of the notion, and against the existing axiomatic characterizations.

The formalism was then used to establish a tight connection between relevance and FSS. For this purpose, we first proposed two requirements for any feature subset to qualify as good: preservation of learning information and minimization of joint entropy. We then showed that, when identified with the variable relevance in the relevance formalism, relevance gives a direct underpinning for FSS: maximizing relevance in both ways (i.e., $r(\mathcal{X}; Y)$ and $r(Y; \mathcal{X})$) will result in an occam-optimal feature subset.

Based on this connection, a heuristic FSS algorithm, RELFSS, was designed and presented. This algorithm selects features incrementally using relevance $r(Y; \mathcal{X})$ until the $r(\mathcal{X}; Y)$ relevance saturates. Although it is not guaranteed to find the globally occam-optimal feature subset, it can find a locally occam-optimal feature subset. It is shown to have a polynomial complexity of $O(mN^2)$, where m is the number of instances in the dataset and N is the number of features.

The RELFSS algorithm is evaluated using 23 public datasets, resulting in an improvement in accuracy on 16 datasets, and a loss in accuracy on only 1 dataset. The experimental results provide evidence that the relevance formalism and the proposed relationship between relevance and FSS are sound.

Appendix

A. *Proof of Theorem 3.2*

By definition we have

$$\begin{aligned} & r(\mathcal{X};\mathcal{Y}) = 1 \\ \iff \quad & \frac{H(\mathcal{Y}) - H(\mathcal{Y}\mid\mathcal{X})}{H(\mathcal{Y})} = 1 \\ \iff \quad & H(\mathcal{Y}\mid\mathcal{X}) = 0 \\ \iff \quad & -E_{p(x,y)} \log p(\mathcal{Y}\mid\mathcal{X}) = 0, \text{ by the definition of conditional entropy} \end{aligned}$$

Consider now the relative entropy between $p(x, y)$ and $p(x)$,[11]

$$D(p(x,y)||p(x)) = E_{p(x,y)} \log \frac{p(\mathcal{X},\mathcal{Y})}{p(\mathcal{X})} = |E_{p(x,y)} \log p(\mathcal{Y}\mid\mathcal{X})|$$

Therefore we have $D(p(x, y) \mid p(x)) = 0 \iff p(x, y) = p(x)$. That is, $\mathcal{Y}$ is fully dependent on $\mathcal{X}$.

Similarly $r(\mathcal{X};\mathcal{Y}) = 0 \iff H(\mathcal{Y}) = H(\mathcal{Y}\mid\mathcal{X}) \iff I(\mathcal{X};\mathcal{Y}) = 0$. Moreover by the definition of mutual information, we have $I(\mathcal{X};\mathcal{Y}) = E_{p(x,y)} \frac{p(x,y)}{p(x)p(y)} \iff p(x, y) = p(x)p(y)$. That is, $\mathcal{Y}$ is independent of $\mathcal{X}$.

For the conditional case, we have

$$\begin{aligned} & r(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 1 \iff H(\mathcal{Y}\mid\mathcal{X},\mathcal{Z}) = 0 \iff -E_{p(x,y\mid z)} \log p(\mathcal{Y}\mid\mathcal{X},\mathcal{Z}) = 0 \\ \iff \quad & D(p(x,y\mid z)\mid p(x\mid z)) = E_{p(x,y\mid z)} \log \frac{p(\mathcal{X},\mathcal{Y}\mid\mathcal{Z})}{p(\mathcal{X}\mid\mathcal{Z})} \\ & \qquad = -E_{p(x,y\mid z)} p(y\mid x,z) = 0 \\ \iff \quad & p(x,y\mid z) = p(x\mid z) \iff \mathcal{Y} \text{ is fully dependent on } \mathcal{X} \text{ given } \mathcal{Z}. \end{aligned}$$

Similarly $r(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0 \iff \mathcal{Y}$ is independent of $\mathcal{X}$ given $\mathcal{Z}$. □

B. *Proof of Theorem 3.3*

$$\begin{aligned} r(\mathcal{X},\mathcal{Y};\mathcal{Z}) &= \frac{I(\mathcal{X},\mathcal{Y};\mathcal{Z})}{H(\mathcal{Z})} = \frac{I(\mathcal{X};\mathcal{Z}) + I(\mathcal{Y};\mathcal{Z}\mid\mathcal{X})}{H(\mathcal{Z})} \\ &= r(\mathcal{X};\mathcal{Z}) + \frac{I(\mathcal{Y};\mathcal{Z}\mid\mathcal{X})}{H(\mathcal{Z})} \end{aligned}$$

$$
\begin{aligned}
&= r(\mathcal{X};\mathcal{Z}) + \frac{I(\mathcal{Y};\mathcal{Z}\mid\mathcal{X})}{H(\mathcal{Z}\mid\mathcal{X})} \times \frac{H(\mathcal{Z}\mid\mathcal{X})}{H(\mathcal{Z})} \\
&= r(\mathcal{X};\mathcal{Z}) + r(\mathcal{Y};\mathcal{Z}\mid\mathcal{X}) \times \frac{H(\mathcal{Z}\mid\mathcal{X})}{H(\mathcal{Z})} \\
&= r(\mathcal{X};\mathcal{Z}) + r(\mathcal{Y};\mathcal{Z}\mid\mathcal{X}) \times \frac{-H(\mathcal{Z}) + H(\mathcal{Z}\mid\mathcal{X}) + H(\mathcal{Z})}{H(\mathcal{Z})} \\
&= r(\mathcal{X};\mathcal{Z}) + r(\mathcal{Y};\mathcal{Z}\mid\mathcal{X}) \times [1 - r(\mathcal{X};\mathcal{Z})] \\
&= r(\mathcal{X};\mathcal{Z}) + r(\mathcal{Y};\mathcal{Z}\mid\mathcal{X}) - r(\mathcal{X};\mathcal{Z}) \times r(\mathcal{Y};\mathcal{Z}\mid\mathcal{X})
\end{aligned}
$$

Similarly $r(\mathcal{X},\mathcal{Y};\mathcal{Z}) = r(\mathcal{Y};\mathcal{Z}) + r(\mathcal{X};\mathcal{Z}\mid\mathcal{Y}) - r(\mathcal{Y};\mathcal{Z}) \times r(\mathcal{X};\mathcal{Z}\mid\mathcal{Y})$. □

C. *Proof of Theorem 3.4*

- Symmetry: We only prove the "sufficient" part. The "necessary" part can be proved similarly. $r(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0$ means either $H(\mathcal{Y}\mid\mathcal{Z}) = 0$ or $I(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0$. For the first case, i.e., $H(\mathcal{Y}\mid\mathcal{Z}) = 0$, by the "conditioning reduces entropy" property we have $H(\mathcal{Y}\mid\mathcal{X},\mathcal{Z}) = H(\mathcal{Y}\mid\mathcal{Z}) = 0$, which leads to $I(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0$. By the symmetry property of mutual information, we have $I(\mathcal{Y};\mathcal{X}\mid\mathcal{Z}) = 0$, which renders $r(\mathcal{Y};\mathcal{X}\mid\mathcal{Z}) = 0$. For the second case, $I(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0 \iff I(\mathcal{Y};\mathcal{X}\mid\mathcal{Z}) = 0$, which also renders $r(\mathcal{Y};\mathcal{X}\mid\mathcal{Z}) = 0$. This concludes our proof.
- Decomposition: If $r(\mathcal{X};\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) = 0$ then

  $$
  \begin{aligned}
  H(\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) &= 0 \text{ or} \\
  I(\mathcal{X};\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) &= 0.
  \end{aligned}
  $$

  For the first case, we have

  $$
  \begin{aligned}
  & H(\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) = 0 \\
  \iff\quad & H(\mathcal{Y}\mid\mathcal{Z}) + H(\mathcal{W}\mid\mathcal{Y},\mathcal{Z}) = 0 \\
  \iff\quad & H(\mathcal{Y}\mid\mathcal{Z}) = 0 \text{ and } H(\mathcal{W}\mid\mathcal{Y},\mathcal{Z}) = 0
  \end{aligned}
  $$

  Therefore $r(\mathcal{X};\mathcal{Y}\mid\mathcal{Z}) = 0$ and $r(\mathcal{X};\mathcal{W}\mid\mathcal{Y},\mathcal{Z}) = 0$.
  Similarly we have

  $$
  \begin{aligned}
  & H(\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) = 0 \\
  \iff\quad & H(\mathcal{W}\mid\mathcal{Z}) + H(\mathcal{Y}\mid\mathcal{W},\mathcal{Z}) = 0 \\
  \iff\quad & H(\mathcal{W}\mid\mathcal{Z}) = 0 \text{ and } H(\mathcal{Y}\mid\mathcal{W},\mathcal{Z}) = 0
  \end{aligned}
  $$

  Therefore $r(\mathcal{X};\mathcal{W}\mid\mathcal{Z}) = 0$ and $r(\mathcal{X};\mathcal{Y}\mid\mathcal{W},\mathcal{Z}) = 0$.
  For the second case, we have

  $$
  \begin{aligned}
  & I(\mathcal{X};\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) = 0 \\
  \iff\quad & H(\mathcal{Y},\mathcal{W}\mid\mathcal{Z}) - H(\mathcal{Y},\mathcal{W}\mid\mathcal{X},\mathcal{Z}) = 0
  \end{aligned}
  $$

$$\begin{aligned}
&\iff H(\mathcal{Y} \mid \mathcal{Z}) + H(\mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) - H(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}) - H(\mathcal{W} \mid \mathcal{X}, \mathcal{Y}, \mathcal{Z}) = 0 \\
&\iff r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) + r((\mathcal{X}; \mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = 0 \\
&\iff r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0 \text{ and } r(\mathcal{X}; \mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = 0.
\end{aligned}$$

Similarly we also have

$$\begin{aligned}
&I(\mathcal{X}; \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) = 0 \\
\iff &H(\mathcal{W}, \mathcal{Y} \mid \mathcal{Z}) - H(\mathcal{W}, \mathcal{Y} \mid \mathcal{X}, \mathcal{Z}) = 0 \\
\iff &H(\mathcal{W} \mid \mathcal{Z}) + H(\mathcal{Y} \mid \mathcal{W}, \mathcal{Z}) - H(\mathcal{W} \mid \mathcal{X}, \mathcal{Z}) - H(\mathcal{Y} \mid \mathcal{X}, \mathcal{W}, \mathcal{Z}) = 0 \\
\iff &r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z}) + r((\mathcal{X}; \mathcal{Y} \mid \mathcal{W}, \mathcal{Z}) = 0 \\
\iff &r(\mathcal{X}; \mathcal{W} \mid \mathcal{Z}) = 0 \text{ and } r(\mathcal{X}; \mathcal{Y} \mid \mathcal{W}, \mathcal{Z}) = 0.
\end{aligned}$$

– Weak union: The proof is already included in the proof of "decomposition" as above.
– Contraction: To show $r(\mathcal{X}; \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) = 0$, we need

$$\begin{aligned}
&H(\mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) = H(\mathcal{Y}, \mathcal{W} \mid \mathcal{X}, \mathcal{Z}) \\
\iff &H(\mathcal{Y} \mid \mathcal{Z}) + H(\mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = H(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}) + H(\mathcal{W} \mid \mathcal{X}, \mathcal{Y}, \mathcal{Z}).
\end{aligned}$$

Let's see if this requirement can be met by the conditions provided. Suppose $H(\mathcal{Y} \mid \mathcal{Z}) \neq 0$ and $H(\mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) \neq 0$. Then

$$\begin{aligned}
&r(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z}) = 0 \text{ and } r(\mathcal{X}; \mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = 0 \\
\Rightarrow\ &H(\mathcal{Y} \mid \mathcal{Z}) = H(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}) \text{ and } H(\mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = H(\mathcal{W} \mid \mathcal{X}, \mathcal{Y}, \mathcal{Z}).
\end{aligned}$$

This certainly meets the needs. If instead $H(\mathcal{Y} \mid \mathcal{Z}) = 0$, we have $H(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}) = 0$ by the "conditioning reduces entropy" property. This also leads to $r(\mathcal{X}; \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) = 0$. Similar for the $H(\mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) = 0$ case. Therefore in any case, given the conditions, we always have $r(\mathcal{X}; \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) = 0$.

This concludes our proof. □

Notes

1. In RELIEFF, the *relevance level* of feature X_i is defined by $E(\delta_i)$ over all instances in the dataset, where $\delta_i = -(x_i - \text{near-hit}_i)^2 + (x_i - \text{near-miss}_i)^2$ and x_i is a member in the domain of X_i. For an instance t in the dataset, a *near-hit* of t is such an instance that belongs to the close neighborhood of t and also to the same class as t, and a *near-miss* is such an instance that belongs to the properly close neighborhood of t but not to the same class as t.
2. In the context of feature subset selection, a *determination* is a set of features that completely determines the decision attribute (Schlimmer, 1993). In other words the decision attribute is fully dependent on a determination.
3. An excellent approach is presented in Lakemeyer (1995).
4. If X is discrete with a probability distribution $p(x)$, the entropy $H(X)$ is defined by $H(X) = -\sum_x p(x) \log p(x)$. If X is continuous with a density function $f(x)$, then the differential entropy $h(X)$ of X is defined

by $h(X) = -\int_S f(x) \log f(x)\, dx$. If X is n-bit quantized (i.e., with 2^n distinct values) as $X^\triangle$, then $h(X)$ is approximated by $H(X^\triangle) - n$. If X and Y are discrete, the mutual information $I(X;Y)$ between X and Y is $I(X;Y) = H(Y) - H(Y \mid X)$; if X and Y are continuous, $I(X;Y) = h(Y) - h(Y \mid X) \approx I(X^\triangle; Y^\triangle)$. $I(X;Y)$ is a measure of the amount of information one variable contains about another. For details, readers are invited to consult Thomas (1991).

5. For a detailed discussion on Gärdenfors' axioms as well as instance relevance and event relevance, readers are invited to consult Wang (1996).
6. A brief proof of $I(X_i; Y \mid \mathcal{S}_i) > 0 \Rightarrow I(X_i; Y) > 0$: Assume $I(X_i; Y \mid \mathcal{S}_i) > 0$ but $I(X_i; Y) = 0$. The latter means X_i and Y are independent. Then $H(X_i \mid \mathcal{S}_i, Y) = H(X_i \mid \mathcal{S}_i)$. By definition we have $I(X_i; Y \mid \mathcal{S}_i) = H(X_i \mid \mathcal{S}_i) - H(X_i \mid \mathcal{S}_i, Y) = 0$, contradicting the assumption.
7. X_1: age, X_2: sex, X_3: chest pain type, X_4: resting blood pressure, X_5: serum cholestoral, X_6: (fasting blood sugar $>$ 120 mg/dl, 1 = true, 0 = false), X_7: resting electrocardiographic results, X_8: maximum heart rate achieved, X_9: exercise induced angina (1 = yes; 0 = no), X_{10}: ST depression induced by exercise relative to rest, X_{11}: the slope of the peak exercise ST segment, X_{12}: number of major vessels (0–3) colored by flourosopy, X_{13}: thal (3 = normal; 6 = fixed defect; 7 = reversable defect), and X_{14}: diagnosis of heart disease.
8. Integral Solutions Limited (ISL): `http://www.isl.co.uk/`
9. The target concept associated to MONK-2: EXACTLY TWO of $a_1 = 1$, $a_2 = 1$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$, $a_6 = 1$.
10. Note that $r(Y; \Pi \cup \{X_i\})$ is used as the individual selection criterion, where Π is the set of previously selected features.
11. All logs in this paper are to base 2.

References

Aha, D. W. & Bankert, R. L. (1994). Feature selection for case-based classification of cloud types. In *Working notes of the AAAI94 Workshop on Case-based Reasoning* (pp. 106–112). AAAI Press.

Almuallim, H. & Dietterich, T. G. (1991). Learning with many irrelevant features. In *Proc. Ninth National Conference on Artificial Intelligence* (pp. 547–552). MIT Press.

Almuallim, H. & Dietterich, T. G. (1994). Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence, 69*, 279–305.

Amirikian, B. & Nishimura, H. (1994). What size network is good for generalization of a specific task of interest? *Neural Networks, 7*(2), 321–329.

Blum, A. (1994). Relevant examples & relevant features: thoughts from computational learning theory. In *Relevance: Proc. 1994 AAAI Fall Symposium* (pp. 14–18). AAAI Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's Razor. *Information Processing Letters, 24*, 377–380.

Carnap, R. (1962). *Logical foundations of probability*. The University of Chicago Press.

Caruana, R. A. & Freitag, D. (1994). Greedy attribute selection. In *Proceedings of the 11th international conference on machine learning* (pp. 28–36). New Brunswick, NJ: Morgan Kaufmann.

Cover, T. M. & Thomas, J. A. 1991. *Elements of information theory*. John Wiley & Sons, Inc.

Davies, S. & Russell, S. 1994. NP-completeness of searches for smallest possible feature sets. In *Proceedings of the 1994 AAAI Fall Symposium on Relevance* (pp. 37–39). AAAI Press.

Fayyad, U. & Irani, K. (1990). What should be minimized in a decision tree? In *AAAI-90: Proceedings of 8th National Conference on Artificial Intelligence* (pp. 749–754).

Fayyad, U. & Irani, K. (1992). The attribute selection problem in decision tree generation. In *AAAI-92: Proceedings of 10th National Conference on Artificial Intelligence* (pp. 104–110).

Gärdenfors, P. (1978). On the logic of relevance. *Synthese, 37*, 351–367.

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–61.

Greiner, R. & Subramanian, D. (Eds.). 1994. In *Relevance: Proc. 1994 AAAI Fall Symposium*. The AAAI Press. AAAI Technical Report FS-94-02.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the 11th international conference on machine learning* (pp. 121–129). New Brunswick, NJ: Morgan Kaufmann.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.

Kira, K. & Rendell, L. A. (1992). The feature selection problem: traditional methods and a new algorithm. In *AAAI-92* (pp. 129–134).

Kohavi, R. & Sommerfield, D. (1995). Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings of KDD'95* (pp. 192–197).

Kohavi, R. (1994). Feature Subset Selection as Search with Probabilistic Estimates. In R. Greiner, & D. Subramanian (Eds.). *Relevance: Proc 1994 AAAI Fall Symposium* (pp. 122–126). The AAAI Press.

Kononenko, I., Simec, E., & Robnik-Sikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, *7*, 39–55.

Kononenko, I. (1994) Estimating attributes: Analysis and extensions of RELIEF. In *Proceedings of the 1994 European Conference on Machine Learning* (pp. 171–182).

Lakemeyer, G. (1995). A Logical account of relevance. In *Proc. of IJCAI-95* (pp. 853–859).

Littlestone, N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. (1988). *Machine learning*, *2*, 285–318.

Liu, H. & Setiono, R. (1998). Feature transformation and multivariate decision tree induction. In *Proceedings of The First International Conference on Discovery Science (DS'98)* (pp. 279–290). Fukuoka, Japan. Springer-Verlag.

Liu, H. & Setiono, R. (1997). Feature selection via discretization of numeric attributes. *IEEE Trans on Knowledge and Data Engineering*, *9*(4), 642–645.

Muggleton, S. (ed.). 1992. *Inductive Logic Programming*. London: Academic Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, California: Morgan Kaufmann Publishers, Inc.

Quinlan, J. & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, *80*, 227–248.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.*, *14*, 1080–1100.

Schlimmer, J. C. (1993). Efficiently inducing determinations: A complete and systematic search algorithm that uses optimal pruning. In *ML93*, pp. 284–290.

Schweitzer, H. (1995). Occam algorithms for computing visual motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(11), 1033–1042.

Shore, J. E. & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Information Theory*, *26*, 26–37.

Skalak, D. B. (1994). Prototype and feature selection by sampling and random mutation hill-climbing algorithms. In *Proceedings of the 11th International Conference on Machine Learning* (pp. 293–301). New Brunswick, N.J.: Morgan Kaufmann.

Subramanian, D. & Genesereth, M. R. (1987). The relevance of irrelevance. In *Proc. of IJCAI-87* (pp. 416–422).

Ullman, J. D. (1989). *Principles of database and knowledgebase systems*. Computer Science Press.

Wallace, C. & Freeman, P. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society (B)*, *49*, 240–265.

Wang, H. (1996). Towards a unified framework of relevance. Ph.D. Thesis, Faculty of Informatics, University of Ulster, N. Ireland, UK. `http://www.infj.ulst.ac.uk/~cbcj23/thesis.ps`.

Wolpert, D. H. (1990). The relationship between Occam's Razor and convergent guessing. *Complex Systems*, *4*, 319–368.

Received November 19, 1997
Revised December 14, 1999
Final manuscript December 14, 1999