Journal of Machine Learning Research 8 (2007) 589-612 Submitted 5/06; Revised 10/06; Published 3/07

Consistent Feature Selection for Pattern Recognition in Polynomial Time

Roland Nilsson ROLLE@IFM.LIU.SE
José M. Peña JMP@IFM.LIU.SE
Division of Computational Biology, Department of Physics, Chemistry and Biology
The Institute of Technology
Linköping University
SE-581 83 Linköping, Sweden

Johan Björkegren JOHAN.BJORKEGREN@KI.SE
The Computational Medicine Group, Center for Molecular Medicine
Department of Medicine
Karolinska Institutet, Karolinska University Hospital Solna
SE-171 76 Stockholm, Sweden

Jesper Tegnér JESPERT@IFM.LIU.SE
Division of Computational Biology, Department of Physics, Chemistry and Biology
The Institute of Technology
Linköping University
SE-581 83 Linköping, Sweden

Editor: Leon Bottou

Abstract

We analyze two different feature selection problems: finding a minimal feature set optimal for classification (MINIMAL-OPTIMAL) *vs.* finding all features relevant to the target variable (ALL-RELEVANT). The latter problem is motivated by recent applications within bioinformatics, particularly gene expression analysis. For both problems, we identify classes of data distributions for which there exist consistent, polynomial-time algorithms. We also prove that ALL-RELEVANT is much harder than MINIMAL-OPTIMAL and propose two consistent, polynomial-time algorithms. We argue that the distribution classes considered are reasonable in many practical cases, so that our results simplify feature selection in a wide range of machine learning tasks.

Keywords: learning theory, relevance, classification, Markov blanket, bioinformatics

1. Introduction

Feature selection (FS) is the process of reducing input data dimension. By reducing dimensionality, FS attempts to solve two important problems: facilitate learning (inducing) accurate classifiers, and discover the most ”interesting” features, which may provide for better understanding of the problem itself (Guyon and Elisseeff, 2003).

The first problem has been extensively studied in pattern recognition research. More specifically, the objective here is to learn an optimal classifier using a minimal number of features; we refer to this as the MINIMAL-OPTIMAL problem. Unfortunately, MINIMAL-OPTIMAL is in general intractable even asymptotically (in the large-sample limit), since there exist data distributions for

which every feature subset must be tested to guarantee optimality (Cover and van Campenhout, 1977). Therefore it is common to resort to suboptimal methods. In this paper, we take a different approach to solving MINIMAL-OPTIMAL: we restrict the problem to the class of *strictly positive* data distributions, and prove that within this class, the problem is in fact polynomial in the number of features. In particular, we prove that a simple backward-elimination algorithm is asymptotically optimal. We then demonstrate that due to measurement noise, most data distributions encountered in practical applications are strictly positive, so that our result is widely applicable.

The second problem is less well known, but has recently received much interest in the bioinformatics field, for example in gene expression analysis (Golub et al., 1999). As we will explain in Section 4, researchers in this field are primarily interested in identifying *all* features (genes) that are somehow related to the target variable, which may be a biological state such as "healthy" *vs.* "diseased" (Slonim, 2002; Golub et al., 1999). This defines the ALL-RELEVANT problem. We prove that this problem is much harder than MINIMAL-OPTIMAL; it is asymptotically intractable even for strictly positive distributions. We therefore consider a more restricted but still reasonable data distribution class, and propose two polynomial algorithms for ALL-RELEVANT which we prove to be asymptotically optimal within that class.

2. Preliminaries

In this section we review some concepts needed for our later developments. Throughout, we will assume a binary classification model where training examples $(x^{(i)},y^{(i)}), i = 1,\ldots,l$ are independent samples from the random variables (X,Y) with density $f(x,y)$, where $X = (X_1,\ldots,X_n) \in \mathbb{R}^n$ is a sample vector and $Y \in \{-1,+1\}$ is a sample label. Capital X_i denote random variables, while lowercase symbols x_i denote observations. We will often treat the data vector X as a set of variables, and use the notation $R_i = X \setminus \{X_i\}$ for the set of all features except X_i. Domains of variables are denoted by calligraphic symbols $\mathcal{X}$. We will present the theory for continuous X, but all results are straightforward to adapt to the discrete case. Probability density functions (continuous variables) or probability mass functions (discrete variables) are denoted $f(x)$ and $p(x)$, respectively. Probability of events are denotes by capital P, for example, $P(Y = 1/2)$.

2.1 Distribution Classes

In the typical approach to FS, one attempts to find heuristic, suboptimal solutions while considering all possible data distributions $f(x,y)$. In contrast, we will restrict the feature selection problem to certain classes of data distributions in order to obtain optimal solutions. Throughout, we will limit ourselves to the following class.

Definition 1 *The class of strictly positive data distributions consists of the $f(x,y)$ that satisfies (i) $f(x) > 0$ almost everywhere (in the Lebesgue measure) and (ii) $P(p(y|X) = 1/2) = 0$.*

Note we do not require $f(x,y) > 0$, which would be more restrictive. The criterion (i) states that a set in $\mathcal{X}$ has nonzero probability iff it has nonzero Lebesgue measure, while (ii) states that the optimal decision boundary has zero measure. These conditions are mainly needed to ensure uniqueness of the Bayes classifier (see below). It is reasonable to assume that this class covers the great majority of practical pattern recognition problems, since most data originates in physical measurements of some kind and is inevitably corrupted by noise. For example, consider the additive Gaussian noise

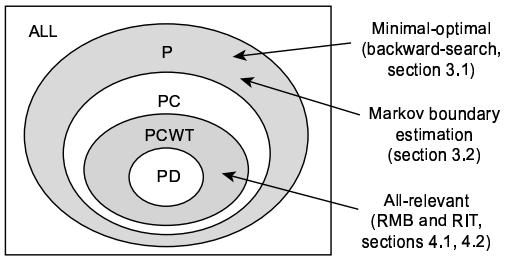

Figure 1: The distribution classes used. P, strictly positive; PC, strictly positive satisfying the composition property; PCWT, strictly positive satisfying composition and weak transitivity; PD, strictly positive and DAG-faithful. Arrows show classes where the various algorithms (right) are proved to be consistent.

model

$$X = x_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma).$$

Since the noise component ε is strictly positive over the domain of X, we immediately obtain $f(x) > 0$. A similar argument holds for binary data with Bernoulli noise, and indeed for any additive noise model with $f(\varepsilon) > 0$. In general, the strictly positive restriction is considered reasonable whenever there is uncertainty about the data (Pearl, 1988). Note that the $f(x) > 0$ criterion by definition only concerns the actual domain $\mathcal{X}$. If the data distribution is known to be constrained for physical reasons to some compact set such as $0 < X < 1$, then naturally $f(x) > 0$ need not hold outside that set. Typical problems violating $f(x) > 0$ involve noise-free data, such as inference of logic propositions (Valiant, 1984).

In Section 4, we will consider the more narrow classes of strictly positive distributions (P) that also satisfy the *composition* property (PC), and those in PC that additionally satisfies the *weak transitivity* property (PCWT). See Appendix B for details on these properties. Although these restrictions are more severe than $f(x) > 0$, these classes still allow for many realistic models. For example, the jointly Gaussian distributions are known to be PCWT (Studený, 2004). Also, the strictly positive and *DAG-faithful* distributions (PD) are contained in PCWT (Theorem 24, Appendix B). However, we hold that the PCWT class is more realistic than PD, since PCWT distributions will remain PCWT when a subset of the features are marginalized out (that is, are not observed), while PD distributions may not (Chickering and Meek, 2002).[1] This is an important argument in favor of PCWT, as in many practical cases we cannot possibly measure all variables. An important example of this is gene expression data, which is commonly modelled by PD distributions (Friedman, 2004). However, it is frequently the case that not all genes can be measured, so that PCWT is a more realistic model class. Of course, these arguments also apply to the larger PC class. Figure 1 summarizes the relations between the distribution classes discussed in this section.

1. The paper by Chickering and Meek (2002) proves that the composition property is preserved under marginalization. A similar proof for the weak transitivity property can be found in Peña et al. (2006).

2.2 Classifiers, Risk and Optimality

A *classifier* is defined as a function $g(x) : \mathcal{X} \mapsto \mathcal{Y}$, predicting a category y for each observed example x. The "goodness" of a classifier is measured as follows.

Definition 2 (risk) *The risk $R(g)$ of a classifier g is*

$$R(g) = P(g(X) \neq Y) = \sum_{y \in \mathcal{Y}} p(y) \int_{\mathcal{X}} 1_{\{g(x) \neq y\}} f(x|y) dx, \tag{1}$$

where $1_{\{\cdot\}}$ is the set indicator function.

For a given distribution $f(x,y)$, an (optimal) *Bayes classifier* is one that minimizes $R(g)$. It is easy to show that the classifier g^* that maximizes the posterior probability,

$$g^*(x) = \begin{cases} +1, & P(Y=1|x) \geq 1/2 \\ -1, & \text{otherwise} \end{cases}, \tag{2}$$

is optimal (Devroye et al., 1996). For strictly positive distributions, $g^*(x)$ is also unique, except on zero-measure subsets of $\mathcal{X}$ (above, the arbitrary choice of $g^*(x) = +1$ at the decision boundary $p(y|x) = 1/2$ is such a zero-measure set). This uniqueness is important for our results, so for completeness we provide a proof in Appendix A (Lemma 19). From now on, we speak of *the* optimal classifier g^* for a given f.

2.3 Feature Relevance Measures

Much of the theory of feature selection is centered around various definitions of feature *relevance*. Unfortunately, many authors use the term "relevant" casually and without a clear definition, which has caused much confusion on this topic. Defining relevance is not trivial, and there are many proposed definitions capturing different aspects of the concept; see for example Bell and Wang (2000) for a recent survey. The definitions considered in this paper are rooted in the well-known concept of (probabilistic) conditional independence (Pearl, 1988, sec. 2.1).

Definition 3 (conditional independence) *A variable X_i is conditionally independent of a variable Y given (conditioned on) the set of variables $S \subset X$ iff it holds that*

$$P\big(p(Y|X_i,S) = p(Y|S)\big) = 1.$$

This is denoted $Y \perp X_i|S$.

In the above, the $P(\dots) = 1$ is a technical requirement allowing us to ignore pathological cases where the posterior differs on zero-measure sets. Conditional independence is a measure of *irrelevance*, but it is difficult to use as an operational definition since this measure depends on the conditioning set S. For example, a given feature X_i can be conditionally independent of Y given $S = \emptyset$ (referred to as *marginal* independence), but still be dependent for some $S \neq \emptyset$. The well-known XOR problem is an example of this. Two well-known relevance definitions coping with this problem were proposed by John et al. (1994).

Definition 4 (strong and weak relevance) *A feature X_i is strongly relevant to Y iff $Y \not\perp X_i|R_i$. A feature X_i is weakly relevant to Y iff it is not strongly relevant, but satisfies $Y \not\perp X_i|S$ for some set $S \subset R_i$.*

Informally, a strongly relevant feature carries information about Y that cannot be obtained from any other feature. A weakly relevant feature also carries information about Y, but this information is "redundant"—it can also be obtained from other features. Using these definitions, relevance and irrelevance of a feature to a target variable Y are defined as

Definition 5 (relevance) *A feature X_i is relevant to Y iff it is strongly relevant or weakly relevant to Y. A feature X_i is irrelevant to Y iff it is not relevant to Y.*

Finally, we will need the following definition of relevance with respect to a classifier.

Definition 6 *A feature X_i is relevant to a classifier g iff*

$$P(g(X_i, R_i) \neq g(X_i', R_i)) > 0,$$

where X_i, X_i' are independent and identically distributed.

This definition states that in order to be considered "relevant" to g, a feature X must influence the value of $g(x)$ with non-zero probability. It is a probabilistic version of that given by Blum and Langley (1997). Note that here, X_i and X_i' are independent samplings of the feature X_i; hence their distributions are identical and determined by the data distribution $f(x,y)$. In the next section, we examine the relation between this concept and the relevance measures in Definition 4.

3. The Minimal-Optimal Problem

In practise, the data distribution is of course unknown, and a classifier must be induced from the training data $D^l = \{(x^{(i)}, y^{(i)})\}_{i=1}^l$ by an *inducer*, defined as a function $I : (\mathcal{X} \times \mathcal{Y})^l \mapsto \mathcal{G}$, where $\mathcal{G}$ is some space of functions. We say that an inducer is *consistent* if the induced classifier $I(D^l)$ converges in probability to g^* as the sample size l tends to infinity,

$$I(D^l) \xrightarrow{P} g^*.$$

Consistency is a reasonable necessary criterion for a sound inducer, and has been verified for a wide variety of algorithms (Devroye et al., 1996). Provided that the inducer used is consistent, we can address the feature selection problem asymptotically by studying the Bayes classifier. We therefore define the optimal feature set as follows.

Definition 7 *The* MINIMAL-OPTIMAL *feature set S^* is defined as the set of features relevant to the Bayes classifier g^* (in the sense of Definition 6).*

Clearly, S^* depends only on the data distribution, and is the minimal feature set that allows for optimal classification; hence its name. Since g^* is unique for strictly positive distributions (Lemma 19), it follows directly from Definition 6 that S^* is then also unique. Our first theorem provides an important link between the MINIMAL-OPTIMAL set and the concept of strong relevance.

Theorem 8 *For any strictly positive distribution $f(x,y)$, the* MINIMAL-OPTIMAL *set S^* contains only strongly relevant features.*

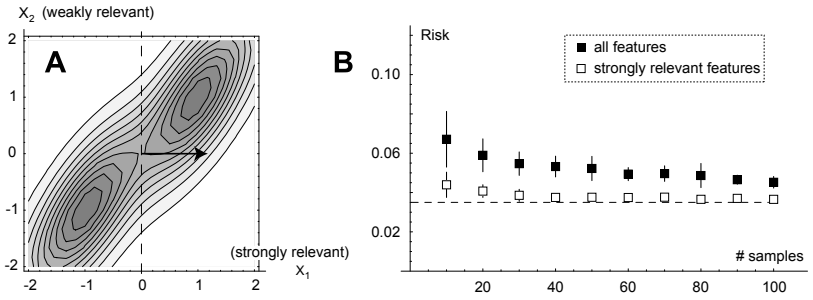

Figure 2: **A:** The example density $f(x_{2i-1}, x_{2i})$ given by (4). Here, X_1 is strongly relevant and X_2 weakly relevant. Arrow and dashed line indicates the optimal separating hyperplane. **B:** The risk functional $R(g)$ for a linear SVM trained on all relevant features (filled boxes) *vs.* on strongly relevant features only (open boxes), for the 10-dimensional distribution (4). Average and standard deviation over 20 runs are plotted against increasing sample size. The Bayes risk (dashed line) is $R(g^*) = 0.035$.

Proof Since f is strictly positive, g^* is unique by Lemma 19 (Appendix A). Consider any feature X_i relevant to g^*, so that $P(g^*(X_i, R_i) \neq g^*(X_i', R_i)) > 0$ by Definition 6. From the form (2) of the Bayes classifier, we find that

$$g^*(x_i, r_i) \neq g^*(x_i', r_i) \Rightarrow p(y|x_i, r_i) \neq p(y|x_i', r_i) \tag{3}$$

everywhere except possibly on the decision surface $\{x : p(y|x) = 1/2\}$. But this set has zero probability due to assumption (ii) of Definition 1. Therefore,

$$P(p(Y|X_i, R_i) \neq p(Y|X_i', R_i)) \geq P(g^*(X_i, R_i) \neq g^*(X_i', R_i)) > 0.$$

By Lemma 20 this is equivalent to $P(p(Y|X_i, R_i) = p(Y|R_i)) < 1$, which is the same as $Y \not\perp X_i | R_i$. Hence, X_i is strongly relevant. ■

Note that uniqueness of g^* is required here: if there would exist a different Bayes classifier g', the implication (3) would not hold. Theorem 8 is important because it asserts that we may safely ignore weakly relevant features, conditioned on the assumption $f(x) > 0$. This leads to more efficient (polynomial-time) algorithms for finding S^* for such problems. We will explore this consequence in Section 3.3.

An example illustrating Theorem 8 is given in Figure 2. Here, f is a 10-dimensional Gaussian mixture

$$f(x_1, \ldots, x_{10}, y) \propto \prod_{i=1}^{5} e^{-\frac{9}{8}((x_{2i-1}-y)^2 + (x_{2i-1}-x_{2i})^2)}. \tag{4}$$

Figure 2A shows the joint distribution of (X_{2i-1}, X_{2i}) (all such pairs are identically distributed). Note that, although the shape of the distribution in Figure 2 suggests that both features are relevant to Y,

it is easy to verify directly from (4) that $X_2, X_4, \ldots, X_{10}$ are weakly relevant: considering for example the pair (X_1, X_2), we have

$$\begin{aligned} p(y|x_1,x_2) &= \frac{f(x_1,x_2,y)}{f(x_1,x_2)} \\ &= \left[1+\exp\left\{-\frac{9}{8}((x_1+y)^2-(x_1-y)^2)\right\}\right]^{-1} \\ &= \left[1+\exp\left\{-\frac{9x_1y}{2}\right\}\right]^{-1} \end{aligned}$$

which depends only on x_1, so X_2 is weakly relevant. The Bayes classifier is easy to derive from the condition $p(y|x) > 1/2$ (Equation 2) and turns out to be $g^*(x) = \mathrm{sgn}(x_1+x_3+x_5+x_7+x_9)$, so that $S^* = \{1,3,5,7,9\}$ as expected.

For any consistent inducer I, Theorem 8 can be treated as an approximation for finite (but sufficiently large) samples. If this approximation is fair, we expect that adding weakly relevant features will degrade the performance of I, since the Bayes risk must be constant while the *design cost* must increase (Jain and Waller, 1978). To illustrate this, we chose I to be a linear, soft-margin support vector machine (SVM) (Cortes and Vapnik, 1995) and induced SVM classifiers from training data sampled from the density (4), with sample sizes $l = 10, 20, \ldots, 100$. Figure 2B shows the risk of $g = I(D^l)$ and $g_{S^*} = I_{S^*}(D^l_{S^*})$ (here and in what follows we take g_{S^*}, I_{S^*}, and $D^l_{S^*}$ to mean classifiers/inducers/data using only the features in S^*). The SVM regularization parameter C was chosen by optimization over a range $10^{-2}, \ldots, 10^2$; in each case, the optimal value was found to be 10^2. We found that I_{S^*} does outperform I, as expected. The risk functional $R(g)$ was calculated by numerical integration of Equation (1) for each SVM hyperplane g and averaged over 20 training data sets. Clearly, adding the weakly relevant features increases risk in this example.

As the following example illustrates, the converse of Theorem 8 is false: there exist strictly positive distributions where even strongly relevant features are not relevant to the Bayes classifier.

Example 1 *Let $\mathcal{X} = [0,1], \mathcal{Y} = \{-1,+1\}$, $f(x) > 0$ and $p(y{=}1|x) = x/2$. Here X is clearly strongly relevant. Yet, X is not relevant to the Bayes classifier, since we have $p(y=1|x) < 1/2$ almost everywhere (except at $x=1$). We find that $g^*(x) = -1$ and $R(g^*) = P(Y=1)$.*

Clearly, this situation occurs whenever a strongly relevant feature X_i affects the value of the posterior $p(y|x)$ but not the Bayes classifier g^* (because the change in $p(y|x)$ is not large enough to alter the decision of $g^*(x)$). In this sense, relevance to the Bayes classifier is *stronger* than strong relevance.

3.1 Related Work

The relevance concepts treated above have been studied by several authors. In particular, the relation between the optimal feature set and strong *vs.* weak relevance was treated in the pioneering study by Kohavi and John (1997), who concluded from motivating examples that "(i) all strongly relevant features and (ii) some of the weakly relevant ones are needed by the Bayes classifier". As we have seen in Example 1, part (i) of this statement is not correct in general. Part (ii) is true in general, but Theorem 8 shows that this is not the case for the class of strictly positive f, and therefore it is rarely true in practise.

A recent study by Yu and Liu (2004) examines the role of weakly relevant features in more detail and subdivides these further into *redundant* and *non-redundant* weakly relevant features, of

which the latter is deemed to be important for the Bayes classifier. However, Yu & Liu consider arbitrary distributions; for strictly positive distributions however, it is easy to see that all weakly relevant features are also "redundant" in their terminology, so that their distinction is not useful in this case.

3.2 Connections with Soft Classification

A result similar to Theorem 8 have recently been obtained for the case of *soft* (probabilistic) classification (Hardin et al., 2004; Tsamardinos and Aliferis, 2003). In soft classification, the objective is to learn the posterior $p(y|x)$ instead of $g^*(x)$. By Definition 6, the features relevant to the optimal soft classifier $p(y|x)$ satisfies

$$P(p(Y|X_i,R_i) \neq p(Y|X_i',R_i)) > 0$$

which is equivalent to $P(p(Y|X_i,R_i) \neq p(Y|R_i)) > 0$ by Lemma 20. Thus, the features relevant to $p(y|x)$ are exactly the strongly relevant ones, so the situation in Example 1 does not occur here.

When learning soft classifiers from data, a feature set commonly encountered is the *Markov boundary* of the class variable, defined as the minimal feature set required to predict the posterior. Intuitively, this is the soft classification analogue of MINIMAL-OPTIMAL. The following theorem given by Pearl (1988, pp. 97) shows that this set is well-defined for strictly positive distributions.[2]

Theorem 9 (Markov boundary) *For any strictly positive distribution $f(x,y)$, there exists a unique minimal set $M \subseteq X$ satisfying $Y \perp X \setminus M|M$. This minimal set is called the Markov boundary of the variable Y (with respect to X) and denoted $M(Y|X)$.*

Tsamardinos & Aliferis recently proved that for the PD distribution class (see Figure 1), the Markov boundary coincides with the set of strongly relevant features (Tsamardinos and Aliferis, 2003). However, as explained in Section 2.1, the PD class is too narrow for many practical applications. Below, we generalize their result to any positive distribution to make it more generally applicable.

Theorem 10 *For any strictly positive distribution $f(x,y)$, a feature X_i is strongly relevant if and only if it is in the Markov boundary $M = M(Y|X)$ of Y.*

Proof First, assume that X_i is strongly relevant. Then $Y \not\perp X_i|R_i$, which implies $M \not\subseteq R_i$, so $X_i \in M$. Conversely, fix any $X_i \in M$ and let $M' = M \setminus \{X_i\}$. If X_i is not strongly relevant, then $Y \perp X_i|R_i$, and by the definition of the Markov boundary, $Y \perp X \setminus M|M$. We may rewrite this as

$$\begin{cases} Y \perp X_i|M' \cup X \setminus M \\ Y \perp X \setminus M|M' \cup \{X_i\}. \end{cases}$$

The intersection property (Theorem 25, Appendix B) now implies $Y \perp X \setminus M'|M'$. Hence, M' is a Markov blanket smaller than M, a contradiction. We conclude that X_i is strongly relevant. ■

The feature set relations established at this point for strictly positive distributions are summarized in Figure 3.

2. Pearl (1988) gives a proof assuming $f(x,y) > 0$. However, it is straightforward to relax this assumption to $f(x) > 0$ in our case, since we only consider Markov boundaries of Y. See Theorem 25 and corollary 26 in Appendix B.

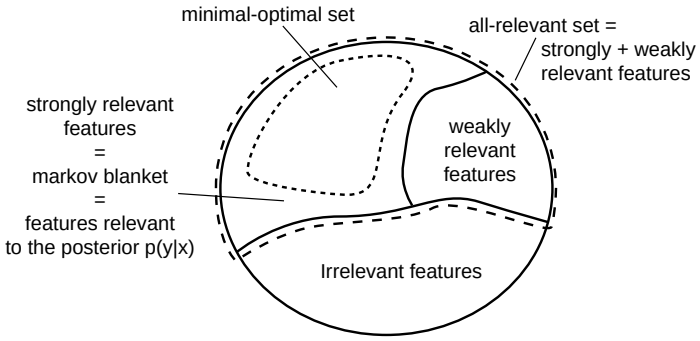

Figure 3: The identified relations between feature sets for strictly positive distributions. The circle represents all features. The dotted line (MINIMAL-OPTIMAL) denotes a subset, while the solid lines denote a partition into disjoint sets.

3.3 Consistent Polynomial Algorithms

By limiting the class of distributions, we have simplified the problem to the extent that weakly relevant features can be safely ignored. In this section, we show that this simplification leads to polynomial-time feature selection (FS) algorithms. A FS algorithm can be viewed as a function $\Phi(D^l) : (\mathcal{X} \times \mathcal{Y})^l \mapsto 2^X$, where 2^X denotes the power-set of X. For finite samples, the optimal Φ depends on the unknown data distribution f and the inducer I (Tsamardinos and Aliferis, 2003). Asymptotically however, we may use consistency as a reasonable necessary criterion for a "correct" algorithm. In analogue with consistency of inducers, we define a FS algorithm $\Phi(D^l)$ to be consistent if it converges in probability to the MINIMAL-OPTIMAL set,

$$\Phi(D^l) \xrightarrow{P} S^*.$$

Conveniently, consistency of Φ depends only on the data distribution f. Next, we propose a polynomial-time FS algorithm and show that it is consistent for any strictly positive f. As before, feature sets used as subscripts denote quantities using only those features.

Theorem 11 *Take any strictly positive distribution $f(x,y)$ and let $\hat{c}(D^l_S)$ be a real-valued criterion function such that, for every feature subset S,*

$$\hat{c}(D^l_S) \xrightarrow{P} c(S), \tag{5}$$

where $c(S)$ depends only on the distribution $f(x,y)$ and satisfies

$$c(S) < c(S') \iff R(g^*_S) < R(g^*_{S'}). \tag{6}$$

Then the feature selection method

$$\Phi(D^l) = \{i : \hat{c}(D^l_{R_i}) > \hat{c}(D^l) + \varepsilon\}$$

where $\varepsilon \in (0, \eta)$ with $\eta = \min_{i \in S^*}(c(R_i) - c(X))$, is consistent.

Proof Since f is strictly positive, S^* is unique by Lemma 19. By Definition 7 and the assumption (6) it holds that $X_i \in S^*$ iff $c(X) < c(R_i)$. First consider the case $X_i \in S^*$. Fix an $\varepsilon \in (0, \eta)$ and let $\varepsilon' = \min\{(\eta - \varepsilon)/2, \varepsilon/2\}$. Choose any $\delta > 0$. By (5) there exist an l_0 such that for all $l > l_0$,

$$P\left(\max_S |\hat{c}(D_S^l) - c(S)| > \varepsilon'\right) \leq \delta/2n$$

Note that since the power-set 2^X is finite, taking the maxima above is always possible even though (5) requires only point-wise convergence for each S. Therefore the events (i) $\hat{c}(D_X^l) < c(X) + \varepsilon'$ and (ii) $\hat{c}(D_{R_i}^l) > c(R_i) - \varepsilon'$ both have probability at least $1 - \delta/2n$. Subtracting the inequality (i) from (ii) yields

$$\begin{aligned} \hat{c}(D_{R_i}^l) - \hat{c}(D_X^l) &> c(R_i) - c(X) - 2\varepsilon' \\ &\geq c(R_i) - c(X) - (\eta - \varepsilon) \geq \varepsilon. \end{aligned}$$

Thus, for every $l > l_0$,

$$\begin{aligned} P(X_i \in \Phi(D^l)) &= P(\hat{c}(D_{R_i}^l) - \hat{c}(D_X^l) > \varepsilon) \\ &\geq P\left(\hat{c}(D_X^l) < c(X) + \varepsilon' \wedge \hat{c}(D_{R_i}^l) > c(R_i) - \varepsilon'\right) \\ &\geq P\left(\hat{c}(D_X^l) < c(X) + \varepsilon'\right) + P\left(\hat{c}(D_{R_i}^l) > c(R_i) - \varepsilon'\right) - 1 \\ &\geq 1 - \delta/n. \end{aligned}$$

For the converse case $X_i \notin S^*$, note that since $c(X) = c(R_i)$,

$$\begin{aligned} P(X_i \in \Phi(D^l)) &= P\left(\hat{c}(D_{R_i}^l) - \hat{c}(D_X^l) > \varepsilon\right) \\ &\leq P(|\hat{c}(D_{R_i}^l) - c(R_i)| + |c(X) - \hat{c}(D_X^l)| > \varepsilon) \\ &\leq P\left(|\hat{c}(D_{R_i}^l) - c(R_i)| > \frac{\varepsilon}{2} \vee |c(X) - \hat{c}(D_X^l)| > \frac{\varepsilon}{2}\right) \\ &\leq P\left(|\hat{c}(D_{R_i}^l) - c(R_i)| > \varepsilon'\right) + P\left(|c(X) - \hat{c}(D_X^l)| > \varepsilon'\right) \leq \delta/n \end{aligned}$$

where in the last line we have used $\varepsilon' \leq \varepsilon/2$. Putting the pieces together, we obtain

$$\begin{aligned} P(\Phi(D^l) = S^*) &= P(\Phi(D^l) \supseteq S^* \wedge \Phi(D^l) \subseteq S^*) \\ &= P(\forall i \in S^* : X_i \in \Phi(D^l) \wedge \forall i \notin S^* : X_i \notin \Phi(D^l)) \\ &\geq |S^*|(1 - \delta/n) + (n - |S^*|)(1 - \delta/n) - (n-1) = 1 - \delta. \end{aligned}$$

Since δ was arbitrary, the required convergence follows. ■

The requirement to choose an $\varepsilon < \eta$ may seem problematic, since in practise η depends on the true distribution $f(x,y)$ and hence is unobservable. For convergence purposes, this can be remedied by choosing a sequence $\varepsilon = \varepsilon(l) \to 0$, so that $\varepsilon < \eta$ will become satisfied eventually. In practise, the parameter ε controls the trade-off between precision and recall; a small ε gives high recall but low precision, and vice versa. If it is possible to estimate the distribution of $\hat{c}$, one might attempt to choose ε so as to control precision as recall as desired. However, this is a difficult issue where further research is necessary.

The algorithm Φ evaluates the criterion $\hat{c}$ precisely n times, so it is clearly polynomial in n provided that $\hat{c}$ is. The theorem applies to both filter and wrapper methods, which differ only in the choice of $\hat{c}(D_S^l)$ (Kohavi and John, 1997). To apply the theorem in a particular case, we need only verify that the requirements (5) and (6) hold. For example, let I be the k-NN rule with training data $D^{l/2} = \{(X_1,Y_1),\ldots,(X_{l/2},Y_{l/2})\}$ and let $\hat{R}$ be the usual empirical risk estimate on the remaining samples $\{(X_{l/2+1},Y_{l/2+1}),\ldots,(X_l,Y_l)\}$. Provided k is properly chosen, this inducer is known to be universally consistent,

$$P(R(I_S(D_S^{l/2})) - R(g_S^*) > \varepsilon) \leq 2e^{-l\varepsilon^2/(144\gamma_S^2)}$$

where γ_S depends on $|S|$ but not on l (Devroye et al., 1996, pp. 170). Next, with a test set of size $l/2$, the empirical risk estimate satisfies

$$\forall g : P(|\hat{R}(g) - R(g)| > \varepsilon) \leq 2e^{-l\varepsilon^2}$$

(Devroye et al., 1996, pp. 123). We choose $\hat{c}(D_S^l) = \hat{R}(I_S(D_S^{l/2}))$ and $c(S) = R(g_S^*)$, so that (6) is immediate. Further, this $\hat{c}(D_S^l)$ satisfies

$$\begin{aligned}
&P\left(|\hat{c}(D_S^l) - c(S)| > \varepsilon\right) = P\left(|\hat{R}(I_S(D_S^{l/2})) - R(g_S^*)| > \varepsilon\right) \\
&\leq P\left(|\hat{R}(I_S(D_S^{l/2})) - R(I_S(D_S^{l/2}))| + |R(I_S(D_S^{l/2})) - R(g_S^*)| > \varepsilon\right) \\
&\leq P\left(|\hat{R}(I_S(D_S^{l/2})) - R(I_S(D_S^{l/2}))| > \frac{\varepsilon}{2}\right) + P\left(|R(I_S(D_S^{l/2})) - R(g_S^*)| > \frac{\varepsilon}{2}\right) \\
&\leq 2e^{-l\varepsilon^2/4} + 2e^{-l\varepsilon^2/(576\gamma_S^2)} \to 0
\end{aligned}$$

for every S as required by (5), and is polynomial in n. Therefore this choice defines a polynomial-time, consistent wrapper algorithm Φ. Note that we need only verify the point-wise convergence (5) for any given $\hat{c}(S)$, which makes the application of the theorem somewhat easier. Similarly, other consistent inducers and consistent risk estimators could be used, for example support vector machines (Steinwart, 2002) and the cross-validation error estimate (Devroye et al., 1996, chap. 24).

The FS method Φ described in Theorem 11 is essentially a backward-elimination algorithm. With slight modifications, the above shows that many popular FS methods that implement variants of backward-search, for example Recursive Feature Elimination (Guyon et al., 2002), are in fact consistent. This provides important evidence of the soundness of these algorithms.

In contrast, forward-search algorithms are not consistent even for strictly positive f. Starting a with feature set S, forward-search would find the feature set $S' = S \cup \{X_i\}$ (that is, add feature X_i) iff $\hat{c}(D_{S'}^l) < \hat{c}(D_S^l)$. But it may happen that $R(g_{S'}^*) \not< R(g_S^*)$ even though S' is contained in S^*. Therefore, forward-search may miss features in S^*. The "noisy XOR problem" (Guyon and Elisseeff, 2003, pp. 1116) is an example of a strictly positive distribution with this property.

A simple example illustrating Theorem 11 is shown in Figure 4. We implemented the feature selection method Φ defined in the theorem, and again used the data density f from Equation (4). Also here, we employed a linear SVM as inducer. We used the leave-one-out error estimate (Devroye et al., 1996) as $\hat{R}$. As sample size increases, we find that the fraction of strongly relevant features selected approaches 1, confirming that $\Phi(D^l) \xrightarrow{P} S^*$. Again, this emphasizes that asymptotic results can serve as good approximations for reasonably large sample sizes.

The algorithm Φ is primarily intended as a constructive proof of the fact that polynomial and consistent algorithms exist; we do not contend that it is optimal in practical situations. Nevertheless,

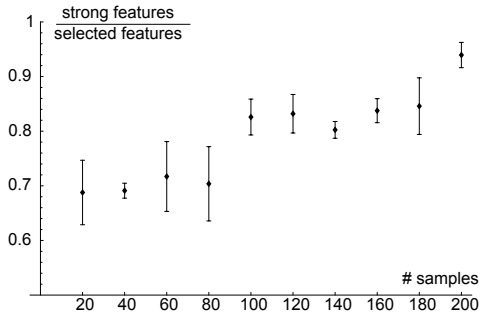

Figure 4: A feature selection example on a 10-dimensional density with 5 strongly and 5 weakly relevant features (Equation 4). Averaged results of 50 runs are plotted for samples sizes $20, \ldots, 200$. Error bars denote standard deviations.

Data set	$l \times n$	No FS	$\Phi_{\varepsilon=0}$	RELIEF	FCBF
Breast cancer	569×30	8	9(7)	69(8)	62(2)
Ionosphere	351×34	11	9(14)	16(26)	14(5)
Liver disorder	345×6	36	39(5)	$-$(0)	43(2)
E.Coli	336×7	36	20(5)	43(1)	57(1)
P.I. Diabetes	768×8	33	36(7)	35(1)	35(1)
Spambase	4601×57	12	17(39)	25(26)	40(4)

Table 1: Feature selection on UCI data sets. Test error rates are given in %, number of features selected in parentheses. Significant differences from the classifier without feature selection ("No FS") are underscored (McNemar's test, $p = 0.05$). l denotes number of samples, n number of features.

we conducted some experiments using Φ on a set of well-known data sets from the UCI machine learning repository (Newman et al., 1998) to demonstrate empirically that weakly relevant features do not contribute to classifier accuracy. We used a 5-NN classifier together with a 10-fold cross-validation error estimate for the criterion function $\hat{c}$. For each case we estimated the final accuracy by holding out a test set of 100 examples. Statistical significance was evaluated using McNemar's test (Dietterich, 1998). We set $\varepsilon = 0$ in this test, as we were not particularly concerned about false positives. For comparison we also tried the RELIEF algorithm (Kira. and Rendell, 1992) and the FBCF algorithm by Yu and Liu (2004), both of which are based on the conjecture that weakly relevant features may be needed. We found that Φ never increased test error significantly compared to the full data set, and significantly improved the accuracy in one case (Table 1). The FCBF and Relief algorithms significantly increased the test error in five cases. Overall, these methods selected very few features (in one case, RELIEF selected no features at all) using the default thresholds recommended by the original papers (for FCBF, $\gamma = n/\log n$ and for RELIEF, $\theta = 0$, in the notation of each respective paper; these correspond to the ε parameter of the Φ algorithm).

In the case of the P.I. Diabetes set, Φ seems to select redundant features, which at first might seem to contradict our theory. This may happen for two reasons. First, at $\varepsilon = 0$, the Φ algorithm is inclined to to include false positives (redundant features) rather than risk any false negatives. Second, it is possible that some of these features are truly in S^*, even though the reduction in classification error is too small to be visible on a small test set.

Theorem 10 also has implications for algorithmic complexity in the case of soft classification. To find the Markov boundary, one need now only test each X_i for strong relevance, that is, for the conditional independence $Y \perp X_i | X_{R_i}$. This procedure is clearly consistent and can be implemented in polynomial time. It is not very practical though, since these tests have very limited statistical power for large n due to the large conditioning sets R_i. However, realistic solutions have recently been devised for more narrow distribution classes such as PC, yielding polynomial and consistent algorithms (Peña et al., 2005; Tsamardinos and Aliferis, 2003).

4. The All-Relevant Problem

Recently, feature selection has received much attention in the field of bioinformatics, in particular in gene expression data analysis. Although classification accuracy is an important objective also in this field, many researchers are more interested in the "biological significance" of features (genes) that depend on the target variable Y (Slonim, 2002). As a rule, biological significance means that a gene is causally involved in the biological process of interest. It is imperative to understand that this biological significance is very different from that in Definition 5. Clearly, genes may be useful for predicting Y without being causally related, and may therefore be irrelevant to the biologist.

Typically, feature selection is performed to optimize classification performance (that is, one attempts to solve MINIMAL-OPTIMAL), and the features chosen are then examined for biological significance (Golub et al., 1999; Guyon et al., 2002). Unfortunately, this strategy ignores the distinction between biological significance and prediction. The features in S^* are typically those with good signal-to-noise ratio (that is, those very predictive of the class), but these need not be more biologically significant than other features dependent on Y. For example, a biologically very important class of genes called *transcription factors* are often present in very small amounts and are therefore difficult to detect with microarrays, leading to poor signal-to-noise ratios (Holland, 2002). Yet, these genes are often implicated in for example cancer development (Darnell, 2002).

Therefore, it is desirable to identify *all* genes relevant to the target variable, rather than the set S^*, which may be more determined by technical factors than by biological significance. Hence, we suggest that the following feature set should be found and examined.

Definition 12 *For a given data distribution, the* ALL-RELEVANT *feature set S^A is the set of features relevant to Y in the sense of Definition 5.*

To the best of our knowledge, this problem has not been studied. We will demonstrate that because S^A includes weakly relevant features (Figure 3), the ALL-RELEVANT problem is much harder than MINIMAL-OPTIMAL. In fact, the problem of determining whether a single feature X_i is weakly relevant requires exhaustive search over all 2^n subsets of X, even if we restrict ourselves to strictly positive distributions.

Theorem 13 *For a given feature X_i and for every $S \subseteq R_i$, there exists a strictly positive $f(x,y)$ satisfying*

$$Y \not\perp X_i|S \quad \wedge \quad \forall S' \neq S : Y \perp X_i|S'. \tag{7}$$

Proof Without loss of generalization we may take $i = n$ and $S = X_1, \ldots X_k$, $k = |S|$. Let $S \cup \{X_{k+1}\}$ be distributed as a $k+1$-dimensional Gaussian mixture

$$\begin{aligned} f(s, x_{k+1}|y) &= \frac{1}{|M_y|} \sum_{\mu \in M_y} N(s, x_{k+1}|\mu, \Sigma), \\ M_y &= \{\mu \in \{1,0\}^{k+1} : \mu_1 \oplus \cdots \oplus \mu_{k+1} = (y+1)/2\}, \end{aligned}$$

where $\oplus$ is the XOR operator (M_y is well-defined since $\oplus$ is associative and commutative). This distribution is a multivariate generalization of the "noisy XOR problem" (Guyon and Elisseeff, 2003). It is obtained by placing Gaussian densities centered at the corners of a $k+1$-dimensional hypercube given by the sets M_y, for $y = \pm 1$. It is easy to see that this gives $Y \not\perp X_{k+1}|S$ and $Y \perp X_{k+1}|S'$ if $S' \subset S$. Next, let $X_{i+1} = X_i + \varepsilon$ for $k < i < n$, where ε is some strictly positive noise distribution. Then it holds that $Y \not\perp X_i|S$ for $k < i < n$, and in particular $Y \not\perp X_n|S$. But it is also clear that $Y \perp X_n|S'$ for $S' \supset S$, since every such S' contain a better predictor $X_i, k < i < n$ of Y. Taken together, this is equivalent to (7), and f is strictly positive. ■

This theorem asserts that the conditioning set that satisfies the relation $Y \not\perp X_i|S$ may be completely arbitrary. Therefore, no search method other than exhaustively examining all sets S can possibly determine whether X_i is weakly relevant. Since ALL-RELEVANT requires that we determine this for every X_i, the following corollary is immediate.

Corollary 14 *The all-relevant problem requires exhaustive subset search.*

Exhaustive subset search is widely regarded as an intractable problem, and no polynomial algorithm is known to exist. This fact is illustrative in comparison with Theorem 8; MINIMAL-OPTIMAL is tractable for strictly positive distributions precisely because S^* does *not* include weakly relevant features.

Since the restriction to strictly positive distributions is not sufficient to render ALL-RELEVANT tractable, we must look for additional constraints. In the following sections we propose two different polynomial-time algorithms for finding S^A, and prove their consistency.

Algorithm 1: Recursive independence test (RIT)

Input: target node Y, features X
Let $S = \emptyset$;
foreach $X_i \in X$ **do**
  if $X_i \not\perp Y|\emptyset$ **then**
    $S = S \cup \{X_i\}$;
  end
end
foreach $X_i \in S$ **do**
  $S = S \cup \mathrm{RIT}(X_i, X \setminus S)$;
end
return S

4.1 Recursive Independence Test

A simple, intuitive method for solving ALL-RELEVANT is to test features pairwise for *marginal* dependencies: first test each feature against Y, then test each feature against every variable found to be dependent on Y, and so on, until no more dependencies are found (Algorithm 1). We refer to this algorithm as Recursive Independence Testing (RIT). We now prove that the RIT algorithm is consistent (converges to S^A) for PCWT distributions, provided that the test used is consistent.

Theorem 15 *For any PCWT distribution, let R denote the set of variables $X_k \in X$ for which there exists a sequence $Z_1^m = \{Z_1, \ldots, Z_m\}$ between $Z_1 = Y$ and $Z_m = X_k$ such that $Z_i \not\perp Z_{i+1}|\emptyset$, $i = 1, \ldots, m-1$. Then $R = S^A$.*

Proof Let $I = X \setminus R$ and fix any $X_k \in I$. Since $Y \perp X_k|\emptyset$ and $X_i \perp X_k|\emptyset$ for any $X_i \in R$, we have $\{Y\} \cup R \perp I|\emptyset$ by the composition property. Then $Y \perp X_k|S$ for any $S \subset X \setminus \{X_k, Y\}$ by the weak union and decomposition properties, so X_k is irrelevant; hence, $S_A \subseteq R$.

For the converse, fix any $X_k \in R$ and let $Z_1^m = \{Z_1, \ldots, Z_m\}$ be a shortest sequence between $Z_1 = Y$ and $Z_m = X_k$ such that $Z_i \not\perp Z_{i+1}|\emptyset$ for $i = 1, \ldots, m-1$. Then we must have $Z_i \perp Z_j|\emptyset$ for $j > i+1$, or else a shorter sequence would exist. We will prove that $Z_1 \not\perp Z_m|Z_2^{m-1}$ for *any* such shortest sequence, by induction over the sequence length. The case $m = 2$ is trivial. Consider the case $m = p$. Assume as the induction hypothesis that, for any $i, j < p$ and any chain Z_i^{i+j} of length j, it holds that $Z_i \not\perp Z_{i+j}|Z_{i+1}^{i+j-1}$. By the construction of the sequence Z_1^m it also holds that

$$Z_1 \perp Z_i|\emptyset, \quad 3 \leq i \leq m \qquad \Longrightarrow \qquad Z_1 \perp Z_3^i|\emptyset \tag{8}$$

(composition)

$$\Longrightarrow \qquad Z_1 \perp Z_i|Z_3^{i-1}. \tag{9}$$

(weak union)

Now assume to the contrary that $Z_1 \perp Z_p|Z_2^{p-1}$. Together with (9), weak transitivity implies

$$Z_1 \perp Z_2|Z_3^{p-1} \quad \vee \quad Z_2 \perp Z_p|Z_3^{p-1}.$$

The latter alternative contradicts the induction hypothesis. The former together with (8) implies $Z_1 \perp Z_2^{p-1}|\emptyset$ by contraction, which implies $Z_1 \perp Z_2|\emptyset$ by decomposition. This is also a contradiction;

Algorithm 2: Recursive Markov Boundary (RMB)

Input: target node Y, data X, visited nodes V
Let $S = M(Y|X)$, the Markov blanket of Y in X;
foreach $X_i \in S \setminus V$ **do**
    $S = S \cup \text{RMB}(Y, X \setminus \{X_i\}, V)$;
    $V = V \cup S$
end
return S

hence $Z_1 \not\perp Z_p | Z_2^{p-1}$, which completes the induction step. Thus X_k is relevant and $R \subseteq S^A$. The theorem follows. ■

Corollary 16 *For any PCWT distribution and any consistent marginal independence test, the RIT algorithm is consistent.*

Proof Since the test is consistent, the RIT algorithm will discover every sequence $Z_1^m = \{Z_1, \ldots, Z_m\}$ between $Z_1 = Y$ and $Z_m = X_k$ with probability 1 as $l \to \infty$. Consistency follows from Theorem 15. ■

Since the RIT algorithm makes up to $n = |X|$ tests for each element of R found, in total RIT will evaluate no more than $n|R|$ tests. Thus, for small R the number of tests is approximately linear in n, although the worst-case complexity is quadratic.

There are many possible alternatives as to what independence test to use. A popular choice in Bayesian networks literature is Fisher's Z-test, which tests for linear correlations and is consistent within the family of jointly Gaussian distributions (Kalisch and Bühlmann, 2005). Typically, for discrete Y a different test is needed for testing $Y \perp X_i | \emptyset$ than for testing $X_i \perp X_j$. A reasonable choice is Student's t-test, which is consistent for jointly Gaussian distributions (Casella and Berger, 2002). More general independence tests can be obtained by considering correlations in kernel Hilbert spaces, as described by Gretton et al. (2005).

4.2 Recursive Markov Boundary

In this section we propose a second algorithm for ALL-RELEVANT called Recursive Markov Boundary (RMB), based on a given consistent estimator of Markov boundaries of Y. Briefly, the RMB algorithm first estimates $M(Y|X)$, then estimates $M(Y|X \setminus \{X_i\})$ for each $X_i \in M(Y|X)$, and so on recursively until no more nodes are found (Algorithm 2). For efficiency, we also keep track of previously visited nodes V to avoid visiting the same nodes several times. We start the algorithm with $\text{RMB}(Y, X, V = \emptyset)$. A contrived example of the RMB algorithm for a PD distribution is given in Figure 5.

Next, we prove that also the RMB algorithm is consistent for any PCWT distribution, assuming that we are using a consistent estimator of Markov boundaries (see below). The proof makes use of the following concept.

Definition 17 *An independence map (I-map) over a set of features $X = \{X_1, \ldots, X_n\}$ is an undirected graph G over X satisfying*

$$R \perp_G S|T \implies R \perp S|T$$

where R, S, T are disjoint subsets of X and $\perp_G$ denotes vertex separation in the graph G, that is, $R \perp_G S|T$ holds iff every path in G between R and S contains at least one $X_i \in T$. An I-map is minimal if no subgraph G' of G (over the same nodes X) is an I-map.

Note that we only consider the minimal I-map over the features X (not over $X \cup \{Y\}$), since in this case, the minimal I-map is unique for any strictly positive distribution (Pearl, 1988).

Theorem 18 *For any PCWT distribution such that a given estimator $M(Y|S)$ of the Markov boundary of Y with respect to a feature set S is consistent for every $S \subseteq X$, the RMB algorithm is consistent.*

Proof For every $S \subseteq X$, the marginal distribution over S, Y is strictly positive, and therefore every Markov boundary $M(Y|S)$ is unique by corollary 26 (Appendix B). Let G be the minimal I-map over the features X, and let $M_1 = M(X|Y)$. Fix an X_k in S^A. If $X_k \in M_1$, we know that X_k is found by RMB. Otherwise, by Lemma 22, there exists a shortest path Z_1^m in G between some $Z_1 \in M_1$ and $Z_m = X_k$. We prove by induction over m that RMB visits every such path. The case $m = 1$ is trivial. Let the induction hypothesis be that Z_p is visited. For Z_{p+1}, Lemma 22 implies $Y \not\perp Z_{p+1}|X \setminus Z_1^{p+1}$. Since Z_p is visited, RMB will also visit all nodes in $M_{p+1} = M(Y|X \setminus Z_1^p)$. However, M_{p+1} contains Z_{p+1}, because it contains all X_i satisfying $Y \not\perp X_i|X \setminus Z_1^p \setminus \{X_i\}$ by Theorem 10. ■

It is easy to see from algorithm 2 that the RMB algorithm requires computing $|S_A|$ Markov blankets. We might attempt to speed it up by marginalizing out several nodes at once, but in that case we cannot guarantee consistency. A general algorithm for estimating Markov boundaries is given by Peña et al. (2005). This estimator is consistent in the PC class, so it follows that RMB is consistent in PCWT in this case (see Figure 1).

At first sight, RMB may seem more computationally intensive that RIT. However, since the Markov boundary is closely related to S^* (see Section 3.2), we anticipate that RMB may be implemented using existing FS methods. In particular, for distribution classes where the Markov boundary coincides with the MINIMAL-OPTIMAL set S^*, one may compute $M(Y|X)$ using any FS method that consistently estimates S^*. For example, this property holds for the well-known class of jointly Gaussian distributions $f(x|y) = N(x|y\mu, \Sigma)$ with $p(y) = 1/2$. To see this, note that the posterior and Bayes classifier are given by

$$\begin{aligned} p(y|x) &= \left[1 + \exp\{-2y\mu^T\Sigma^{-1}x\}\right]^{-1}, \\ g^*(x) &= \operatorname{sgn}(\mu^T\Sigma^{-1}x). \end{aligned}$$

Clearly, both $g^*(x)$ and $p(y|x)$ are constant with respect to an x_i iff $(\mu^T\Sigma^{-1})_i = 0$. Thus, in this case S^* equals the Markov boundary. SVM-based FS methods are one attractive option, as there exist efficient optimization methods for re-computation of the SVM solution after marginalization (Keerthi, 2002).

4.3 Related Work

We have not been able to find any previous work directly aimed at solving the ALL-RELEVANT problem. It is related to inference of graphical probability models: for the PD distributions class,

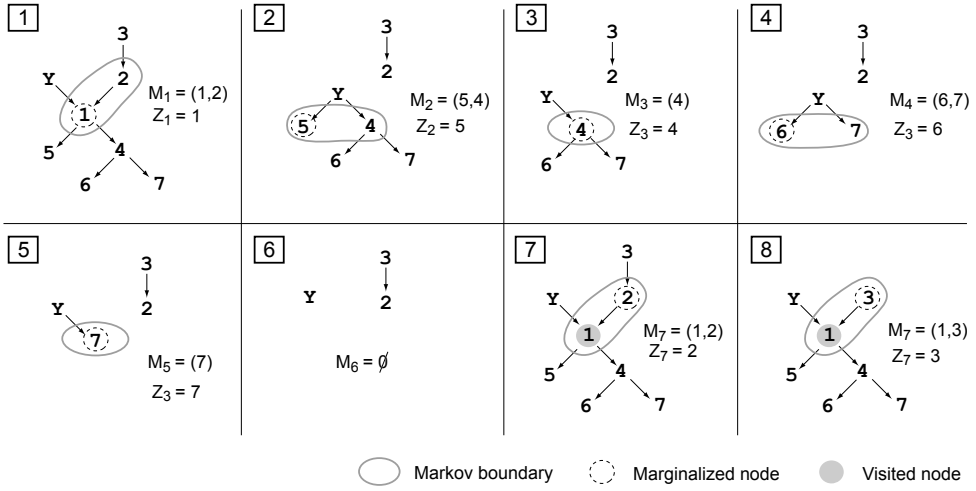

Figure 5: A contrieved example of the RMB algorithm for a PD distribution faithful to the DAG shown in black arrows. Numbers denote the relevant features, Y denotes the target variable. As in Theorem 18, M_i denotes Markov boundaries and Z_i denotes marginalized nodes. Note that the marginal distributions from step 2 onwards may not be PD, so absence of arrows should not be interpreted as independencies.

ALL-RELEVANT can be solved by inferring a Bayesian network and then taking S_A to be the connected component of Y in that network. However, this is less efficient than our approach, since Bayesian network inference is asymptotically NP-hard, even in the very restricted PD class (Chickering et al., 2004). Certainly, such a strategy seems inefficient as it attempts to "solve a harder problem as an intermediate step" (by inferring a detailed model of the data distribution merely to find the set S^A), thus violating Vapnik's famous principle (Vapnik, 2000, pp. 39).

On the other hand, several methods have been proposed for solving MINIMAL-OPTIMAL that in fact attempt to find all relevant features, since they do not assume $f > 0$ and therefore cannot rule out the weakly relevant ones. These include FOCUS (Almuallim and Dietterich, 1991), which considers the special case of binary $\mathcal{X}$ and noise-free labels; RELIEF (Kira. and Rendell, 1992), a well-known approximate procedure based on nearest-neighbors; Markov blanket filtering (Koller and Sahami, 1996; Yu and Liu, 2004), which considers the special case of marginal dependencies (and is therefore fundamentally different from RMB, despite the similar name). All known methods are either approximate or have exponential time-complexity.

5. Conclusion

In this paper, we have explored an alternative approach to the feature selection (FS) problem: instead of designing suboptimal methods for the intractable full problem, we propose consistent and efficient (polynomial-time) methods for a restricted data distribution class. We find that a very mild

restriction to *strictly positive* distributions is sufficient for the MINIMAL-OPTIMAL problem to be tractable (Figure 1). Therefore, we conclude that it is tractable in most practical settings.

We have also identified a different feature selection problem, that of discovering all relevant features (ALL-RELEVANT). This problem is much harder than MINIMAL-OPTIMAL, and has hitherto received little attention in the machine learning field. With the advent of major new applications in the bioinformatics field, where identifying features *per se* is often a more important goal than building accurate predictors, we anticipate that ALL-RELEVANT will become a very important research problem in the future. We have herein provided a first analysis, proved that the problem is intractable even for strictly positive distributions, and proposed two consistent, polynomial-time algorithms for more restricted classes (Figure 1). We hope that these results will inspire further research in this novel and exciting direction.

Acknowledgments

We would like to thank the editor and the anonymous reviewers for helpful comments and insight, as well as for pointing out an initial problem with the proof of Lemma 19 and also suggesting a remedy to that problem. We also thank Timo Koski, Department of Mathematics, Linköping University for proof-reading and Peter Jonsson, Department of Computer and Information Science, Linköping University for valuable discussions. This work was supported by grants from the Swedish Foundation for Strategic Research (SSF), the Ph.D. Programme in Medical Bioinformatics, the Swedish Research Council (VR-621-2005-4202), Clinical Gene Networks AB and Linköping University.

Appendix A. Lemmas

For completeness, we here give a proof of the uniqueness of the Bayes classifier for strictly positive distributions.

Lemma 19 *For any strictly positive distribution $f(x,y)$, the Bayes classifier g^* is unique in the sense that, for every classifier g, it holds that $R(g) = R(g^*)$ iff the Lebesgue measure of $\{x : g(x) \neq g^*(x)\}$ is zero.*

Proof By Theorem 2.2 of Devroye et al. (1996), the risk of any classifier g can be written as

$$R(g) = R(g^*) + \int_{\mathcal{X}} \left| \max_y p(y|x) - \frac{1}{2} \right| 1_{\{g(x) \neq g^*(x)\}} f(x) dx.$$

By property (ii) of Definition 1, we have $\max_y p(y|x) \neq 1/2$ almost everywhere. Thus, the integral is zero iff the Lebesgue measure of $\{x : g(x) \neq g^*(x)\}$ is zero. ■

Lemma 20 *For any conditional distribution $p(y|x)$, it holds that*

$$P(p(Y|X_i,R_i) = p(Y|X_i',R_i)) = 1 \iff P(p(Y|X_i,R_i) = p(Y|R_i)) = 1$$

provided that X_i, X_i' are independent and identically distributed.

Proof Assume that the left-hand side holds. Then we must have

$$P(p(Y|X_i,R_i) = p_0) = 1$$

for some p_0 constant with respect to X_i. But

$$p(y|r_i) = \frac{f(r_i,y)}{f(r_i)} = \frac{\int_{\mathcal{X}_i} p(y|x) f(x)}{\int_{\mathcal{X}_i} f(x)} = \frac{p_0 \int_{\mathcal{X}_i} f(x)}{\int_{\mathcal{X}_i} f(x)} = p_0$$

with probability 1, which implies the right-hand side. The converse is trivial. ■

The following lemmas are needed for the correctness proof for the RMB algorithm.

Lemma 21 *Let $f(x)$ be any PCWT distribution, so that a unique undirected minimal I-map G of f exists. Then, for any shortest path $Z_1^m = \{Z_1, \dots Z_m\}$ between Z_1 and Z_m in G, it holds that $Z_1 \not\perp Z_m | X \setminus Z_1^m$.*

Proof The proof is by induction. For $m = 2$, the lemma follows immediately from the definition of the minimal I-map (Pearl, 1988). Also, it holds that

$$Z_i \not\perp Z_{i+1} | X \setminus \{Z_i, Z_{i+1}\} \tag{10}$$

$$Z_i \perp Z_j | X \setminus \{Z_i, Z_j\}, \quad j > i+1. \tag{11}$$

Take any distinct Z_i^{i+1}, Z_k and assume that $Z_i \perp Z_{i+1} | X \setminus \{Z_i, Z_{i+1}, Z_k\}$. Then $Z_i \perp \{Z_{i+1}, Z_k\} | X \setminus \{Z_i, Z_{i+1}, Z_k\}$ by contraction with (11), and therefore $Z_i \perp Z_{i+1} | X \setminus \{Z_i, Z_{i+1}\}$ by weak union. This contradicts (10), so we conclude that

$$Z_i \not\perp Z_{i+1} | X \setminus \{Z_i, Z_{i+1}, Z_k\}. \tag{12}$$

Next, take any sequence Z_i^{i+2}. Applying (12), we obtain $Z_i \not\perp Z_{i+1} | X \setminus Z_i^{i+2}$ and $Z_{i+1} \not\perp Z_{i+2} | X \setminus Z_i^{i+2}$. Using weak transitivity implies either $Z_i \not\perp Z_{i+2} | X \setminus Z_i^{i+2}$ or $Z_i \not\perp Z_{i+2} | X \setminus \{Z_i, Z_{i+2}\}$. The latter alternative contradicts (11), so we conclude

$$Z_i \not\perp Z_{i+2} | X \setminus Z_i^{i+2}. \tag{13}$$

Finally, take any Z_i, Z_j, Z_k such that neither Z_i, Z_j nor Z_j, Z_k are consecutive in the path Z_1^m. Using (11) with intersection (Theorem 25) and decomposition (Theorem 23), we find

$$\left.\begin{array}{l} Z_i \perp Z_j | X \setminus \{Z_i, Z_j\} \\ Z_j \perp Z_k | X \setminus \{Z_j, Z_k\} \end{array}\right\} \Longrightarrow Z_i \perp Z_j | X \setminus \{Z_i, Z_j, Z_k\}. \tag{14}$$

Equations (12),(13) and (14) show that the properties (10) and (11) hold also for the shortened path $Z'_1, \dots, Z'_{m-1}$ given by $Z'_1 = Z_1$ and $Z'_i = Z_{i+1}, 2 \leq i < m$ (removing Z_2). The lemma follows from (10) by induction. ■

Lemma 22 *For any PCWT distribution $f(x,y)$, a feature X_k is relevant iff there exists a path $Z_1^m = \{Z_1, \dots Z_m\}$ in the minimal I-map of $f(x)$ between some $Z_1 \in M = M(Y|X)$ and $Z_m = X_k$. In particular, for such a path it holds that $Y \not\perp Z_m | X \setminus Z_1^m$.*

Proof If $Z_m \in M$ (that is, $m = 1$), the lemma is trivial. Consider any $Z_m \notin M$. First, assume that there exists no path Z_1^m. Then $Z_m \perp M|S$ for any $S \subseteq X \setminus M \setminus \{Z_m\}$ by Lemma 21. Fix such an S. Since $Z_m \perp Y|M \cup S$, contraction and weak union gives $Z_m \perp M|\{Y\} \cup S$. Again using $Z_m \perp M|S$, weak transitivity gives

$$Z_m \perp Y|S \quad \vee \quad Y \perp M|S.$$

The latter alternative is clearly false; we conclude $Z_m \perp Y|S$. Next, fix any $S' \subseteq M$. By decomposition, $Z_m \perp M|S \implies Z_m \perp S'|S$. Combining with the above result, by the composition property

$$\left.\begin{array}{l} Z_m \perp S'|S \\ Z_m \perp Y|S \end{array}\right\} \implies Z_m \perp \{Y\} \cup S'|S.$$

Finally, weak union gives $Z_m \perp Y|S \cup S'$. Since $S \cup S'$ is any subset of $X \setminus \{Z_m\}$, we conclude that Z_m is irrelevant.

For the converse, assume that there exists a path Z_1^m. By Lemma 21, we have $Z_1 \not\perp Z_m|X \setminus Z_1^m$. Also, since $Z_1 \in M$ and $Z_2^m \cap M = \emptyset$, it holds that $Y \not\perp Z_1|S$ for any S that contains $M \setminus \{Z_1\}$. In particular, take $S = X \setminus Z_1^m$. Weak transitivity then gives

$$\left.\begin{array}{l} Z_1 \not\perp Z_m|X \setminus Z_1^m \\ Y \not\perp Z_1|X \setminus Z_1^m \end{array}\right\} \implies Z_m \not\perp Y|X \setminus Z_1^m \quad \vee \quad Z_m \not\perp Y|X \setminus Z_2^m.$$

But the latter alternative is false, since $X \setminus Z_2^m$ contains M by assumption. We conclude that $Z_m \not\perp Y|X \setminus Z_1^m$ and that Z_m is relevant. ■

Appendix B. Distribution Classes and Properties

The following two theorems are given by Pearl (1988) and concern any probability distribution.

Theorem 23 *Let R,S,T,U denote any disjoint subsets of variables. Any probability distribution satisfies the following properties:*

$$\begin{array}{ll} \textit{Symmetry:} & S \perp T|R \implies T \perp S|R \\ \textit{Decomposition:} & S \perp T \cup U|R \implies S \perp T|R \\ \textit{Weak union:} & S \perp T \cup U|R \implies S \perp T|R \cup U \\ \textit{Contraction:} & S \perp T|R \cup U \wedge S \perp U|R \implies S \perp T \cup U|R \end{array}$$

Theorem 24 *Let R,S,T,U denote any disjoint subsets of variables and let γ denote a single variable. Any DAG-faithful probability distribution satisfies the following properties:*

$$\begin{array}{ll} \textit{Composition:} & S \perp T|R \wedge S \perp U|R \implies S \perp T \cup U|R \\ \textit{Weak transitivity:} & S \perp T|R \wedge S \perp T|R \cup \gamma \implies S \perp \gamma|R \vee \gamma \perp T|R \end{array}$$

The next theorem is a slight modification of that found in Pearl (1988), adapted to our classification setting and our Definition 1.

Theorem 25 *Let R,S,T be disjoint subsets of X and let Y be the target variable. Any strictly positive distribution $f(x,y)$ satisfies the intersection property*

$$Y \perp R|(S\cup T) \ \wedge \ Y \perp T|(S\cup R) \Rightarrow Y \perp (R\cup T)|S.$$

Proof Using $Y \perp R|(S\cup T)$ we find

$$\begin{aligned} f(r,s,t,y) &= p(y|r,s,t)\,f(r,s,t) \\ &= p(y|s,t)\,f(r,s,t). \end{aligned}$$

Similarly, from $Y \perp T|(S\cup R)$ we find that $f(r,s,t,y) = p(y|s,r)\,f(r,s,t)$. Because $f(r,s,t) > 0$, it follows that $p(y|s,r) = p(y|s,t)$. Therefore both of these probabilities must be constant w.r.t. R and T, that is,

$$p(y|s,t) = p(y|s,r) = p(y|s).$$

Hence, $Y \perp R|S$ and $Y \perp T|S$ holds. The intersection property follows by contraction. ■

Corollary 26 *For any strictly positive distribution $f(x,y)$, the Markov boundary $M(Y|X)$ is unique.*

Proof Let S be the set of all Markov blankets of Y, $S = \{T \subseteq X : Y \perp X \setminus T|T\}$. Let T_1, T_2 be any two Markov blankets in S. By Theorem 25 the intersection property holds, so with $T' = T_1 \cap T_2$ we obtain

$$\left\{ \begin{array}{l} Y \perp X \setminus T_1|T' \cup (T_1 \setminus T') \\ Y \perp X \setminus T_2|T' \cup (T_2 \setminus T') \end{array} \right. \implies Y \perp X \setminus T'|T'$$

Hence T' is a Markov blanket of Y. Continuing in this fashion for all members of S, we obtain the unique $M(Y|X) = T_1 \cap T_2 \cap \cdots \cap T_{|S|}$. ■

References

Hussein Almuallim and Thomas G. Dietterich. Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, volume 2, pages 547–552, 1991. AAAI Press.

David A. Bell and Hui Wang. A formalism for relevance and its application in feature subset selection. *Machine Learning*, 41:175–195, 2000.

Avril L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2nd edition, 2002.

David Chickering and Christopher Meek. Finding optimal Bayesian networks. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence*, pages 94–102, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Thomas M. Cover and Jan M. van Campenhout. On the possible orderings of the measurement selection problem. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(9):657–661, 1977.

James E. Darnell. Transcription factors as targets for cancer therapy. *Nature Reviews Cancer*, 2: 740–749, 2002.

Luc Devroye, Lázló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

Nir Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303: 799–805, 2004.

Todd R. Golub, Donna K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and Eric S. Lander. Molecular classifiation of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

Douglas P. Hardin, Constantin Aliferis, and Ioannis Tsamardinos. A theoretical characterization of SVM-based feature selection. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Michael J. Holland. Transcript abundance in yeast varies over six orders of magnitude. *Journal of Biological Chemistry*, 277:14363–66, 2002.

Anil K. Jain and William G. Waller. On the optimal number of features in the classification of multivariate gaussian data. *Pattern Recognition*, 10:365–374, 1978.

George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129, 1994.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Technical report, Seminar für Statistik, ETH Zürich, Switzerland, 2005.

S. Sathiya Keerthi. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. *IEEE Transactions on Neural Networks*, 13(5):1225–1229, 2002.

Kenji Kira. and Larry A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 129–134, 1992.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97: 273–324, 1997.

Daphne Koller and Mehran Sahami. Towards optimal feature selection. In *Proceedings of the 13th International Conference of Machine Learning*, pages 248–292, 1996.

David J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Judea Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kauffman Publishers, Inc., San Fransisco, California, 1988.

José M. Peña, Johan Björkegren, and Jesper Tegnér. Scalable, efficient and correct learning of markov boundaries under the faithfulness assumption. In *Proceedings of the Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning under Uncertainty*, pages 136–147, 2005.

Jose M. Peña, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Identifying the relevant nodes before learning the structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 367–374, 2006.

Donna K. Slonim. From patterns to pathways: Gene expression comes of age. *Nature Genetics*, 32: 502–508, 2002. Supplement.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

Milan Studený. *Probabilistic Conditional Independence Structures*. Springer, 1st edition, 2004.

Ioannis Tsamardinos and Constantin Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2nd edition, 2000.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.