
Towards Principled Feature Selection: Relevancy, Filters and Wrappers

Ioannis Tsamardinos

Dept. of Biomedical Informatics
Vanderbilt University
ioannis.tsamardinos@vanderbilt.edu

Constantin F. Aliferis

Dept. of Biomedical Informatics
Vanderbilt University
constantin.aliferis@vanderbilt.edu

Abstract

In an influential paper Kohavi and John [7] presented a number of disadvantages of the filter approach to the feature selection problem, steering research towards algorithms adopting the wrapper approach. We show here that neither approach is inherently better and that any practical feature selection algorithm needs to at least consider the learner used for classification and the metric used for evaluating the learner’s performance. In the process we formally define the feature selection problem, re-examine the relationship between relevancy and filter algorithms, and establish a connection between Kohavi and John’s definition of relevancy to the Markov Blanket of a target variable in a Bayesian Network faithful to some data distribution. The theoretical results lead to principled ways of designing optimal filter algorithms of which we present one example.

1 Feature Selection for Classification

Feature (also called variable) selection for classification is an important problem that has been given considerable attention during the last three decades [8]. In this section we formally define two versions of the feature selection problem and discuss how previous definitions do not meet the needs of a practitioner.

The purpose of feature selection is many-fold: First, while theoretically, in the sample limit, the more features we have, the better, in learning experiments involving many practical algorithms a good selection of features often yields models with better generalization performance than when using the full feature set [5]. A second reason for desiring to reduce the number of variables required for learning is that they may be unnecessarily expensive to observe. For example, in

medicine, an unnecessary variable may cost hundreds or even thousands of dollars per observation, and may, in addition, entail risks for the patients’ health [4]. Third, parsimonious models are easier to understand and less computationally expensive for performing inference and prediction. Finally, when feature selection is used as a tool to deepen researchers’ understanding of the characteristics and structure of some domain (e.g., to orient them towards subsequent experimentation and eventually development of a detailed domain theory), unnecessary variables make such interpretation more difficult.

Despite significant research in the field, a standard and acceptable general definition of the feature selection problem has not yet been reached. According to a recent call for papers for the special issue on variable and feature selection of the Journal of Machine Learning Research (JMLR), the “Variable selection refers to the problem of selecting input variables that are most predictive of a given outcome”. Kohavi and John [7] (hereafter KJ) define it as the “subset of features such that the accuracy of the induced classifier ... is maximal”. A similar definition is given in [13]: the feature selection problem is to discover the set of features *and* the parameters of the classifier to be used such that it minimizes the expected loss, according to a loss function. Among other problems that we discuss below, these definitions do not distinguish among all possible features sets with maximal accuracy (or minimal loss) and cannot incorporate instances of the problem where the cost of features is important and a trade-off between accuracy and cost of observations is required.

Because of the lack of standard definitions it is hard to analyze arguments in favor or against feature selection algorithms, e.g., the argument of KJ against filter algorithms (defined below). Hence, we attempt a definition of the problem that enables the analysis of the theoretical properties of feature selection.

Definition 1. Feature Selection Problem 1 (FSP_1). A feature selection problem is a tuple

$\langle X, \Phi, T, A, M \rangle$, where X is a sample of input patterns defined over a feature set Φ , $T \in \Phi$ a target variable, A a classification algorithm producing a prediction model for T given T and X , and M a performance metric of the classifier’s model and the selected features. A *solution* to the problem is a feature subset $\phi \subseteq \Phi$ that maximizes $M(\phi, A(T, X \downarrow \phi))$, where $X \downarrow \phi$ is the projection of the data X on only the features in ϕ .

For technical simplicity and without loss of generality we impose the constraint that T always has to be included in the selected features. Typical feature selection algorithms that have appeared in the literature tackle specific instances of the above definition, but not the problem in its full generality. For example, the KJ definition matches the above one when M is any loss function and there is no preference among feature sets that exhibit the same loss (accuracy). Unlike their definition however, the metric M requires as a parameter the selected feature set to account not only for any loss function, but also to incorporate feature observation costs and possibly trade-off accuracy for smaller feature subsets.

Notice that the classification algorithm A in the definition is fixed and given as part of the problem. That is, a solution to the problem is a feature subset that optimizes the metric for the given learner A . However, *this is not the problem that practitioners would really like to solve*. The practitioner is free to choose any classifier and thus he or she is interested in optimizing the metric over *all* possible or available classifiers. The confusion over feature selection is depicted in the definition of feature selection, given by the editors of the recent JMRL’s special issue on feature selection, as the process of selecting the features “that are most predictive of a given outcome”: according to which classifier is prediction power measured by? Is it measured by a given classifier or is it the maximum over all classifiers? The above discussion suggests that a more appropriate definition is the following:

Definition 2. Feature Selection Problem 2 (FSP_2). A feature selection problem is a tuple $\langle X, \Phi, T, M \rangle$ with semantics as in Definition 1. A *solution* to the problem is a feature subset $\phi \subseteq \Phi$ and a learning algorithm A that maximizes $M(\phi, A(T, X \downarrow \phi), \phi)$.

To distinguish between the two definitions we will refer to the first problem as FSP_1 and the second as FSP_2 . There is no third definition possible that also maximizes over all possible metrics M because the metric is a characteristic of the problem: a practitioner is allowed to choose any known classifier A but whether the 0/1-Loss, the Mean-Squared-Loss, or some trade-off of accuracy and cost of features should be optimized depends on the nature of the problem

and the goals and ultimate real-life uses of the learnt model.

An established distinction of feature selection algorithms is KJ’s filters and wrappers.

Definition 3. Types of Feature Selection Algorithms. A **wrapper** feature selection algorithm for FSP_1 is a search procedure in the space of all possible feature subsets that uses the classification algorithm A and the evaluation metric M for assessing the quality of states (i.e., feature subsets). A **filter** variable selection algorithm is an algorithm that selects variables without evaluating the metric M on the output of the classification algorithm A . For FSP_2 , a wrapper is a search procedure in the space of all possible subsets and all possible classification algorithms.

Notice that a filter may use a learner A' and a metric M' to select features. In this case, the difference with a wrapper is that A' and M' may be different than the A and M used in the definition of the feature selection problem. If however $A' = A$ and $M' = M$, then the filter algorithm becomes a wrapper. One such example is the Recursive Feature Elimination algorithm (RFE) [6] where a linear Support Vector Machine (SVM) is trained on the data and then the first half of the features corresponding to the smallest weights in the vector normal to the optimal hyperplane is eliminated recursively. Of the $\log_2 |\Phi|$ linear SVMs models and corresponding feature sets produced this way, the one with the maximum performance is selected, where performance is measured by a combination of the success rate, the acceptance rate, and other SVM model characteristics. If the classifier A used to induce the final model is again a linear SVM and the same performance metric is used, then RFE is a wrapper method, otherwise it is used as a filter. In such settings the distinction between wrappers and filters blurs.

2 Relevancy and Filters

In this section we suggest that relevancy is defined over feature subsets instead of individual features, so that subsets labeled as relevant correspond to solutions of the feature selection problems and that a filter algorithm essentially implements a relevancy definition.

2.1 What is Relevancy for Feature Selection

Volume 97 of the Artificial Intelligence Journal was devoted to the concept of relevancy. Why are researchers interested in this concept (in the context of feature selection)? One justification is that it is perhaps the first step to feature selection: the semantics of “relevancy” suggest to our intuition that a relevant variable should be included in the selected variables and all irrelevant

variables should not. Moreover, an implicit consensus in the community is that relevancy can and should be defined independent of both the classifier to be used and the evaluation metric: the relevance of a variable should depend on the probability distribution of the data, not whether SVMs, for example, will be used to build the final model.

Formally, the relevant set of variables should be the *solution* to $FSP_1 \langle X, \Phi, T, A, M \rangle$ (or $FSP_2 \langle X, \Phi, T, M \rangle$) for a given X, Φ and T but independent of any learner A and performance metric M , at least in the sample limit of X where the joint distribution of X is a close approximation to the real distribution of the whole population of the data instances.

Since filters are independent of A and M , *each (computable) definition of relevancy corresponds to a number of exact or approximate filtering algorithms that implement the definition and return all relevant features*. Conversely, any filter algorithm corresponds to some definition of relevancy that only employs the distribution of the data.

If the scientific community could agree on a satisfying and appropriate definition of relevancy it is then a matter of designing efficient filters for determining the relevant set of variables. An attractive property of such an ideal definition would be that we can select the relevant variables independent of A and M . The above discussion suggests the following:

Definition 4. A *definition of relevancy* \mathcal{R}_T for target T is a set of functions $f : P_\Phi \rightarrow 2^\Phi$, where Φ is a feature set and P_Φ is the set of all possible distributions defined over Φ .

In other words, a definition of relevancy is a rule for labeling features as relevant or irrelevant given T and the probability distribution of the data. Indeed, all definitions given in the KJ paper [7] and in Blum and Langley [3] intentionally define functions $f : P_\Phi \rightarrow 2^\Phi$ for each feature set Φ and target T . We present some of these definitions in the sections to follow.

Notice however, that an FSP_1 might have two (or more) solutions ϕ_1 and ϕ_2 . In addition, a feature might belong to both of these sets, just one, or none. In case we label as relevant the features in the union of these two sets, the correspondance between relevant features and solutions to the FSP_1 is lost: $\phi_1 \cup \phi_2$ may no longer be a solution. The same is true for the intersection of the solution sets. This loss of correspondance can lead to confusion and further attempts to distinguish sets of features, e.g., KJ define weakly and strongly relevant features with the intent that strongly relevant ones are always required for maximum accuracy, while weakly relevant ones may or may not be needed.

We counter-suggest that a relevancy definition, instead of labeling individual features as relevant or irrelevant, should label whole feature subsets as relevant or irrelevant, specifically, those subsets that are solutions to the feature selection problem.

Definition 5. A *definition of relevancy* \mathcal{R}_T for a target T is a set of functions $f : P_\Phi \rightarrow 2^{2^\Phi}$ for each feature set Φ .

That is, f returns a set of feature subsets given the probability distribution of the data.

2.2 There are No Relevancy Definitions Independent of the Learner or Metric That Solve the FSP

In this section we examine the KJ argument against filter algorithms, we prove that every definition of relevancy necessarily needs to consider the classifier A and metric M , and we provide an alternative definition of relevancy.

KJ closely examine a number of previous definitions of relevancy and present examples where the definitions fail to classify variables as relevant with the desired properties. Subsequently, they provide an improved definition of relevancy that satisfies the intuition of relevancy over the previous examples, nevertheless, even for this definition there exist learners for which the set of relevant variables is not the solution to the variable selection problem. KJ argue convincingly that in the general case (i.e., when relevancy is not tied to a specific classifier) the concept of relevancy is of little use: *both* relevant and irrelevant features may be required for optimal classifications. They quote: “Relevance of a feature does not imply that it is in the optimal feature subset and, somewhat surprisingly, irrelevance does not imply that it should not be in the optimal feature subset”, and “Different algorithms have different biases and a feature that may help one algorithm may hurt another”. Thus, the feature selection process interacts with the classifier used to induce the final model.

Of course, when they talk about relevancy they mean their definition of it, but the implicit conjecture that lingers in the the paper is that *there is no definition of relevancy independent of the learner A* with the desired property that the relevant variables according to this definition are the solution to $FSP_1 \langle X, \Phi, T, A, M \rangle$ (they do not consider FSP_2). We now formally prove that :

Theorem 1. *There is no concept of relevancy \mathcal{R}_T defined independent of the metric M , such that the set of relevant feature subsets given a probability distribution p are solutions to the $FSP_1 \langle X, \Phi, T, A, M \rangle$ or $FSP_2 \langle X, \Phi, T, M \rangle$, where X a data sample drawn from p .*

Proof. Let R be the set of relevant feature subsets for target T , probability distribution p of the data, and according to the relevancy definition \mathcal{R}_T . Let $\phi_1 \notin R$ be a feature subset. Define $M(\phi_1, A(T, X)) = 1$ and $M(\phi, A(T, X)) = 0$, for any other feature subset ϕ . If there is no such set ϕ_1 , that means that \mathcal{R}_T labels all subsets as relevant. In this case, pick one feature selection subset arbitrarily and create a metric M which assigns score 1 to it, and 0 to any other subset (if no other subset exists then $\Phi = \{T\}$). In all cases not all relevant feature subsets maximize the metric. \square

The theorem proves that relevancy should take into consideration (i.e., should be a function of) the metric used in the definition of the feature selection problem. We now prove that for FSP_1 it should also take into consideration the algorithm A .

Theorem 2. *There is no concept of relevancy $\mathcal{R}_T(M)$ (i.e., given the metric M) defined independent of the classifier A , such that the set of relevant feature subsets given a probability distribution p are solutions to the $FSP_1 \langle X, \Phi, T, A, M \rangle$, where X is a data sample drawn from p .*

Proof. Let $M(\phi, A(T, X \downarrow \phi)) = -l$, where l is the 0/1-Loss of the classifier A , thus M maximizes the accuracy of the classifier. Let $\Phi = \{T, V\}$, T and V binary with distribution p such that $P(T = 1|V = 1) = 0.8$, $P(T = 1|V = 0) = 0.4$, and $P(V = 1) = 0.5$ from which we infer that $P(T = 1) = 0.6$. Let S be a Simple Bayes classifier (SBC). Any definition of relevancy may label $\{V, T\}$, $\{T\}$, or both sets as the relevant feature subsets. Selecting $\{V, T\}$ has the minimum 0/1-Loss and $\{T\}$ the maximum for the SBC S in the sample limit. Thus, given enough sample $M(S(X \downarrow \{V, T\}, T), \{V, T\}) > M(S(X \downarrow \{T\}, T), \{T\})$. Let S' be the classifier that runs S on the input and returns the reverse classification from S .

If the definition of relevancy returns $\{T\}$ as the only relevant variable, then for the algorithm S there is a better subset $\{V, T\}$. If a definition of relevancy returns $\{V, T\}$ as the relevant feature subset, then for algorithm S' there is a better subset $\{T\}$ (since S performs worse on $\{T\}$ so S' performs better). It is wrong (or at least uninformative) for a definition of relevancy to return both sets as relevant since they get different scores for different algorithms. \square

In other words, there are cases where even if we know the metric M and the probability distribution p of the data, there is a classifier A such that we could still not label the feature sets as relevant or irrelevant in such a way that there exists an algorithm that solves the feature selection problem.

KJ's implicit conjecture was right. Knowing the metric M is a necessary condition for defining relevancy for both feature selection problems and knowing the classifier A is a necessary condition for defining relevancy for FSP_1 . In the same fashion, *knowing M and A is a necessary condition when designing optimal filter algorithms.*

To the significant number of definitions of relevancy for feature selection that have appeared in the literature, and in particular are mentioned in [3, 7] we counter-suggest:

Definition 6. A feature subset ϕ is *relevant* for FSP_1 for target T , classifier A , metric M and data X if it is the solution to the problem FSP_1 . Similarly, for FSP_2 .

3 No Free Lunch Theorems for Wrappers

In this section we consider limitations of the wrapper approach and in particular prove they are subjected to the constraints of the No Free Lunch Theorem [14] for search optimization.

Wrapper approaches display two considerable shortcomings: (a) they require¹ training and evaluation of the performance of the classifier used for every variable subset considered during search, which is computationally expensive, and (b) it is necessary to repeat the feature selection for every different classifier used to solve the classification problem (for FSP_1 wrappers). The advantage of the wrapper approach is that if the whole feature subset space for FSP_1 is explicitly or implicitly searched, then they are guaranteed to discover the optimal feature subset for classification for any learner, evaluation metric, and data distribution. Nevertheless, on all but the smallest problems an exhaustive search is computationally prohibited and so *wrappers provide no optimality guarantees.*

Wrappers that are designed independent of the algorithm and the metric used treat the objective function $M(\phi, A(T, X \downarrow \phi))$ to be optimized as a black-box. On each FSP_1 they perform a heuristic search in the space of all possible feature subsets. As black-box optimization searches, wrappers are subject to the results of the No Free Lunch (NFL) theorem for optimization [14] that we are about to discuss.

NFL states that for any measure of performance (e.g., proximity of the output value to the maximum value) a black-box optimization search is as good as any other, when averaged out on all possible (but finite number

¹We have already mentioned that filters may also train and evaluate classifiers in their attempt to select features, but in the case of filters this is optional.

of) objective functions (problem landscapes). Specifically, let \mathcal{X} be a finite search space of states, \mathcal{Y} a finite space of objective values, and $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ the set of all possible objective functions defined on these two spaces. NFL states that the performance of any pair of algorithms A_1 and A_2 averaged out on all $f \in \mathcal{F}$ is the same. For a wrapper, the objective function is the metric M that for all practical purposes takes a finite number of output values, e.g., all numbers that can be represented by 32 or 64 bits on some machine.

Theorem 3. *NFL holds for FSP_1 wrappers if the choice of the metric M or the classifier A is unconstrained.*

Proof. \mathcal{X} and \mathcal{Y} are both finite. It suffices to show that any objective function $f \in \mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ is realizable. When the metric M is unconstrained and it can be any function, then for any $f \in \mathcal{F}$, with $f(\phi) = a_\phi \in \mathcal{Y}$ we can simply assign $M(\phi, A(T, X \downarrow \phi)) = a_\phi$.

When the metric M is given and fixed, we need to show for any function $f = M(\phi, A(T, X \downarrow \phi)) \in \mathcal{Y}^{\mathcal{X}}$, with $f(\phi) = a_\phi$, there is a distribution p of data and an algorithm A that realizes it.

Let M be minus the 0/1-Loss of the classifier A (accuracy), let the distribution p be $P(V_i = 2^i) = 1$ for any feature $V_i \in \Phi \setminus \{T\}$, and $P(T = 1) = 1$. Let the classifier A consist of rules of the form: “if $\sum_{i=0}^n V_i = j$, then output b_j ”, for $j = 1 \dots 2^{|\Phi|+1}$ (we assume states containing A are included in the search space). Notice that each rule j fires for one and only feature subset ϕ_j , namely the one whose binary encoding is the number j , e.g., feature subset 1101 containing the first, third, and fourth feature (starting from the right) corresponds to $j=13=2^0 + 2^2 + 2^3$. By setting b_j to 1 we get $M(\phi_j, A(T, X \downarrow \phi_j)) = 1$ and by setting b_j to 0, we get $M(\phi_j, A(T, X \downarrow \phi_j)) = 0$. For $\mathcal{Y} = \{0, 1\}$ and $a_\phi \in \{0, 1\}$ we get $f(\phi) = a_\phi$ when $a_\phi = b_j$. Thus, any function is realizable even if the metric M is fixed. \square

4 Implications for Designing Filters and Wrappers

All of our theorems and KJ’s arguments employ extreme classifiers to make their point. For example, KJ show that there is a handicapped classifier (the limited perceptron classifier [9]) that requires irrelevant variables to be included in the selected features for some problems, where relevance is defined according to Definition 9. But, that does not mean that there are no filtering algorithms for FSP_1 that perform well with the problems and classifiers that are typically used in practice. In a similar fashion, just because NFL holds for wrappers does not mean that averaged out on *all*

problems that occur in practice there is no wrapper with superior performance over all other wrappers.

KJ’s argument is valid in principle for FSP_1 . However, as we have already mentioned in practice we are interested in solving FSP_2 . It is very unlikely that a researcher or modeler will use a filtering algorithm and then apply *only* the limited perceptron classifier without trying any other more powerful classifiers. Even though FSP_2 was not spelled out by researchers as the real problem they try to tackle, in a number of feature selection papers *a manual meta-search is performed over a limited number of classifiers*, essentially attempting to optimize the metric over both feature subsets and classifiers simultaneously. For example, KJ use wrappers on both the Naive Bayes and Decision Trees. A practitioner would choose the feature subset and algorithm that maximizes the accuracy.

The conclusions we draw from this discussion are the following: in principle both filters and wrappers need to consider the metric and algorithm in order to be optimal and efficient respectively. *Thus, designing feature selection algorithms should target specific classes of metrics and algorithms.* Neither approach to feature selection is inherently superior or should be dismissed.

5 Designing Optimal Filters

Armed with a new understanding of relevancy and filters, we now proceed to design optimal filters for special cases. First, the relevancy definitions of KJ and the concept of Markov Blanket of a target variable T , $MB(T)$ are presented. The relationship between the relevant features and $MB(T)$ is explored. $MB(T)$ is the solution of several classes of feature selection problems.

Regarding notation, we will denote a variable by $V_i \in \Phi$, its values by v_i , the target variable T , a value of T by t , the remaining set of variables by S_i , i.e., $S_i = \Phi \setminus \{V_i, T\}$, and a joint value of S_i as s_i . We will also use the shorthand $P(T | V_1, V_2) = P(T | V_2)$ to denote that for every instantiation t of T , and every instantiation v_1 , and v_2 of the set of variables V_1 and V_2 the following equation holds: $P(T = t | V_1 = v_1, V_2 = v_2) = P(T = t | V_2 = v_2)$. Similarly, $P(T | V_1, V_2) \neq P(T | V_2)$ expresses the fact that there is an instantiation of T, V_1 , and V_2 for which the equation does not hold. Also, we will denote the conditional independence of T and V_1 given V_2 as $\mathbf{I}(T; V_1 | V_2) \equiv P(T | V_1, V_2) = P(T | V_2)$. Finally, to avoid technical difficulties when the conditional probability $P(T | V)$ is not defined (which is the case when $P(V) = 0$), we will assume that the joint probability distribution has no structural zeros, i.e., all the possible instantiations of the variables Φ have a positive

probability, no matter how small. According to KJ [7]:

Definition 7. KJ-Strong relevancy. A variable V_i is **KJ-strongly relevant** to T if and only if there exists some v_i, t , and s_i for which $p(V_i = x_i, S_i = s_i) > 0$, such that $p(T = t | V_i = v_i, S_i = s_i) \neq p(T = t | S_i = s_i)$.

Definition 8. KJ-Weak relevancy. A variable V_i is **KJ-weakly relevant** to T if and only if it is not KJ-strongly relevant, and there exists a subset of variables S'_i of S_i for which there exists some v_i, t , and s'_i with $p(V_i = v_i, S'_i = s'_i) > 0$ such that: $p(T = t | V_i = v_i, S'_i = s'_i) \neq p(T = t | S'_i = s'_i)$.

Definition 9. KJ-Relevancy. A feature is **KJ-relevant** to T if it is KJ-weakly or KJ-strongly relevant to T . A feature is **KJ-irrelevant** to T if it not KJ-relevant to T .

Definition 10. Markov Blanket. The Markov Blanket of T , denoted as $MB(T)$ is a minimal set of variables, such that every other variable is independent of T given $MB(T)$, i.e., $\forall V \in \Phi \setminus \{V, T\}, P(T | MB(T), V) = P(T | MB(T))$.

It turns out that the $MB(T)$ is unique and coincides with the set of KJ-strongly relevant features in distributions that are faithful to some Bayesian Network. In distributions not faithful to any BN the KJ-strongly relevant features are the ones that belong in the intersection of the Markov Blankets. We now define these concepts and prove the properties.

Definition 11. Bayesian Network $\langle \Phi, G, J \rangle$. Let Φ be a set of discrete variables and J be a joint probability distribution over all possible instantiations of Φ . Let G be a directed acyclic graph over a set of variables $S \subset \Phi$. Let all nodes of G correspond one-to-one to members of Φ . We require that for every node $V \in \Phi$, V is probabilistically independent of all non-descendants of V , given the parents of V (**Markov Condition**). Then we call the triplet $\langle \Phi, G, J \rangle$ a Bayesian Network (BN) [10].

A well known property of BNs is that *any* joint probability distribution can be represented with a BN.

Definition 12. d -separation. Two variables V_1 and V_2 are *d -separated* given a set of variables V_3 in a BN if and only if there exists no adjacency path p (i.e., a path ignoring the ordering of the edges) such that (i) every collider of p (a collider being a node with two incoming edges that belong in the path) is in V_3 or has a descendant in V_3 , and (ii) no other nodes on path p are in V_3 [11].

From the definition and graph theory we infer:

Proposition 1. A variable with a direct edge to T is never d -separated given any subset of the variables; a

parent of a common child with T is never d -separated given any subset of the variables that contains the common child.

Definition 13. Faithfulness. The graph G of some BN N is faithful to a joint probability distribution J over feature set V if and only if every dependence entailed by G is also present in J . We say that a data-generating process K is faithfully represented by N , if K in the sample limit produces data with joint probability distribution P , and N is faithful to P . A BN N is faithful if there is a probability distribution J to which it is faithful.

It follows from the Markov Condition that every conditional independence entailed by G is also present in J . Thus, together Faithfulness and the Markov Condition establish a close relationship between a graph G and some probability distribution J and allow us to associate statistical properties of J with graph properties of G . In the terminology of Spirtes, Glymour, and Scheines [12] G and J are *faithful to one another*, and in the terminology of Pearl [11] G is a *perfect-map* of P and P is a *DAG-isomorph* of G .

Proposition 2. In a faithful BN, d -separation captures all conditional dependence and independence relations that are encoded in the graph [11, 10] which implies that two nodes are d -separated with each other given V , if and only if they are conditionally independent given V .

Theorem 4. *The unique $MB(T)$ in a faithful BN is the set of parents, children, and parents of children of T .*

Proof. Neapolitan shows [10] that the set of parents, children, and parents of children of T d -separates T from any other variable. Let us call this set L and prove that it is minimal and thus a $MB(T)$. By Proposition 1 nothing d -separates a parent or a child of T so we cannot remove them from L . Again, by Proposition 1, if the children belong in L nothing can d -separate a parent of a common child with T . Thus, parents of common children also cannot be removed from L . We now prove that L is also unique. Assume that another set $L' \neq L$ d -separates T from any other variable. With a similar argument as above, we see that L' has to contain all parents and children of T , and also all parents of children of T . Thus, $L \subseteq L'$ and since L is minimal, $L' = L$ and so L is unique. \square

The next theorems associate KJ-relevancy and $MB(T)$.

Theorem 5. *In a faithful BN, a variable $V \in \Phi$ is KJ-strongly relevant, if and only if $V \in MB(T)$.*

Proof. Suppose that V_i is KJ-strongly relevant, but does not belong in $MB(T)$. Recall that in the definition of KJ-strong relevancy (Definition 7), $S_i = \Phi - \{V_i, T\}$, and so $MB(T) \subseteq S_i$. Since $MB(T)$ is a subset of S_i it follows that $P(T = t | V = v, S_i = s_i) = P(T = t | S_i)$. Therefore, according to Definition 7, V_i can never be KJ-strongly relevant, contrary to what we assumed.

Conversely, we can prove that if $V_i \in MB(T)$, then it is KJ-strongly relevant. By the definition of d -separation (Definition 12), we can see that each member V_i of $MB(T)$ is not d -separated from T given S_i , i.e., given the remaining set of variables. In turn this implies (by Proposition 2) that T and V_i are conditionally dependent given S_i , i.e., $P(T = t | V_i, S_i) \neq P(T = t | S_i)$ and so V_i is KJ-strongly relevant. \square

Theorem 6. *In a faithful BN, a variable $V \in \Phi$ is KJ-weakly relevant, if and only if it is not KJ-strongly relevant and there is an undirected path from V to T .*

Proof. Consider an undirected path p from V to T and let Z be the set of colliders in p . Then V and T are not d -separated given Z , so they are conditionally dependent (Proposition 2) and thus, $P(T | V, Z) \neq P(T | Z)$. By definition then, if V is not KJ-strongly relevant, it is KJ-weakly relevant. Conversely, if V is KJ-weakly relevant, then there is a set Z such that $P(T | V, Z) \neq P(T | Z)$. If there is no path between V and T then they are d -separated given any set, and by Proposition 2 conditionally independent given any set and so V cannot be KJ-weakly relevant. \square

A corollary of the above is that KJ-irrelevant features have no path to T in the BN faithful to the probability distribution. If faithfulness and thus Proposition 2 do not hold (as is typical when there are deterministic relations for example) it is possible that there are multiple Markov Blankets for T , $MB_1(T), \dots, MB_n(T)$. In this case the KJ-strongly relevant features are the ones in the intersection of all the Markov Blankets, i.e., the set $\bigcap_i MB_i(T)$ (we omit a proof due to lack of space).

6 $MB(T)$ as the Solution of a Feature Selection Problem

This section summarizes the conditions under which $MB(T)$ is the solution to feature selection problems.

By definition, $MB(T)$ carries all information required to estimate the probability distribution of T given the data. The exact distribution is required *only* for calibrated classification, i.e., when the output of the classifier is not the most probable class of T , but the distribution over class membership. Calibration is re-

quired in many learning applications such as when cost-sensitive decisions must be made and corresponds to Mean-Squared Loss. For example, in order to apply decision theory, an agent should know the probability distribution of T and not just the most probable classification.

It is quite probable that when 0/1-Loss is used instead as the metric only some of the features in $MB(T)$ are required or features that do not belong in $MB(T)$. But, for calibrated classification *all* features in $MB(T)$ are required.

Proposition 3. $MB(T)$ is the solution to $FSP_1 \langle X, \Phi, T, A, M \rangle$ in the sample limit of X , where X is drawn from a faithful BN, A is any calibrated classifier that can approximate any probability distribution² and M is a metric strictly decreasing with the Mean-Squared Loss with a preference for smaller subsets. $MB(T)$ and A is the solution to the $FSP_2 \langle X, \Phi, T, M \rangle$. If X is not drawn from a faithful BN then the solution to the FSP_1 problem above is the smallest among all $MB(T)$.

7 An Optimal Filter Algorithm

The definitions and theoretical results presented above are not just of academic interest. They directly lead to the design of optimal filter algorithms for the special case of Proposition 3: any algorithm that provably identifies $MB(T)$ is an optimal filter algorithm under the conditions stated in the proposition.

We now present the Incremental Association Markov Blanket (IAMB) algorithm (Figure 1). IAMB was first introduced in [2] (available by request from the authors) and identifies the $MB(T)$ under the following assumptions: **1.** All data is generated by processes that can be faithfully represented by Bayesian Networks. **2.** There exist reliable statistical conditional independence tests and measures of associations for checking independence and strength of association of T with some other variable X given a set of variables Y . When the assumptions are violated its output serves as a heuristic approximation of the MB. Experimental results on IAMB are reported elsewhere [2].

IAMB consists of both a forward phase and a backward phase. An estimate of the MB is kept in the set **CMB**. In the forward phase all the variables that belong in $MB(T)$ and possibly more (false positives) enter **CMB**, while in the backward phase the false positives are identified and removed so that **CMB**= $MB(T)$ in the end.

The heuristic used in IAMB to identify potential MB

²For example A , can be calibrated Neural Networks, Bayesian Network Learners, etc.

```

Phase I (forward)
CMB =  $\emptyset$ ; Cont = True
While Cont = True
  Cont = False
   $F = \arg \max_{V \in \Phi - \{T\} - \mathbf{CMB}} \text{assoc}(V; T | \mathbf{CMB})$ 
  If  $\neg \mathbf{I}(F; T | \mathbf{CMB})$ 
    CMB = CMB  $\cup$   $F$ 
    Cont = True
  End If
End While

Phase II(backward)
For each variable  $F$  in CMB
  If  $\mathbf{I}(F; T | \mathbf{CMB} - \{F\})$ 
    Remove  $F$  from CMB
  EndIf
EndFor
return CMB

```

Figure 1: The Incremental Association Markov Blanket (**IAMB**) Algorithm.

members in phase I is the following: start with an empty candidate set for the $MB(T)$, i.e., $\mathbf{CMB} = \emptyset$, and admit into it (in the next iteration) the variable that has the largest association with T conditioned on \mathbf{CMB} . Function *assoc* in the figure measures the strength of association between F and T given the features in \mathbf{CMB} . We stop when the association of every variable conditioned on \mathbf{CMB} vanishes (F and T are independent given \mathbf{CMB} , i.e., $\mathbf{I}(F; T | \mathbf{CMB})$). This heuristic is admissible in the sample limit because all members of MB will enter \mathbf{CMB} eventually.

A number of parametric and non-parametric measures of associations and conditional independence tests can be used to implement functions *assoc* and \mathbf{I} in the figure, that are sound in the sample limit under various data sampling assumptions [1]. In the sample limit IAMB will provably output the correct MB using any of these metrics (see proof in [2]).

8 Conclusions

In this paper we re-examine the concepts of relevancy, the feature selection problem and the distinction between wrappers and filters. We prove there is no concept of relevancy, defined independent of either the classifier used for the final induced model or the metric used for evaluating performance, that corresponds to solutions of the feature selection problem. Thus, filter algorithms need to consider these two parameters to be optimal. Similarly, we prove that wrappers are subject to the No Free Lunch theorem unless they consider these two parameters. *Optimal feature selection is possible only for special cases; design of optimal feature selection algorithms is attainable only by constraining the application domain in terms of classifiers*

and loss functions used and tailoring the algorithms in those terms. For calibrated classification, the Markov Blanket of the target variable is the optimal feature set. The Markov Blanket corresponds to the strongly relevant features, as defined by Kohavi and John, in data faithful to some Bayesian Network. We present an algorithm that provably discovers the Markov Blanket and thus optimally solves a special case of the feature selection problem.

References

- [1] A Agresti. *Categorical Data Analysis: Probability and Mathematical Statistics*. John Wiley and Sons, 1990.
- [2] C. F. Aliferis and Ioannis Tsamardinos. Markov blanket induction for feature selection. Technical Report DSL-02-02, Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, 2002.
- [3] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.
- [4] G. F. Cooper and et al. An evaluation of machine learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9, 1997.
- [5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, second edition, 2001.
- [6] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
- [7] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
- [8] Huan Liu and Hiroshi Motoda, editors. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. International Series in Engineering and Computer Science. Kluwer, 1998.
- [9] Marvin L. Minsky and S. Papert. *Perceptrons: an Introduction to Computational Geometry*. MIT Press, expanded edition, 1988.
- [10] R. E. Neapolitan. *Probabilistic Reasoning in Expert Systems*. John Wiley and Sons, 1990.
- [11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.
- [12] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, second edition, 2000.
- [13] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for SVMs. In *NIPS*, pages 668–674, 2000.
- [14] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.