On Interacting Features in Subset Selection

Zheng Zhao
Arizona State University, USA

Huan Liu
Arizona State University, USA

INTRODUCTION

The high dimensionality of data poses a challenge to learning tasks such as classification. In the presence of many irrelevant features, classification algorithms tend to overfit training data (Guyon & Elisseeff, 2003). Many features can be removed without performance deterioration, and feature selection is one effective means to remove irrelevant features (Liu & Yu, 2005). Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. Usually a feature is relevant due to two reasons: (1) it is strongly correlated with the target concept; or (2) it forms a feature subset with other features and the subset is strongly correlated with the target concept. Optimal feature selection requires an exponentially large search space ($O(2^n)$, where n is the number of features) (Almual-lim & Dietterich, 1994). Researchers often resort to various approximations to determine relevant features, and in many existing feature selection algorithms, feature relevance is determined by correlation between individual features and the class (Hall, 2000; Yu & Liu, 2003). However, a single feature can be considered irrelevant based on its correlation with the class; but when combined with other features, it can become very relevant. Unintentional removal of these features can result in the loss of useful information and thus may cause poor classification performance, which is studied as attribute interaction in (Jakulin & Bratko, 2003). Therefore, it is desirable to consider the effect of feature interaction in feature selection.

BACKGROUND

The goal of feature selection is to remove irrelevant features and retain relevant ones. We first give the definition of feature relevance as in (John et al., 1994).

Definition 1 (Feature Relevance):

Let F be the full set of features, F_i be a feature and $S_i = F - \{F_i\}$. Let $P(C|S)$ denote the conditional probability of class C given a feature sets. A feature F_i is relevant *iff*

$$\exists S_i' \in S_i, \text{ such that } P(C \mid F_i, S_i') \neq P(C \mid S_i') \qquad (1)$$

Definition 1 suggests that a feature can be relevant, if its removal from a feature set reduces the prediction power of the feature set. A feature, whose removal does not reduce the prediction power of any feature set, is an irrelevant feature and can be removed from the whole feature set without any side-effect. From Definition 1, it can be shown that a feature can be relevant due to two reasons: (1) it is strongly correlated with the target concept; or (2) it forms a feature subset with other features and the subset is strongly correlated with the target concept. If a feature is relevant because of the second reason, there exists feature interaction. Feature interaction is characterized by its irreducibility (Jakulin & Bratko, 2004). We give the definition of *k*th-order below.

Definition 2 (*k*th order Feature Interaction):

Let F be a feature subset with k features $F_1, F_2, \ldots, F_k$. Let $\Im$ denote a metric that measures the relevance of a feature or a feature subset with the class label. Features $F_1, F_2, \ldots F_k$ are said to interact with each other *iff*: for an arbitrary partition $S = \{S_1, S_2, S_3, \ldots, S_l\}$ of F, where $2 \leq l \leq k$ and $S_i \neq \phi$, we have $\forall i \in [1, l]$, $\Im(F) > \Im(S_i)$.

It is clear that identifying either relevant features or *k*th-order feature interaction requires exponential time. Therefore Definitions 1 and 2 cannot be directly applied to identify relevant or interacting features when the dimensionality of a data set is huge. Many efficient feature selection algorithms identify relevant features based on the evaluation of the correlation between the

class and a feature (or a current, selected feature subset). Representative measures used for evaluating feature relevance includes (Liu & Motoda, 1998): distance measures (Kononenko, 1994; Robnik-Sikonja & Kononenko, 2003), information measures (Fleuret, 2004), and consistency measures (Almuallim & Dietterich, 1994), to name a few. Using these measures, feature selection algorithms usually start with an empty set and successively add "good" features to the selected feature subset, the so-called sequential forward selection (SFS) framework. Under this framework, features are deemed relevant mainly based on their individually high correlations with the class, and relevant interacting features of high order may be removed (Hall, 2000; Bell & Wang, 2000), because the irreducible nature of feature interaction cannot be attained by SFS. This motivates the necessity of handling feature interaction in feature selection process.

MAIN FOCUS

Finding high-order feature interaction using Definitions 1 and 2 entails exhaustive search of all feature subsets. Existing approaches often determine feature relevance using the correlation between individual features and the class, thus cannot effectively detect interacting features. Ignoring feature interaction and/or unintentional removal of interacting features might result in the loss of useful information and thus may cause poor classification performance. This problem arouses the research attention to the study of interacting features. There are mainly two directions for handling feature interaction in the process of feature selection: using *information theory* or through *margin maximization*.

Detecting Feature Interaction via Information Theory

Information theory can be used to detect feature interaction. The basic idea is that we can detect feature interaction by measuring the information loss of removing a certain feature. The measure of information loss can be achieved by calculating *interaction information* (McGill, 1954) or McGill's multiple mutual information (Han, 1980). Given three variables, *A, B* and *C*, the interaction information of them is defined as:

$$I(A;B;C) = H(AB) + H(BC) + H(AC) - H(A) - H(B) - H(C) - H(ABC)$$
$$= I(A,B;C) - I(A;C) - I(B;C) \quad (2)$$

Here $H(\cdot)$ denotes the entropy of a feature or a feature set. $I(X; Y)$ is the mutual information between X and Y, where X can be a feature set, such as $\{X_1, X_2\}$. Interaction information among features can be understood as the amount of information that is common to all the attributes, but not present in any subset. Like mutual information, interaction information is symmetric, meaning that I(A;B;C) = I(A;C;B) = I(C;B;A) = …. However, interaction information can be negative.

If we set one of the features in the interaction information to be the class, then the interaction information can be used to detect the *2*-way feature interaction as defined in Definition 2. $\Im$, the metric that measures the relevance of a feature or a feature subset with the class label, is defined as the mutual information between a feature or a feature set and the class. Positive interaction information indicates the existence of interaction between features.

Using the interaction information defined in Formula (2), we can only detect *2*-way feature interaction. To detect high order feature interaction, we need to generalize the concept to interactions involving an arbitrary number of attributes. In (Jakulin, 2005) the *k-way interaction information* is defined as:

$$I(S) = -\sum_{T \subseteq S} (-1)^{|S \setminus T|} H(T) \quad (3)$$

Where S is a feature set with k features in it, T is a subset of S and "\" is the set difference operator. $|\cdot|$ measures the cardinality of the input feature set, and $H(T)$ is the entropy for the feature subset T and is defined as:

$$H(T) = -\sum_{v \in T} P(v) \log_2 P(v) \quad (4)$$

According to Definition 3, the bigger the $I(S)$ is, the stronger the interaction between the features in S is. The k-way multiple mutual information defined in (Jakulin, 2005) is closely related to the lattice-theoretic derivation of multiple mutual information (Han, 1980), $\Delta h(S) = -I(S)$, and to the set-theoretic derivation of multiple mutual information (Yeung, 1991) and co-information (Bell, 2003) as $I'(S) = (-1)^{|S|} \times I(S)$.