Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Can high-order dependencies improve mutual information based feature selection?

Nguyen Xuan Vinh *[,1], Shuo Zhou, Jeffrey Chan, James Bailey

Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised form
7 September 2015
Accepted 10 November 2015

Keywords:
Feature selection
Mutual information
High-order dependency

ABSTRACT

Mutual information (MI) based approaches are a popular paradigm for feature selection. Most previous methods have made use of low-dimensional MI quantities that are only effective at detecting low-order dependencies between variables. Several works have considered the use of higher dimensional mutual information, but the theoretical underpinning of these approaches is not yet comprehensive. To fill this gap, in this paper, we systematically investigate the issues of employing high-order dependencies for mutual information based feature selection. We first identify a set of assumptions under which the original high-dimensional mutual information based criterion can be decomposed into a set of low-dimensional MI quantities. By relaxing these assumptions, we arrive at a principled approach for constructing higher dimensional MI based feature selection methods that takes into account higher order feature interactions. Our extensive experimental evaluation on real data sets provides concrete evidence that methodological inclusion of high-order dependencies improve MI based feature selection.

1. Introduction

Feature selection is an important task in data mining and knowledge discovery. Effective feature selection can improve performance while reducing the computational cost of learning systems. In this paper, we focus on mutual information (MI) based feature selection, which is a very popular *filter* paradigm. Compared to *wrapper* and *embedded* approaches [1], filter methods, such as those based on the MI criteria, are generally less optimized, but possess the major advantage of being learning-model independent and also typically less computationally intensive.

MI based feature selection is concerned with identifying a subset $\mathbf{S}$ of m features $\{X_1,\ldots,X_m\}$ within the original set $\mathbf{X}$ of M features in a data set, that maximizes the multidimensional joint MI between features and the class variable C, defined as

$$I(\mathbf{S};C) \triangleq \sum_{X_1,\ldots,X_m,C} P(X_1,\ldots,X_m,C)\log\frac{P(X_1,\ldots,X_m,C)}{P(X_1,\ldots,X_m)P(C)} \quad (1)$$

This criterion possesses a solid theoretical foundation, in that the MI can be used to write both an upper and lower bound on the Bayes error rate [2,3]. Nevertheless, the problems of estimating high-dimensional joint MI, and more generally estimating high-dimensional probability distribution, especially from small samples, are long-standing challenges in statistics. Therefore, a rich body of work in the MI-based feature selection literature approaches this difficulty by approximating the high-dimensional joint MI with low-dimensional MI terms. A particularly popular and successful class of methods makes use of the following criterion, which is the combination of low-dimensional MI terms known as '*relevancy*' and '*redundancy*',

$$f(X_m) \triangleq I(X_m;C) - \beta\sum_{X_j\in\mathbf{S}} I(X_m;X_j) \quad (2)$$

Under this framework, the features are often selected in an incremental manner: given a set $\mathbf{S}$ of $m-1$ already selected features $\{X_1,\ldots,X_{m-1}\}$, the next feature X_m is selected so that $f(X_m)$ is maximized. The term $I(X_m;C)$ measures the relevancy of X_m to the class variable C, while $\sum_{X_j\in\mathbf{S}} I(X_m;X_j)$ quantifies the redundancy between X_m and the selected features in $\mathbf{S}$, and β plays the role of a balancing factor. Many MI-based feature selection heuristics can be shown to be variations of (2) [3], including highly influential methods such as the Mutual Information Feature Selection (MIFS) criterion ($\beta\in[0,1]$) [4], and the Minimum Redundancy Maximum Relevance (MRMR) criterion ($\beta=1/|\mathbf{S}|$) [5].

It is noted that the two-dimensional MI can only detect pairwise variable interactions, either between two features or between a feature and the class variable. More complicated variable interactions cannot be identified with the two-dimensional MI. Fig. 1 provides an illustrative example of two variables (switches) that

* Corresponding author. Tel.: +614 3220 8948.
E-mail address: vinh.nguyen@unimelb.edu.au (N.X. Vinh).
[1] Postal address: Department of Computing and Information Systems, The University of Melbourne, Parkville VIC 3010, Australia.

http://dx.doi.org/10.1016/j.patcog.2015.11.007
0031-3203/

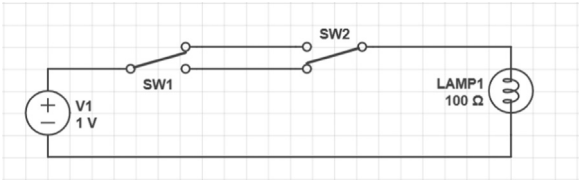

Fig. 1. An example of high-order variable interaction.

jointly control the target variable (the lamp). Knowing the state of either switch alone provides no information about whether the lamp is on or off. Only the joint state of both switches provides comprehensive knowledge on the state of the lamp. The pairwise mutual information cannot detect this type of multi-variable interaction.

To address this shortcoming, several works have considered the use of higher-dimensional MI quantities, such as the joint relevancy $I(X_iX_j;C)$ [6], the conditional relevancy $I(X_i;C|X_j)$ [3] and the conditional redundancy $I(X_i;X_j|C)$ [7]. Brown et al. [3] showed that many such proposed methods can fit within the parameterized criterion:

$$J(X_m) \triangleq I(X_m;C) - \beta \sum_{X_j \in \mathbf{S}} I(X_m;X_j) + \gamma \sum_{X_j \in \mathbf{S}} I(X_m;X_j|C). \tag{3}$$

For example, the Joint Mutual Information (JMI) criterion [6] can be obtained with $\beta = \gamma = 1/|\mathbf{S}|$. The Conditional Informative Feature Extraction (CIFE) criterion [8] is obtained with $\beta = \gamma = 1$. The extended MRMR criterion [9] is a special case when $\beta = \gamma$. The objective in (2), including MRMR and MIFS, are clearly special cases where $\gamma = 0$. These methods can detect higher order variable dependencies, in particular those between two features and the class variable. However, all the mentioned criteria were handcrafted and their theoretical underpinning is not well understood. In particular, (i) *in retrospect*, we would like to understand how these criteria are related to the original full joint MI criterion in (1), and (ii) *moving forward*, we would like to leverage this understanding to design higher-order MI based feature selection methods in a more systematic and methodological manner. Recent work has partially elucidated the former question [10,3], while to our knowledge, the latter question has not been investigated.

Contributions: To address the identified gap, in this paper, we study the connection between the low-dimensional MI based criteria, such as the ones in (2) and (3), and the ultimate high-dimensional MI objective in (1). The benefit of such an investigation is two-fold: (i) to establish the theoretical underpinnings for heuristics based on (2) and (3), and (ii) to inspire a systematic and methodological development of higher-dimensional MI-based feature selection techniques by relaxing the identified assumptions. We take a first step towards this direction by proposing several novel MI based feature selection approaches that take into account higher-order dependency between features, in particular three-way feature interaction $I(X_i;X_j|X_k)$. Our extensive experimental evaluation shows that systematic inclusion of higher-dimensional MI quantities improves the feature selection performance.

2. Assumptions underlying low-dimensional MI-based feature selection heuristics

Our first goal in this paper is to strive for a more comprehensive understanding of the theoretical underpinnings behind various MI based feature selection heuristics. Several recent works have partially addressed this question. Balagani and Proha [10] identified a set of assumptions underlying the objective (2) while Brown et al. [3] investigated the assumptions underlying the more general objective (3). In this section, we continue to develop further along these lines, while making some new connections between the previous work.

In [10], Balagani and Proha set out to identify the conditions under which the high-dimensional MI in (1) could be decomposed exactly as a sum of low-dimensional relevancy and redundancy MI terms, i.e.,

$$I(\mathbf{S};C) \equiv \sum_{i=1}^{m} I(X_i;C) - \sum_{i=2}^{m} \sum_{j<i} I(X_i;X_j) \tag{4}$$

They showed that under the following three assumptions, the identity (4) holds true.

Assumption 1. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are independent, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}) = \prod_{i=1}^{m-1} P(X_i) \tag{5}$$

Assumption 2. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are conditionally independent given the feature X_m, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}|X_m) = \prod_{i=1}^{m-1} P(X_i|X_m). \tag{6}$$

Assumption 3 (*Naive Bayes independence assumption*). Each feature independently influences the class variable, i.e.,

$$P(X_m|X_1, \ldots, X_{m-1}, C) = P(X_m|C). \tag{7}$$

We will argue here briefly that, of these three assumptions, Assumption 1 is a strong condition. More specifically, the condition in (5) implies that all features in $\mathbf{S}$ are pairwise independent, indeed

$$\begin{aligned} \forall X_i, X_j \in \mathbf{S} : P(X_i, X_j) &= \sum_{\mathbf{S}\setminus\{X_i,X_j\}} P(X_1, X_2, \ldots, X_{m-1}) \\ &= \sum_{\mathbf{S}\setminus\{X_i,X_j\}} P(X_1)P(X_2)\ldots P(X_{m-1}) = P(X_i)P(X_j) \end{aligned}$$

Furthermore, since at design time, it is not possible to anticipate which features of $\mathbf{X}$ will be selected in $\mathbf{S}$, it is necessary that all features in the original feature set $\mathbf{X}$ are also pairwise independent, for the identity (4) to hold true on any selected subset of $\mathbf{X}$. Therefore, with this assumption, we effectively have $I(X_i;X_j) = 0 \ \forall i \neq j$, implying that the incremental objective in (2) reduces to the simplistic objective of $f(X_m) = I(X_m;C)$, i.e., selecting the m-th highest ranking feature, in terms of the MI shared with C, without taking into account the redundancy with the selected features.

2.1. An alternative view

In this section, we present an alternative view on the issue of approximating high-dimensional MI with low-dimensional MI terms. First, note that even if the high-dimensional MI were easily estimable, the problem of identifying a subset $\mathbf{S}$ that shares the maximal MI with C remains a challenging combinatorial optimization problem without known efficient solution. An exhaustive search will be of $O(2^M)$ time complexity, while restricting the maximum size of $\mathbf{S}$ to $k < M$ will reduce the cost to $O(M^k)$, but will still be expensive. As such, an obvious iterative greedy strategy is to select one feature at a time: given the set $\mathbf{S} = \{X_1, \ldots, X_{m-1}\}$ of $m-1$ already selected features, the m-th feature is chosen maximizing the following objective function:

$$\arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S} \cup X_m; C) \tag{8}$$

We will now try to understand under what conditions, low-order MI based heuristics such as MRMR and MIFS in (2) will produce

Please cite this article as: N.X. Vinh, et al., Can high-order dependencies improve mutual information based feature selection? Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.11.007