Variational Information Maximization for Feature Selection

Shuyang Gao  Greg Ver Steeg  Aram Galstyan
University of Southern California, Information Sciences Institute
gaos@usc.edu, gregv@isi.edu, galstyan@isi.edu

Abstract

Feature selection is one of the most fundamental problems in machine learning. An extensive body of work on information-theoretic feature selection exists which is based on maximizing mutual information between subsets of features and class labels. Practical methods are forced to rely on approximations due to the difficulty of estimating mutual information. We demonstrate that approximations made by existing methods are based on unrealistic assumptions. We formulate a more flexible and general class of assumptions based on variational distributions and use them to tractably generate lower bounds for mutual information. These bounds define a novel information-theoretic framework for feature selection, which we prove to be optimal under tree graphical models with proper choice of variational distributions. Our experiments demonstrate that the proposed method strongly outperforms existing information-theoretic feature selection approaches.

1 Introduction

Feature selection is one of the fundamental problems in machine learning research [1, 2]. Its problematic issues include a large number of features that are either irrelevant or redundant for the task at hand. In these cases, it is often advantageous to pick a smaller subset of features to avoid over-fitting, to speed up computation, or simply to improve the interpretability of the results.

Feature selection approaches are usually categorized into three groups: wrapper, embedded and filter [3, 4, 5]. The first two methods, wrapper and embedded, are considered classifier-dependent, i.e., the selection of features somehow depends on the classifier being used. Filter methods, on the other hand, are classifier-independent and define a scoring function between features and labels in the selection process.

Because filter methods may be employed in conjunction with a wide variety of classifiers, it is important that the scoring function of these methods is as general as possible. Since mutual information (MI) is a general measure of dependence with several unique properties [6], many MI-based scoring functions have been proposed as filter methods [7, 8, 9, 10, 11, 12]; see [5] for an exhaustive list.

Owing to the difficulty of estimating mutual information in high dimensions, most existing MI-based feature selection methods are based on various low-order approximations for mutual information. While those approximations have been successful in certain applications, they are heuristic in nature and lack theoretical guarantees. In fact, as we demonstrate in Sec. 2.2, a large family of approximate methods are based on two assumptions that are mutually inconsistent.

To address the above shortcomings, in this paper we introduce a novel feature selection method based on a variational lower bound on mutual information; a similar bound was previously studied within the Infomax learning framework [13]. We show that instead of maximizing the mutual information, which is intractable in high dimensions (hence the introduction of many heuristics), we can
maximize a lower bound on the MI with the proper choice of tractable variational distributions. We use this lower bound to define an objective function and derive a forward feature selection algorithm.

We provide a rigorous proof that the forward feature selection is optimal under tree graphical models by choosing an appropriate variational distribution. This is in contrast with previous information-theoretic feature selection methods which lack any performance guarantees. We also conduct empirical validation on various datasets and demonstrate that the proposed approach outperforms state-of-the-art information-theoretic feature selection methods.

In Sec. 2 we introduce general MI-based feature selection methods and discuss their limitations. Sec. 3 introduces the variational lower bound on mutual information and proposes two specific variational distributions. In Sec. 4, we report results from our experiments, and compare the proposed approach with existing methods.

2 Information-Theoretic Feature Selection Background

2.1 Mutual Information-Based Feature Selection

Consider a supervised learning scenario where \( x = \{x_1, x_2, \ldots, x_D\} \) is a \( D \)-dimensional input feature vector, and \( y \) is the output label. In filter methods, the mutual information-based feature selection task is to select \( T \) features \( S^* = \{x_{f_1}, x_{f_2}, \ldots, x_{f_T}\} \) such that the mutual information between \( x_{S^*} \) and \( y \) is maximized. Formally,

\[
S^* = \arg \max_S I(x_S : y) \quad \text{s.t. } |S| = T
\]  

where \( I(\cdot) \) denotes the mutual information [6].

**Forward Sequential Feature Selection** Maximizing the objective function in Eq. 1 is generally NP-hard. Many MI-based feature selection methods adopt a greedy method, where features are selected incrementally, one feature at a time. Let \( S^{t-1} = \{x_{f_1}, x_{f_2}, \ldots, x_{f_{t-1}}\} \) be the selected feature set after time step \( t-1 \). According to the greedy method, the next feature \( f_t \) at step \( t \) is selected such that

\[
f_t = \arg \max_{i \notin S^{t-1}} I(x_{S^{t-1} \cup i} : y)
\]  

where \( x_{S^{t-1} \cup i} \) denotes \( x \)'s projection into the feature space \( S^{t-1} \cup i \). As shown in [5], the mutual information term in Eq. 2 can be decomposed as:

\[
I(x_{S^{t-1} \cup i} : y) = I(x_{S^{t-1}} : y) + I(x_i : y | x_{S^{t-1}})
\]

\[
= I(x_{S^{t-1}} : y) + I(x_i : y) - I(x_i : x_{S^{t-1}}) + I(x_i : x_{S^{t-1}} | y)
\]

\[
= I(x_{S^{t-1}} : y) + I(x_i : y) - (H(x_{S^{t-1}}) - H(x_{S^{t-1}} | x_i)) + (H(x_{S^{t-1}} | y) - H(x_{S^{t-1}} | x_i, y))
\]  

where \( H(\cdot) \) denotes the entropy [6]. Omitting the terms that do not depend on \( x_i \) in Eq. 3, we can rewrite Eq. 2 as follows:

\[
f_t = \arg \max_{i \notin S^{t-1}} I(x_i : y) + H(x_{S^{t-1}} | x_i) - H(x_{S^{t-1}} | x_i, y)
\]  

The greedy learning algorithm has been analyzed in [14].

2.2 Limitations of Previous MI-Based Feature Selection Methods

Estimating high-dimensional information-theoretic quantities is a difficult task. Therefore, most MI-based feature selection methods propose low-order approximation to \( H(x_{S^{t-1}} | x_i) \) and \( H(x_{S^{t-1}} | x_i, y) \) in Eq. 4. A general family of methods rely on the following approximations [5]:

\[
H(x_{S^{t-1}} | x_i) \approx \sum_{k=1}^{t-1} H(x_{f_k} | x_i)
\]

\[
H(x_{S^{t-1}} | x_i, y) \approx \sum_{k=1}^{t-1} H(x_{f_k} | x_i, y)
\]
The approximations in Eq. 5 become exact under the following two assumptions [5]:

**Assumption 1. (Feature Independence Assumption)**

\[
p(x_{S^{t-1}} | x_i) = \prod_{k=1}^{t-1} p(x_{f_k} | x_i)
\]

**Assumption 2. (Class-Conditioned Independence Assumption)**

\[
p(x_{S^{t-1}} | x_i, y) = \prod_{k=1}^{t-1} p(x_{f_k} | x_i, y)
\]

Assumption 1 and Assumption 2 mean that the selected features are independent and class-conditionally independent, respectively, given the unselected feature \(x_i\) under consideration.

![Assumption 1](image)
![Assumption 2](image)

Figure 1: The first two graphical models show the assumptions of traditional MI-based feature selection methods. The third graphical model shows a scenario when both Assumption 1 and Assumption 2 are true. Dashed line indicates there may or may not be a correlation between two variables.

We now demonstrate that the two assumptions cannot be valid simultaneously unless the data has a very specific (and unrealistic) structure. Indeed, consider the graphical models consistent with either assumption, as illustrated in Fig. 1. If Assumption 1 holds true, then \(x_i\) is the only common cause of the previously selected features \(S^{t-1} = \{x_{f_1}, x_{f_2}, \ldots, x_{f_{t-1}}\}\), so that those features become independent when conditioned on \(x_i\). On the other hand, if Assumption 2 holds, then the features depend both on \(x_i\) and class label \(y\); therefore, generally speaking, distribution over those features does not factorize by solely conditioning on \(x_i\)—there will be remnant dependencies due to \(y\). Thus, if Assumption 2 is true, then Assumption 1 cannot be true in general, unless the data is generated according to a very specific model shown in the rightmost model in Fig. 1. Note, however, that in this case, \(x_i\) becomes the most important feature because \(I(x_i : y) > I(x_{S^{t-1}} : y)\); then we should have selected \(x_i\) at the very first step, contradicting the feature selection process.

As we mentioned above, most existing methods implicitly or explicitly adopt both assumptions or their stronger versions, as shown in [5]—including mutual information maximization (MIM) [15], joint mutual information (JMI) [8], conditional mutual information maximization (CMIM) [9], maximum relevance minimum redundancy (mRMR) [10], conditional Infomax feature extraction (CIFE) [16], etc. Approaches based on global optimization of mutual information, such as quadratic programming feature selection (QPFS) [11] and the state-of-the-art conditional mutual information-based spectral method (SPEC) [12], are derived from the previous greedy methods and therefore also implicitly rely on those two assumptions.

In the next section we address these issues by introducing a novel information-theoretic framework for feature selection. Instead of estimating mutual information and making mutually inconsistent assumptions, our framework formulates a tractable variational lower bound on mutual information, which allows a more flexible and general class of assumptions via appropriate choices of variational distributions.

## 3 Method

### 3.1 Variational Mutual Information Lower Bound

Let \(p(x, y)\) be the joint distribution of input (\(x\)) and output (\(y\)) variables. Barber & Agkov [13] derived the following lower bound for mutual information \(I(x : y)\) by using the non-negativity of KL-divergence, i.e., \(\sum_x p(x|y) \log \frac{p(x|y)}{q(x|y)} \geq 0\) gives:

\[
I(x : y) \geq H(x) + (\ln q(x|y))_{p(x,y)}
\]
where angled brackets represent averages and \( q(x|y) \) is an arbitrary variational distribution. This bound becomes exact if \( q(x|y) \equiv p(x|y) \).

It is worthwhile to note that in the context of unsupervised representation learning, \( p(y|x) \) and \( q(x|y) \) can be viewed as an encoder and a decoder, respectively. In this case, \( y \) needs to be learned by maximizing the lower bound in Eq. 6 by iteratively adjusting the parameters of the encoder and decoder, such as [13, 17].

### 3.2 Variational Information Maximization for Feature Selection

Naturally, in terms of information-theoretic feature selection, we could also try to optimize the variational lower bound in Eq. 6 by choosing a subset of features \( S^* \) in \( x \), such that,

\[
S^* = \arg \max_S \{ H(x_S) + \langle \ln q(x_S|y) \rangle_{p(x_S,y)} \}
\]

(7)

However, the \( H(x_S) \) term in RHS of Eq. 7 is still intractable when \( x_S \) is very high-dimensional.

Nonetheless, by noticing that variable \( y \) is the class label, which is usually discrete, and hence \( H(y) \) is fixed and tractable, by symmetry we switch \( x \) and \( y \) in Eq. 6 and rewrite the lower bound as follows:

\[
I(x:y) \geq H(y) + \langle \ln q(y|x) \rangle_{p(x,y)} = \left\langle \ln \left( \frac{q(y|x)}{p(y)} \right) \right\rangle_{p(x,y)}
\]

(8)

The equality in Eq. 8 is obtained by noticing that \( H(y) = \langle -\ln p(y) \rangle_{p(y)} \).

By using Eq. 8, the lower bound optimal subset \( S^* \) of \( x \) becomes:

\[
S^* = \arg \max_S \left\langle \ln \left( \frac{q(y|x_S)}{p(y)} \right) \right\rangle_{p(x_S,y)}
\]

(9)

#### 3.2.1 Choice of Variational Distribution

\( q(y|x_S) \) in Eq. 9 can be any distribution as long as it is normalized. We need to choose \( q(y|x_S) \) to be as general as possible while still keeping the term \( \langle \ln q(y|x_S) \rangle_{p(x_S,y)} \) tractable in Eq. 9.

As a result, we set \( q(y|x_S) \) as

\[
q(y|x_S) = \frac{q(x_S,y)}{q(x_S)} = \frac{q(x_S|y)p(y)}{\sum_{y'} q(x_S|y')p(y')}
\]

(10)

We can verify that Eq. 10 is normalized even if \( q(x_S|y) \) is not normalized.

If we further denote,

\[
q(x_S) = \sum_{y'} q(x_S|y')p(y')
\]

(11)

then by combining Eqs. 9 and 10, we get,

\[
I(x_S:y) \geq \left\langle \ln \left( \frac{q(x_S|y)}{q(x_S)} \right) \right\rangle_{p(x_S,y)} \equiv I_{LB}(x_S:y)
\]

(12)

And we also have the following equation which shows the gap between \( I(x_S:y) \) and \( I_{LB}(x_S:y) \),

\[
I(x_S:y) - I_{LB}(x_S:y) = \langle KL(p(y|x_S) \| q(y|x_S)) \rangle_{p(x_S)}
\]

(13)

**Auto-Regressive Decomposition.** Now that \( q(y|x_S) \) is defined, all we need to do is model \( q(x_S|y) \) under Eq. 10, and \( q(x_S) \) is easy to compute based on \( q(x_S|y) \). Here we decompose \( q(x_S|y) \) as an auto-regressive distribution assuming \( T \) features in \( S \):

\[
q(x_S|y) = q(x_{f_1}|y) \prod_{t=2}^{T} q(x_{f_t}|x_{f_{<t}}, y)
\]

(14)
where \( x_{f_c} \) denotes \( \{ x_{f_1}, x_{f_2}, \ldots, x_{f_c}, \ldots \} \). The graphical model in Fig. 2 demonstrates this decomposition. The main advantage of this model is that it is well-suited for the forward feature selection procedure where one feature is selected at a time (which we will explain in Sec. 3.2.3). And if \( q(x_{f_c} | x_{f_{<c}}, y) \) is tractable, then so is the whole distribution \( q(x_S | y) \). Therefore, we would find tractable \( Q \)-distributions over \( q(x_{f_c} | x_{f_{<c}}, y) \). Below we illustrate two such \( Q \)-distributions.

**Naive Bayes Q-distribution.** A natural idea would be to assume \( x_i \) is independent of other variables given \( y \), i.e.,

\[
q(x_{f_c} | x_{f_{<c}}, y) = p(x_{f_c} | y)
\]  

(15)

Then the variational distribution \( q(y | x_S) \) can be written based on Eqs. 10 and 15 as follows:

\[
q(y | x_S) = \frac{p(y) \prod_{j \in S} p(x_j | y)}{\sum_{y'} p(y') \prod_{j \in S} p(x_j | y')}
\]

(16)

And we also have the following theorem:

**Theorem 3.1 (Exact Naive Bayes).** Under Eq. 16, the lower bound in Eq. 8 becomes exact if and only if data is generated by a Naive Bayes model, i.e., \( p(x, y) = p(y) \prod_i p(x_i | y) \).

The proof for Theorem 3.1 becomes obvious by using the mutual information definition. Note that the most-cited MI-based feature selection method mRMR [10] also assumes conditional independence given the class label \( y \) as shown in [5, 18, 19], but they make additional stronger independence assumptions among only feature variables.

**Pairwise Q-distribution.** We now consider an alternative approach that is more general than the Naive Bayes distribution:

\[
q(x_{f_c} | x_{f_{<c}}, y) = \left( \prod_{i=1}^{t-1} p(x_{f_i} | x_{f_{<i}}, y) \right)^{1/t}
\]

(17)

In Eq. 17, we assume \( q(x_{f_c} | x_{f_{<c}}, y) \) to be the geometric mean of conditional distributions \( q(x_{f_c} | x_{f_{<c}}, y) \). This assumption is tractable as well as reasonable because if the data is generated by a Naive Bayes model, the lower bound in Eq. 8 also becomes exact using Eq. 17 due to \( p(x_{f_c} | x_{f_c}, y) \equiv p(x_{f_c} | y) \) in that case.

### 3.2.2 Estimating Lower Bound From Data

Assuming either Naive Bayes Q-distribution or pairwise Q-distribution, it is convenient to estimate \( q(x_S | y) \) and \( q(x_S) \) in Eq. 12 by using plug-in probability estimators for discrete data or one/two-dimensional density estimators for continuous data. We also use the sample mean to approximate the expectation term in Eq. 12. Our final estimator for \( I_{LB} (x_S : y) \) is written as follows:

\[
\hat{I}_{LB} (x_S : y) = \frac{1}{N} \sum_{x^{(k)}, y^{(k)}} \ln \frac{\hat{q}(x_S^{(k)} | y^{(k)})}{\hat{q}(x_S^{(k)})}
\]

(18)

where \( \{ x^{(k)}, y^{(k)} \} \) are samples from data, and \( \hat{q}(\cdot) \) denotes the estimate for \( q(\cdot) \).
3.2.3 Variational Forward Feature Selection Under Auto-Regressive Decomposition

After defining $q(y|x_S)$ in Eq. 10 and auto-regressive decomposition of $q(x_S|y)$ in Eq. 15, we are able to do the forward feature selection previously described in Eq. 2, but replace the mutual information with its lower bound $\hat{I}_{LB}$. Recall that $S^{t-1}$ is the set of selected features after step $t-1$, then the feature $f_t$ will be selected at step $t$ such that

$$f_t = \arg \max_{i \in S^{t-1}} \hat{I}_{LB} (x_{S^{t-1} \cup i} : y)$$

where $\hat{I}_{LB} (x_{S^{t-1} \cup i} : y)$ can be obtained from $\hat{I}_{LB} (x_{S^{t-1}} : y)$ recursively by auto-regressive decomposition $q(x_{S^{t-1}}|y) = q(x_{S^{t-1}|y}) q(x_i|x_{S^{t-1}}, y)$ where $q(x_{S^{t-1}}|y)$ is stored at step $t-1$. This forward feature selection can be done under auto-regressive decomposition in Eqs. 10 and 14 for any $Q$-distribution. However, calculating $q(x_i|x_{S^t}, y)$ may vary according to different $Q$-distributions. We can verify that it is easy to get $q(x_i|x_{S^t}, y)$ recursively from $q(x_i|x_{S^{t-1}}, y)$ under Naive Bayes or pairwise $Q$-distribution. We call our algorithm under these two $Q$-distributions $VMI_{naive}$ and $VMI_{pairwise}$ respectively.

It is worthwhile noting that the lower bound does not always increase at each step. A decrease in lower bound at step $t$ indicates that the $Q$-distribution would approximate the underlying distribution worse than it did at previous step $t-1$. In this case, the algorithm would re-maximize the lower bound from zero with only the remaining unselected features. We summarize the concrete implementation of our algorithms in supplementary Sec. A.

Time Complexity. Although our algorithm needs to calculate the distributions at each step, we only need to calculate the probability value at each sample point. For both $VMI_{naive}$ and $VMI_{pairwise}$, the total computational complexity is $O(NDT)$ assuming $N$ as number of samples, $D$ as total number of features, $T$ as number of final selected features. The detailed time analysis is left for the supplementary Sec. A. As shown in Table 1, our methods $VMI_{naive}$ and $VMI_{pairwise}$ have the same time complexity as mRMR [10], while the state-of-the-art global optimization method SPEC [12] is required to precompute the pairwise mutual information matrix, which gives a time complexity of $O(ND^2)$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRMR</td>
<td>$O(NDT)$</td>
</tr>
<tr>
<td>$VMI_{naive}$</td>
<td>$O(NDT)$</td>
</tr>
<tr>
<td>$VMI_{pairwise}$</td>
<td>$O(NDT)$</td>
</tr>
<tr>
<td>SPEC</td>
<td>$O(ND^2)$</td>
</tr>
</tbody>
</table>

Optimality Under Tree Graphical Models. Although our method $VMI_{naive}$ assumes a Naive Bayes model, we can prove that this method is still optimal if the data is generated according to tree graphical models. Indeed, both of our methods, $VMI_{naive}$ and $VMI_{pairwise}$, will always prioritize the first layer features, as shown in Fig. 3. This optimality is summarized in Theorem B.1 in supplementary Sec. B.

4 Experiments

Synthetic Data. We begin with the experiments on a synthetic model according to the tree structure illustrated in the left part of Fig. 3. The detailed data generating process is shown in supplementary section D. The root node $Y$ is a binary variable, while other variables are continuous. We use $VMI_{naive}$ to optimize the lower bound $I_{LB}(x : y)$. 5000 samples are used to generate the synthetic data, and variational $Q$-distributions are estimated by the kernel density estimator. We can see from the plot in the right-hand part of Fig. 3 that our algorithm, $VMI_{naive}$, selects $x_1$, $x_2$, $x_3$ as the first three features, although $x_2$ and $x_3$ are only weakly correlated with $y$. If we continue to add deeper level features $\{x_4, ..., x_6\}$, the lower bound will decrease. For comparison, we also illustrate the mutual information between each single feature $x_i$ and $y$ in Table 2. We can see from Table 2 that it would choose $x_1$, $x_4$ and $x_5$ as the top three features by using the maximum relevance criteria [15].
Figure 3: (Left) This is the generative model used for synthetic experiments. Edge thickness represents the relationship strength. (Right) Optimizing the lower bound by $VMI_{naive}$. Variables under the blue line denote the features selected at each step. Dotted blue line shows the decreasing lower bound if adding more features. Ground-truth mutual information is obtained using $N = 100,000$ samples.

Table 2: Mutual information between label $y$ and each feature $x_i$ for Fig. 3. $I(x_i : y)$ is estimated using $N=100,000$ samples. Top three variables with highest mutual information are highlighted in bold.

<table>
<thead>
<tr>
<th>feature</th>
<th>$I(x_i : y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.111</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.052</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.022</td>
</tr>
<tr>
<td>$x_4$</td>
<td><strong>0.058</strong></td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.025</td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.029</td>
</tr>
<tr>
<td>$x_7$</td>
<td>0.012</td>
</tr>
<tr>
<td>$x_8$</td>
<td>0.013</td>
</tr>
<tr>
<td>$x_9$</td>
<td></td>
</tr>
</tbody>
</table>

Real-World Data. We compare our algorithms $VMI_{naive}$ and $VMI_{pairwise}$ with other popular information-theoretic feature selection methods, including mRMR [10], JMI [8], MIM [15], CMIM [9], CIFE [16], and $SPEC_{CMI}$ [12]. We use 17 well-known datasets in previous feature selection studies [5, 12] (all data are discretized). The dataset summaries are illustrated in supplementary Sec. C. We use the average cross-validation error rate on the range of 10 to 100 features to compare different algorithms under the same setting as [12]. Tenfold cross-validation is employed for datasets with number of samples $N \geq 100$ and leave-one-out cross-validation otherwise. The 3-nearest-neighbor classifier is used for Gisette and Madelon, following [5]. For the remaining datasets, the chosen classifier is Linear SVM, following [11, 12].

The experimental results can be seen in Table 3. The entries with * and ** indicate the best performance and the second best performance, respectively (in terms of average error rate). We also use the paired t-test at 5% significant level to test the hypothesis that $VMI_{naive}$ or $VMI_{pairwise}$ perform significantly better than other methods, or vice versa. Overall, we find that both of our methods, $VMI_{naive}$ and $VMI_{pairwise}$, strongly outperform other methods. This indicates that our variational feature selection framework is a promising addition to the current literature of information-theoretic feature selection.

Figure 4: Number of selected features versus average cross-validation error in datasets Semeion and Gisette.

---

1We omit the results for MIM and CIFE due to space limitations. The complete results are shown in the supplementary Sec. C.
and developing more powerful variational approaches for capturing complex dependencies. 

for information-theoretic feature selection, such as optimizing the variational lower bound globally. 

We hope that our framework will inspire new mathematically rigorous algorithms 

VMI manner by progressively maximizing the lower bound. We also presented two concrete methods 

issues and showed that by auto-regressive decomposition, feature selection can be done in a forward 

distributions. We introduced a unifying variational mutual information lower bound to address these 

Feature selection has been a significant endeavor over the past decade. Mutual information gives 

the dimensionality is high. The proposed variational lower bound can be viewed as a way of esti-

more scalable feature selection methods have been developed, but these methods also require the 

combinations of so-called 

twenty years: [5, 7, 8, 9, 10, 15, 11, 12, 20], to name a few. Most of these methods are based on 

5 Related Work 

shows the two most distinguishable data sets, Semeion and Gisette. We can see that both of our 

We also plot the average cross-validation error with respect to number of selected features. Fig. 4 

only a few examples exist in literature [22] under this setting. We hope our method will shed light 

the calculation of high-dimensional mutual information as a basic scoring function. 

5 Related Work 

There has been a significant amount of work on information-theoretic feature selection in the past 

twenty years: [5, 7, 8, 9, 10, 15, 11, 12, 20], to name a few. Most of these methods are based on 

combinations of so-called relevant, redundant and complimentary information. Such combinations 

representing low-order approximations of mutual information are derived from two assumptions, 

and it has proved unrealistic to expect both assumptions to be true. Inspired by group testing [21], 

more scalable feature selection methods have been developed, but these methods also require the 

calculation of high-dimensional mutual information as a basic scoring function. 

Estimating mutual information from data requires a large number of observations—especially when 

the dimensionality is high. The proposed variational lower bound can be viewed as a way of esti-

mating mutual information between a high-dimensional continuous variable and a discrete variable. 

Only a few examples exist in literature [22] under this setting. We hope our method will shed light 

on new ways to estimate mutual information, similar to estimating divergences in [23]. 

6 Conclusion 

Feature selection has been a significant endeavor over the past decade. Mutual information gives 

general basis for quantifying the informativeness of features. Despite the clarity of mutual in-

formation, estimating it can be difficult. While a large number of information-theoretic methods 

exist, they are rather limited and rely on mutually inconsistent assumptions about underlying data 

distributions. We introduced a unifying variational mutual information lower bound to address these 

issues and showed that by auto-regressive decomposition, feature selection can be done in a forward 

manner by progressively maximizing the lower bound. We also presented two concrete methods 

using Naive Bayes and pairwise Q-distributions, which strongly outperform the existing methods. 

VMI naive only assumes a Naive Bayes model, but even this simple model outperforms the existing 

information-theoretic methods, indicating the effectiveness of our variational information maxima-

tion framework. We hope that our framework will inspire new mathematically rigorous algorithms 

for information-theoretic feature selection, such as optimizing the variational lower bound globally 

and developing more powerful variational approaches for capturing complex dependencies.
References


