Conditional Likelihood Maximization: A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown, Adam Pocock, Mingjie Zhao and Mikel Lujan
School of Computer Science
University of Manchester

Presented by Wenzhao Lian
July 27, 2012
Main Contribution

Feature selection problem: selecting the feature set which is most relevant and least redundant. What’s the criterion of selection? Existing criteria provide scoring functions to measure relevancy and redundancy of features.
Main Contribution

Feature selection problem: selecting the feature set which is most relevant and least redundant.
What’s the criterion of selection?
Existing criteria provide scoring functions to measure relevancy and redundancy of features.
In this paper:
- Deriving a scoring function, instead of defining.
- Proposing a unifying framework for information theoretic feature selection.
- This general criterion can be naturally extended to existing criteria under different assumptions.
Entrophy and Mutual Information

\[ H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \]

\[ H(X|Y) = - \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y) \]

\[ I(X; Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy) \log \frac{p(xy)}{p(x)p(y)} \]

\[ I(X; Y|Z) = H(X|Z) - H(X|YZ) \]

\[ = \sum_{z \in \mathcal{Z}} p(z) \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy|z) \log \frac{p(xy|z)}{p(x|z)p(y|z)} \]
Previous Feature Selection Criteria

- Mutual Information Maximization (MIM)
  \[ J_{mim}(X_k) = I(X_k; Y) \]  
  \( J_{mim} \): relevance index. \( X_k \): \( k \)-th feature. \( Y \): class label.

- Mutual Information Feature Selection (MIFS)
  \[ J_{mifs}(X_k) = I(X_k; Y) - \beta \sum_{X_j \in S} I(X_k; X_j) \]  
  \( J_{mifs} \): relevance index. \( S \): set of currently selected features. \( \beta \) controlling redundancy penalty.

- Joint Mutual Information (JMI)
  \[ J_{jmi}(X_k) = \sum_{X_j \in S} I(X_kX_j; Y) \]  
  Indicating that the candidate feature which is complementary with existing features should be included.
Conditional Likelihood Problem

\[ \mathcal{D} = \{x^i, y^i; i = 1..N\} \]
\[ x^i = [x^i_1, x^i_2, \ldots, x^i_d]^T \]
\[ x = \{x_\theta, x_{\tilde{\theta}}\} \]
\[ \tau: \text{parameters used to predict } y \]

Conditional log likelihood of the labels given parameters \( \theta, \tau \) is

\[ \ell = \frac{1}{N} \sum_{i=1}^{N} \log q(y^i|x^i_\theta, \tau) \] (5)
Conditional Likelihood Problem

Introduce $p(y|x_\theta)$ and $p(y|x)$: the true distribution of the class labels given the selected features $x_\theta$ and of the class labels given all features.

$$\ell = \frac{1}{N} \sum_{i=1}^{N} \log \frac{q(y^i|x^i_\theta, \tau)}{p(y^i|x^i_\theta)} + \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(y^i|x^i_\theta)}{p(y^i|x^i)} + \frac{1}{N} \sum_{i=1}^{N} \log p(y^i|x^i)$$ (6)
Conditional Likelihood Problem

Introduce \( p(y|x_\theta) \) and \( p(y|x) \): the true distribution of the class labels given the selected features \( x_\theta \) and of the class labels given all features.

\[
\ell = \frac{1}{N} \sum_{i=1}^{N} \log \frac{q(y^i|x^i_\theta, \tau)}{p(y^i|x^i_\theta)} + \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(y^i|x^i_\theta)}{p(y^i|x^i)} + \frac{1}{N} \sum_{i=1}^{N} \log p(y^i|x^i) \tag{6}
\]

Taking the limit, the objective function becomes minimizing

\[
-\ell = E_{x,y} \{ \log \frac{p(y|x_\theta)}{q(y|x_\theta, \tau)} \} + I(X_\theta; Y|X_\theta) + H(Y|X) \tag{7}
\]

The first term depends on the model.

The final term gives a lower bound on the Bayes error.

Based on the Filter assumption, which means optimizing the feature set and optimizing the classifier are two independent stages, we can minimize the second term not caring about the first term.
For the second term, we have

\[ I(X_{\tilde{\theta}}; Y|X_\theta) = I(X; Y) - I(X_{\theta}; Y) \] (8)

Thus, minimizing \( I(X_{\tilde{\theta}}; Y|X_\theta) \) equals to maximizing \( I(X_\theta; Y) \).

Using the greedy approach
First, initialize the selected set as a null set.
Then, at each step the feature that has the highest score is selected.
Repeat the second step until a stopping criterion is reached.

\( S \) is the currently selected set, and the score for a feature \( X_k \) is

\[ J_{cmi}(X_k) = I(X_k; Y|S) \] (9)
Unifying criterion

To bring score functions proposed in previous work into this framework, three assumptions are needed.

**Assumption 1** For all unselected features $X_k \in X_{\tilde{\theta}}$, assume

$$p(x_{\theta}|x_k) = \prod_{j \in S} p(x_j|x_k)$$

$$p(x_{\theta}|x_ky) = \prod_{j \in S} p(x_j|x_ky)$$

(10)

Under Assumption 1, an equivalent criterion can be written as

$$J'_{cmi}(X_k) = I(X_k; Y) - \sum_{j \in S} I(X_j; X_k) + \sum_{j \in S} I(X_j; X_k|Y)$$

(11)
Unifying criterion

**Assumption 2** For all features, assume

\[ p(x_i x_j | y) = p(x_i | y) p(x_j | y) \]  

(12)

**Assumption 3** For all features, assume

\[ p(x_i x_j) = p(x_i) p(x_j) \]  

(13)

Depending on how strong the belief in Assumption 2 and 3 is, different criteria are obtained.

\[ J_{mim}(X_k) = I(X_k; Y) \]

\[ J_{mifs}(X_k) = I(X_k; Y) - \beta \sum_{X_j \in S} I(X_k; X_j) \]

\[ J_{mrmr}(X_k) = I(X_k; Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_k; X_j) \]  

(14)

\[ J_{jmi}(X_k) = I(X_k; Y) - \frac{1}{|S|} \sum_{X_j \in S} [I(X_k; X_j) - I(X_k; X_j | Y)] \]
Unifying criterion

A general form of the unifying criterion:

\[ J'_{cmi}(X_k) = I(X_k; Y) - \beta \sum_{j \in S} I(X_j; X_k) + \gamma \sum_{j \in S} I(X_j; X_k|Y) \]  \hspace{1cm} (15)
1. Main Contribution
2. Background
3. Main work
4. Experiments
5. Conclusion
Criteria

Criteria:
- Stability or Consistency
- Similarity between different methods
- Performance in limited and extreme small-sample situations.
- Stability and Accuracy Tradeoff

Classifier: A nearest neighbour classifier (k=3) is used.
Criteria:

- Stability or Consistency
- Similarity between different methods
- Performance in limited and extreme small-sample situations.
- Stability and Accuracy Tradeoff

Classifier: A nearest neighbour classifier (k=3) is used.
Figure 3: Kuncheva's Stability Index across 15 data sets. The box indicates the upper/lower quartiles, the horizontal line within each shows the median value, while the dotted crossbars indicate the maximum/minimum values. For convenience of interpretation, criteria on the x-axis are ordered by their median value.

Figure 4: Yu et al's Information Stability Index across 15 data sets. For comparison, criteria on the x-axis are ordered identically to Figure 3. The general picture emerges similarly, though the information stability index is able to take feature redundancy into account, showing that some criteria are slightly more stable than expected.
5.2 How Similar are the Criteria?

Two criteria can be directly compared with the same methodology: by measuring the consistency and information consistency between selected feature subsets on a common set of data. We calculate the mean consistencies between two feature sets of size 10, repeatedly selected over 50 bootstraps from the original data. This is then arranged in a similarity matrix, and we use classical multi-dimensional scaling to visualise this as a 2-d map, shown in Figures 5a and 5b. Note again that while the indices may return different absolute values (one is a normalized mean of a hypergeometric distribution and the other is a pairwise sum of mutual information terms) they show very similar relative 'distances' between criteria.

Both diagrams show a cluster of several criteria, and 4 clear outliers: MIFS, CIFE, ICAP and CondRed. The 5 criteria clustering in the upper left of the space appear to return relatively similar feature sets. The 4 outliers appear to return quite significantly different feature sets, both from the clustered set, and from each other. A common characteristic of these 4 outliers is that they do not scale the redundancy or conditional redundancy information terms. In these criteria, the upper bound on the redundancy term $\sum_{j \in S} I(X_k; X_j)$ grows linearly with the number of selected features, whilst the upper bound on the relevancy term $I(X_k; Y)$ remains constant. When this happens the relevancy term is overwhelmed by the redundancy term and thus the criterion selects features with minimal redundancy, rather than trading off between the two terms. This leads to strongly divergent feature sets being selected, which is reflected in the stability of the criteria. Each of the outliers are different from each other as they have different combinations of redundancy and conditional redundancy. We will see this 'balance' between relevancy and redundancy emerge as a common theme in the experiments over the next few sections.

Figure: Stability Comparison
Average ranks of criteria in terms of test error, selecting 10 features, across 11 data sets. Note the clear dominance of criteria which do not allow the redundancy term to overwhelm the relevancy term (unfilled markers) over those that allow redundancy to grow with the size of the feature set (filled markers).

**Figure:** Limited and Extreme Small-sample
Stability Accuracy Tradeoff

Figure: Stability Accuracy Tradeoff
Present a unifying framework for information theoretic feature selection via optimization of the conditional likelihood.

Clarify the implicit assumptions made when using different feature selection criteria.

Conduct empirical study on 9 heuristic mutual information criteria across data sets to analyze their properties.