Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy

Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy

H Peng, F Long, C Ding - IEEE Transactions on pattern ..., 2005 - ieeexplore.ieee.org

Feature selection is an important problem for pattern classification systems. We study how to select good features according to the maximal statistical dependency criterion based on mutual information. Because of the difficulty in directly implementing the maximal dependency condition, we first derive an equivalent form, called minimal-redundancy-maximal-relevance criterion (mRMR), for first-order incremental feature selection. Then, we present a two-stage feature selection algorithm by combining mRMR and other more ...

☆ 99 被引用次数：5430 相关文章 所有 14 个版本

- **First, derive an minimal-redundancy-maximal-relevance criterion**
- **Then, present a two-stage feature selection algorithm by combining mRMR and other more sophisticated feature selectors (e.g., wrappers)**

Given the input data D tabled as N samples and M features $X = \{x_i, i = 1, \ldots, M\}$, and the target classification variable c, the feature selection problem is to find from the M-dimensional observation space, R^M, a subspace of m features, R^m, that "optimally " characterizes c.

Minimal classification error usually requires the maximal statistical dependency of the target class c on the data distribution in the subspace R^m (and vice versa). This scheme is *maximal dependency*.

One of the most popular approaches to realize *maximal dependency* is *maximal relevance* feature selection: selecting the features with the highest relevance to the target class c. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$:

$$I(x, y) = \iint p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy \qquad (1)$$

In *maximal relevance*, the selected features x_i are required, individually, to have the largest mutual information $I(x_i; c)$ with the target class c, reflecting the largest dependency on the target class. However, it has been recognized that the combinations of individually good features do not necessarily lead to good classification performance. In other words, **"the m best features are not the best m features"**.

Relationships on Max-Dependency, Max-Relevance, and Min-Redundancy

Max-Dependency

Max-Dependency has the following form:

$$maxD(S,c), D = I(\{x_i, i = 1, \ldots, m\}; c) \tag{2}$$

$$\begin{aligned} I(S_m; c) &= \iint p(S_m, c) \log \frac{p(S_m, c)}{p(S_m)p(c)} dS_m dc \\ &= \iint p(S_{m-1}, x_m, c) \log \frac{p(S_{m-1}, x_m, c)}{p(S_{m-1}, x_m)p(c)} dS_{m-1} dx_m dc \\ &= \int \cdots \int p(x_1, \cdots, x_m, c) \log \frac{p(x_1, \cdots, x_m, c)}{p(x_1, \cdots, x_m)p(c)} dx_1 \cdots dx_m dc \end{aligned} \tag{3}$$

Max-Relevance and Min-Redundancy

As Max-Dependency criterion is hard to implement, an alternative is to select features based on *maximal relevance* criterion. Max-Relevance is to search features satisfying (4), which approximates $D(S, c)$ in (2) with the mean value of all mutual information values between individual features x_i and class c.

$$\max D(S, c),\ D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{4}$$

The following *minimal redundancy* condition can be added to select mutually exclusive features:

$$\min R(S),\ R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{5}$$

The criterion combing the above two constraints is called "minimal-redundancy-maximal-relevance" (mRMR). We define operator $\Phi(D, R)$ to combine D and R and consider the following simplest form to optimize D and R simultaneously:

$$\max \Phi(D, R),\ \Phi = D - R \tag{6}$$

In practice, incremental search methods can be used to find the near-optimal features defined by $\Phi(\cdot)$. Suppose we already have S_{m-1}, the feature set with $m - 1$ features. The task is to select the mth feature from the set $\{X - S_{m-1}\}$. This is done by selecting the feature that maximizes $\Phi(\cdot)$. The respective incremental algorithm optimizes the following condition:

$$\max_{x_j \in X - S_{m-1}} [I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i)] \tag{7}$$

The computational complexity of this incremental search method is $O(|S| \cdot M)$.

Optimal First-Order Incremental Selection

We have the following theorem:

Theorem. For the first-order incremental search, mRMR is equivalent to Max-Dependency (2).

Proof. By definition of the first-order search, we assume that S_{m-1}, i.e., the set of $m-1$ features, has already been obtained. The task is to select the optimal mth feature x_m from set $\{X - S_{m-1}\}$.

The dependency D in (2) and (3) is represented by mutual information, i.e., $D = I(S_m; c)$, where $S_m = \{S_{m-1}, x_m\}$ can be treated as a multivariate variable. Thus, by the definition of mutual information, we have:

$$\begin{aligned} I(S_m; c) &= H(c) + H(S_m) - H(S_m, c) \\ &= H(c) + H(S_{m-1}, x_m) - H(S_{m-1}, x_m, c) \end{aligned} \tag{8}$$

where $H(\cdot)$ is the entropy of the respective multivariate (or univariate) variables.

Now, we define the following quantity $J(S_m) = J(x_1, \cdots, x_m)$ for scalar variables $x_1, \cdots, x_m$,

$$\begin{aligned} &J(x_1, x_2, \cdots, x_m) = \\ &\int \cdots \int p(x_1, \cdots, x_m) \log \frac{p(x_1, x_2, \cdots, x_m)}{p(x_1) \cdots p(x_m)} dx_1 \cdots dx_m \end{aligned} \tag{9}$$

Similarly, we define $J(S_m, c) = J(x_1, \cdots, x_m, c)$ as

$$\begin{aligned} &J(x_1, x_2, \cdots, x_m, c) = \\ &\int \cdots \int p(x_1, \cdots, x_m, c) \log \frac{p(x_1, x_2, \cdots, x_m, c)}{p(x_1) \cdots p(x_m) p(c)} dx_1 \cdots dx_m dc \end{aligned} \tag{10}$$

We can easily derive (11) and (12) from (9) and (10),

$$H(S_{m-1}, x_m) = H(S_m) = \sum_{i=1}^{m} H(x_i) - J(S_m) \tag{11}$$

$$H(S_{m-1}, x_m, c) = H(S_m, c) = H(c) + \sum_{i=1}^{m} H(x_i) - J(S_m, c) \tag{12}$$

By substituting them to the corresponding terms in (8), we have

$$\begin{aligned} I(S_m; c) &= J(S_m, c) - J(S_m) \\ &= J(S_{m-1}, x_m, c) - J(S_{m-1}, x_m) \end{aligned} \tag{13}$$

Obviously, Max-Dependency is equivalent to simultaneously maximizing the first term and minimizing the second term.

We can use the Jensen's Inequality to show the second term $J(S_{m-1}, x_m)$ is lower-bounded by 0.

Consider the inequality $\log(z) \leq z - 1$ with the equality if and only if $z = 1$. We see that

$$
\begin{aligned}
& -J(x_1, x_2, \cdots, x_m) \\
&= \int \cdots \int p(x_1, \cdots, x_m) \log \frac{p(x_1) \cdots p(x_m)}{p(x_1, \cdots, x_m)} dx_1 \cdots dx_m \\
&\leq \int \cdots \int p(x_1, \cdots, x_m) [\frac{p(x_1) \cdots p(x_m)}{p(x_1, \cdots, x_m)} - 1] dx_1 \cdots dx_m \qquad (10) \\
&= \int \cdots \int p(x_1) \cdots p(x_m) dx_1 \cdots dx_m - \int \cdots \int p(x_1, \cdots, x_m) dx_1 \cdots dx_m \\
&= 1 - 1 = 0
\end{aligned}
$$

It is easy to verify that the minimum is attained when $p(x_1, \cdots, x_m) = p(x_1) \cdots p(x_m)$, i.e., all the variables are independent of each other. As all the $m-1$ features have been selected this pair-wise independence condition means that the mutual information between x_m and any selected feature $x_i (i = 1, \cdots, m-1)$ is minimized. This is the **Min-Redundancy** criterion.

We can also derive the upper bound of the first term in (13), $J(S_{m-1}, c, x_m)$.

It is easy to verify the maximum of $J(y_1, \cdots, y_n)$ or, similarly, the first term in (13), $J(S_{m-1}, c, x_m)$, is attained when all variables are maximally dependent. When S_{m-1} has been fixed, this indicates that x_m and c should have the maximal dependency. This is the **Max-Relevance criterion**.

Therefore, according to (13), as a combination of Max-Relevance and Min-Redundancy, mRMR is equivalent to Max-Dependency for first-order selection.

> Note that the quantity $J(\cdot)$ in (9) and (10) has also been called "mutual information" for multiple scalar variables.

Feature Selection Algorithms

Selecting the Candidate Feature Set

1. Use mRMR incremental selection (7) to select n (a preset large number) sequential features from the input X. This leads to n sequential feature sets $S_1 \subset S_2 \subset \cdots \subset S_{n-1} \subset S_n$.
2. Compare all the n sequential feature sets $S_1, \cdots, S_k, \cdots, S_n, (1 \leq k \leq n)$ to find the range of k, called Ω, within which the respective (cross-validation classification) error e_k is consistently small (i.e., has both small mean and small variance).
3. Within Ω, find the smallest classification error $e^* = \min e_k$. The optimal size of the candidate feature set, n^*, is chosen as the smallest k that corresponds to e^*.

Selecting Compact Feature Subsets

1. The backward selection tries to exclude one redundant feature at a time from the current feature set S_k (initially, k is set to n^* obtained in former subsection), with the constraint that the resultant feature set S_{k-1} leads to a classification error e_{k-1} no worse than e_k.
2. The forward selection tries to select a subset of m features from S_{n^*} in an incremental manner. Initially, the classification error is set to the number of samples, i.e., N. The wrapper first searches for the feature subset with one feature, denoted as Z_1, by selecting the feature x_1^* that leads to the largest error reduction. Then, from the set $\{S_n - Z_1\}$, the wrapper selects the feature x_2^* so that the feature set $Z_2 = \{Z_1, x_2^*\}$ leads to the largest

error reduction. This incremental selection repeats until the classification error begins to increase, i.e., $e_{k+1} > e_k$.

提供计算氨基酸理化性质的服务器：http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/PROTEIN/PC-PseAAC/ 之前提供生物信息基础知识各种的网站：http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/doc/#PROTEIN_Amino利用二项分布选择特征的参考论文：https://www.academia.edu/3675136/A_new_feature_selection_algorithm_based_on_binomial_hypothesis_testing_for_spam_filtering 和 林老师的论文 Sequence-based predictive modeling to identify cancerlectins