[Feature selection considering the composition of feature relevancy]

[Zahra Fathollahi] [Second Homework of Information Theory]

[Highlights]

1. A novel feature selection method based on information theory is proposed.

2. The composition of feature relevancy is taken into account.

3. Our method maximizes new information while minimizing redundancy information.

4. Our method outperforms five competing methods in terms of average accuracy.

5. The proposed method achieves the highest accuracy.

[Preliminaries]

Consider three discrete random variables X, Y and Z. The mutual information between X and Y is defined as:

$$I(X;Y) = H(Y)-H(Y|X) \quad (1)$$

Where H(Y) and H(Y|X) are entropy and conditional entropy. Entropy describes the uncertainty of a random variable, and conditional entropy describes the amount of uncertainty left when another variable is introduced. Therefore, the mutual information represents the uncertainty reduced when another variable is introduced; it also indicates the amount of information that both variables share. Another important concept of information theory is conditional mutual information, which is expressed as follows:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z)-H(X,Y|Z) \quad (2)$$

Conditional mutual information represents the amount of information between X and Y when Z is introduced. Similarly, joint mutual information between (X,Y) and Z is defined as:

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \quad (3)$$

Joint mutual information indicates the mutual information between (X,Y) and Z. Interaction information is used to measure feature redundancy in the process of feature selection. It is defined as:

$$I(X;Y;Z) = I(X;Z) + I(X;Y)-I(X;Y,Z) \quad (4)$$

Interaction information can be positive, zero or negative. It is positive when the variables Y and Z can provide the same information; it is negative when Y and Z provide more information than they provide individually; it is zero when Y and Z are independent in the context of X. Eq. (4) can be rewritten as follows (Wang et al., 2017a):

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \qquad (5)$$

In other sources, interaction information is defined using the opposite sign in Eq. (5); however, this is for mathematical convenience and does not change the final conclusion.

[The proposed Composition of Feature Relevancy method]

one group of feature selection methods aims to maximize new classification information, i.e., $I(X_k;Y|X_j)$. According to Eq. (5),

$$I(X_k;Y|X_j) = I(X_k;Y) - I(X_k;Y;X_j) \qquad (16)$$

The first item on the right side of Eq. (16) is feature relevancy: mutual information between a candidate feature and the class. Similarly, the second item represents feature redundancy. Eq. (16) can be rewritten as follows:

$$I(X_k;Y) = I(X_k;Y|X_j) + I(X_k;Y;X_j) \qquad (17)$$

Therefore, we can conclude that feature relevancy consists of two parts: new classification information and redundant information. Feature selection methods intend to maximize the new classification information while minimizing redundant information. Intuitively, we propose a novel feature selection method based on information theory named Composition of Feature Relevancy (CFR). Its criterion is as follows:

$$J(X_k) = \sum_{X_j \in S} \{I(X_k;Y|X_j) - I(X_k;Y;X_j)\} \qquad (18)$$

the criterion of CFR is equivalent to the following form; the derivation process is as follows:

$$\begin{aligned}
J(X_k) &= \sum_{X_j \in S} \{I(X_k;Y) - 2I(X_k;Y;X_j)\} \\
&= \sum_{X_j \in S} \{I(X_k;Y) - 2\{I(X_k;X_j) - I(X_k;X_j|Y)\}\} \\
&= |S| I(X_k;Y) - 2\sum_{X_j \in S} I(X_k;X_j) + 2\sum_{X_j \in S} I(X_k;X_j|Y) \\
&\simeq I(X_k;Y) - \frac{2}{|S|}\sum_{X_i \in S} I(X_k;X_j) + \frac{2}{|S|}\sum_{X_i \in S} I(X_k;X_j|Y)
\end{aligned}$$

CFR can also be regarded as a specific form of this Formula (Brownetal.,2012)

$$J(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_j;X_k) + \lambda \sum_{X_j \in S} I(X_j;X_k|Y)$$

For the sake of fairness, our method and the comparison methods both employ the same greedy searching strategy, namely, Sequential Forward Search(SFS).The feature subset S starts as an empty set, then, selects one informative feature based on the feature selection method each time until the number of selected features is larger than the user-specified threshold k. The steps of our method are described as follows:

1. (Initialization) Set F ← "Original feature set of n features", S ←"empty set".

2. (Calculate mutual information between the class with each candidate feature) For each feature $X_k \in F$, calculate $I(X_k;Y)$.

3. (Select the first feature) Find the feature that maximizes $I(X_k;Y)$, $F \leftarrow F\backslash\{X_k\}$; $S \leftarrow \{X_k\}$.

4. (Greedy selection) Repeat until$|S| = k$

(a) (Calculate conditional mutual information and interaction information) For all pairs of variables X_k and X_j such that $X_k \in F$, $X_j \in S$, calculate $I(X_k;Y|X_j)$ and $I(X_k;Y;X_j)$.

(b) (Select the next feature) choose the feature X_k that maximizes Formula (18).

5. The output is the set S that includes the selected features.

The method chooses the most relevant feature as the first feature. Then, interaction information and conditional mutual information are calculated to select the feature that maximizes the criterion in Formula (18). The procedure is ended when $|S| = k$.

Suppose the number of features to be selected is k, M represents the number of instances in the data set, and the total number of features is N. The time complexity of CFR is O(kMN).

[Results]

Knn with neighbor=5

I use python dictionary and Loop unrolling for efficiency

Accuracy train

	0.1N	0.2N	0.3N	0.4N	0.5N
COIL20	0.9339	0.9860	0.9972	0.9999	0.9999
ORL	0.9000	0.9057	0.9200	0.9375	0.9381
Musk	0.7263	0.8017	0.8471	0.8811	0.8843
BREAST	0.9470	0.9470	0.9650	0.9672	0.9717
COLON	0.7813	0.8072	0.8408	0.8387	0.8693

Accuracy test

	0.1N	0.2N	0.3N	0.4N	0.5N
COIL20	0.8986	0.9631	0.9791	0.9895	0.9861
ORL	0.7725	0.9057	0.8175	0.8025	0.8225
Musk	0.6680	0.7310	0.7794	0.8067	0.7899
BREAST	0.9413	0.9413	0.9499	0.9570	0.9542
COLON	0.7258	0.7096	0.7580	0.7419	0.7741

F1

	0.1N	0.2N	0.3N	0.4N	0.5N
COIL20	0.8956	0.9627	0.9790	0.9895	0.9860
ORL	0.7672	0.7700	0.8100	0.7961	0.8216
Musk	0.6630	0.7301	0.7792	0.8065	0.7899
BREAST	0.9354	0.9354	0.9447	0.9526	0.9494
COLON	0.7257	0.7084	0.7528	0.7375	0.7654