

WJMI: A New Feature Selection Algorithm Based on Weighted Joint Mutual Information ¹

Qi Xiuli^{1,a}, Yin Chengxiang^{1,b*}, Cheng Kai^{1,c}, Liao Xianglin^{1,d}, Kang Xingdang^{1,2,e}

¹PLA University of Science and Technology, Nanjing China

²Force 61175 of PLA, China

^aqixiuli@189.cn, ^b15951769048@163.com, ^cchengkai911@126.com, ^d876754706@qq.com, ^emailkxd@163.com

Keywords: feature selection, feature interaction, pruning rule, WJMI

Abstract. Mutual information based feature selection algorithms are very popular currently. Though they perform well in many cases, they suffer from two drawbacks: (1) the neglect of feature interaction; (2) the overestimation of some features. To overcome these shortcomings, a new feature evaluation criterion considering feature interaction is proposed and a pruning rule is designed. Based on the criterion and pruning rule, a new feature selection algorithm WJMI is proposed. Experiments carried out on UCI real world dataset against other four algorithms demonstrate the effectiveness of WJMI.

Introduction

Feature selection is a key procedure of machine learning and pattern recognition and a lot of researchers have paid their attention on it. Feature selection aims at selecting a feature subset that could represent the characteristics of the original feature space. A well selected feature subset could significantly reduce the dimension of origin problem and help to improve the efficiency and performance of learning algorithms.

Feature selection methods are usually divided into four categories: filter, wrapper, embedded and hybrid. In embedded methods, feature selection is accomplished in the training process such as decision tree and the selected feature subset works well on the specified classifier [1]. As they are restricted to one classifier, embedded methods are limited in terms of generalization. Wrapper methods search the feature space by testing all feature subsets on a predefined classifier and they perform well on that classifier. Besides the weakness that they are restricted to specified classifier and are easy to cause over fitting, wrapper methods also suffer from expensive computational complexity, moreover, the parameters in the predefined classifier make the problem even harder [2]. Filter methods rank features according to some criteria such as distance, correlation, consistency, etc. Filter methods are classifier-independent, so they have the best generalization ability. Meanwhile, they are much less expensive in computational complexity than wrapper methods. Hybrid methods are combinations of filters and wrappers.

Due to the advantages mentioned above, filter methods received the most attention. In filter methods, one of the key factors that affects the algorithm performance is the feature evaluation criterion. Information theory based criteria have been widely applied in filter methods, such as MIFS [3], JMI [4], CMIM [5], FOU [6], JMIM [7], etc. These algorithms use mutual information, conditional mutual information, joint mutual information, and normalized mutual information to measure relevance between candidate features and the class and redundancy between candidate features and selected features, then construct an evaluate criterion to balance the relevance and redundancy. Though these existing algorithms have performed well in many cases, their evaluate criteria have some drawbacks: (1) feature interaction is neglected; (2) some features may be overestimated.

In this paper we propose a new feature evaluation criterion based on weighted joint mutual information to overcome the shortcomings mentioned above. Moreover, we designed a pruning rule to avoid overestimating some features. On the basis of the evaluation criterion and the pruning rule, a

new filter feature selection method WJMI is proposed. The experiment carried out on UCI dataset against other algorithms proves the effectiveness of WJMI.

The rest of this paper is organized as follows. In Section 2, some preliminary of information theory is represented. In Section 3, related work is reviewed. In Section 4, we introduce our WJMI algorithm. In Section 5, experiments carried out on UCI dataset against other feature selection algorithms are discussed. In Section 6, we make a brief conclusion and discuss the future research direction.

Preliminary

This section introduces some basic information theory concepts related to this paper.

For a random variable $X = (x_1, x_2, \dots, x_n)$, the entropy of X is denoted by $H(X)$.

$$H(X) = -\sum_{i=1}^n p(x_i) \log(p(x_i)) \quad (1)$$

For two random variables X and $Y = (y_1, y_2, \dots, y_m)$, the joint entropy is denoted by $H(X, Y)$.

$$H(X, Y) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log(p(x_i, y_j)) \quad (2)$$

The conditional entropy of X given Y is denoted by $H(X | Y)$.

$$H(X | Y) = H(X, Y) - H(Y) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log(p(x_i | y_j)) \quad (3)$$

The mutual information of two random variables X and Y is denoted as $I(X; Y)$.

$$I(X; Y) = H(X) - H(X | Y)$$

$$I(X; Y) = H(Y) - H(Y | X) \quad (4)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

The conditional mutual information of two random variables X and Y given Z is denoted as $I(X; Y | Z)$.

$$I(X; Y | Z) = H(X | Z) - H(X | Y, Z) \quad (5)$$

The joint mutual information of three random variables X, Y, Z is denoted as $I(X, Y; Z)$.

$$I(X, Y; Z) = I(X; Z | Y) + I(Y; Z) \quad (6)$$

The interaction information of three random variables X, Y, Z is denoted as $I(X; Y; Z)$.

$$I(X; Y; Z) = I(X, Y; Z) - I(X; Z) - I(Y; Z) \quad (7)$$

Related work

Our work focuses on filter methods based on information theory, so in this section, we mainly review work belong to this context. In information theory context, feature selection in filter methods can be regarded as an optimization problem whose goal is to find a feature subset to maximize the mutual information between the selected feature subset and the class.

$$S_{opt} = \arg \max_{S \subseteq F} I(S; C) \quad (8)$$

Eq.8 gives the formal description of the discussion above where F is the full feature set and C is the class. Though the target is very clear, it is impractical to find the best feature subset due to the exponential scale of search space, especially when the dimension of data is high. As a result, researchers adopt forward selection strategy based on various criteria to get approximate optimized solution.

Battiti [3] proposed the MIFS algorithm to select features based on the criterion J_{MIFS} .

$$J_{MIFS} = I(f_i; C) - b \sum_{f_j \in S} I(f_i; f_j) \quad (9)$$

In the process of forward selection, MIFS selects the feature f_i maximizes J_{MIFS} each time.

Kwak [8] proposed mutual information feature selector under uniform information distribution (MIFS-U) based on the criterion:

$$J_{MIFS-U} = I(f_i; C) - b \sum_{f_j \in S} \frac{I(f_j; C)}{H(f_j)} I(f_i; f_j) \quad (10)$$

Yang [4] proposed JMI algorithm whose criterion is:

$$J_{JMI} = \sum_{f_j \in S} I(f_i, f_j; C) \quad (11)$$

Peng [9] proposed mRMR algorithm based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.

$$J_{mRMR} = I(f_i; C) - \frac{1}{|S|-1} \sum_{f_j \in S} I(f_i; f_j) \quad (12)$$

Based on conditional mutual information, Fleuret [5] proposed CMIM algorithm whose criterion is:

$$J_{CMIM} = \min_{f_j \in S} [I(f_i; C | f_j)] \quad (13)$$

Gavin Brown [6] proposed the FOU algorithm based on the criterion J_{FOU} .

$$J_{FOU} = I(f_i; C) - b \sum_{f_j \in S} I(f_i; f_j) + g \sum_{f_k \in S} I(f_i; f_k | C) \quad (14)$$

Bennasar [7] proposed the JMIM algorithm based on J_{JMIM} .

$$J_{JMIM} = \min_{f_s \in S} I(f_i, f_s; C) \quad (15)$$

Of all algorithms mentioned above, FOU is the only one that takes feature interaction into consideration. However, the parameters b and g in J_{FOU} bring additional difficulties. Interacting features are those that appear to be irrelevant with the class individually, but when it combined with other features, it may highly correlate to the class [10]. In addition to the neglect of feature interaction, MIFS, JMI, MIFS-U, mRMR, FOU share another drawback that some of features would be overestimated. This may occur when a candidate feature is completely correlated with one or several pre-selected features, but is almost independent with other pre-selected features, then the criteria J_{MIFS} , J_{JMI} , J_{MIFS-U} , J_{mRMR} , J_{FOU} of this feature would be high despite the redundancy of this feature[7].

WJMI Algorithm

In order to overcome the drawbacks existing in current mutual information based filter methods, we design a new feature evaluation criterion J_{WJMI} .

$$J_{WJMI} = \sum_{f_j \in S} \frac{I(f_i, f_j; C)}{I(f_i; C) + I(f_j; C)} I(f_i, f_j; C) \quad (16)$$

J_{WJMI} is the weighted joint mutual information criterion with the weight factor

$w(f_i, f_j) = \frac{I(f_i, f_j; C)}{I(f_i; C) + I(f_j; C)}$. By the discussion of Zeng [10], interaction information can be

positive, negative or zero. When $I(f_i, f_j; C) > 0$, $w(f_i, f_j) > 1$, the combination of feature f_i and feature f_j brings additional information, then f_i should has larger probability to be selected. When $I(f_i, f_j; C) = 0$, $w(f_i, f_j) = 1$. When $I(f_i, f_j; C) < 0$, $w(f_i, f_j) < 1$, the combination of feature f_i and feature f_j brings some redundant information. From the discussion, it is clear that our weighting strategy is consistent with feature interaction theory.

With respect to the problem that some features may be overestimated, we introduce a pruning rule: for a candidate feature f_i , if there exists a feature f_j in the pre-selected feature subset S satisfying $w(f_i, f_j) \leq q (q \geq \frac{1}{2})$, f_i can be removed from candidate set directly. Now we give the explanation of the pruning rule. When $w(f_i, f_j)$ is less than some predefined parameter q , f_i and f_j is highly correlated with each other. For the extreme case that $q = \frac{1}{2}$, f_i and f_j are completely correlated with each other, f_i is obviously unnecessary to be considered.

Based on J_{WJMI} and the pruning rule, we propose the WJMI algorithm.

Algorithm 1: WJMI algorithm

Input: full feature set F , max count of feature to be selected K , the class C , weight threshold q .

Output: selected feature subset S .

1. Initialize $S = \emptyset$, $k = 1$;
2. Select the first feature $f_1 = \arg \max_{f_i \in F} I(f_i; C)$;
3. Set $S = S \cup \{f_1\}$, $F = F \setminus \{f_1\}$
4. While $k < K$ and $F \neq \emptyset$
5. For each $f_i \in F \setminus S$
6. For each $f_j \in S$
7. If $w(f_i, f_j) \leq q$
8. $F = F \setminus \{f_i\}$ break;
9. Set $k = k + 1$;
10. $f_k = \arg \max_{f_i \in F \setminus S} \sum_{f_j \in S} \frac{I(f_i, f_j; C)}{I(f_i; C) + I(f_j; C)} I(f_i, f_j; C)$;
11. Set $S = S \cup \{f_k\}$, $F = F \setminus \{f_k\}$.

WJMI begins with selecting the first feature (line 1-3), then WJMI adopts forward selection strategy (line 4-11) and selects the feature maximize J_{WJMI} each time. The pruning rule is represented in line 7 and 8.

Experiment

In order to validate the effectiveness of WJMI, we carry out experiment on 11 UCI dataset (Table 1) against four other algorithms (JMI, CFS, FCBF, mRMR) to compare the selected features by classification accuracy on three classifiers (C4.5, PART, IB1). In the implementation process of WJMI, we set $q = \frac{1}{2}$.

Table 1 Dataset used in the experiment

No.	Dataset Name	Number of Instances	Number of features	Number of classes
1	anneal	898	38	6
2	colic	368	22	2
3	hypothyroid	3772	29	4
4	ionosphere	351	34	2
5	kr-vs-kp	3196	36	2
6	lymphography	148	18	4
7	sonar	208	60	2
8	spectf	269	44	2
9	splice	3190	60	3
10	vehicle	846	18	4
11	wine	178	13	3

The experiment is carried out in WEKA environment, and we adopt the default parameters of the classifiers set by WEKA. All datasets are partitioned as training (70%) and testing (30%), the classification accuracies presented here are the average of 100 trials with different random initializations each time. Average accuracies and standard deviations are shown in Tables 2–4, respectively. The best result is represented in boldface.

Table 2 Average accuracy (%) on C4.5

Dataset	WJMI	JMI	CFS	FCBF	mRMR
anneal	98.80±0.61	98.87±0.58	97.29±0.88	97.28±0.88	98.78±0.63
colic	86.15±2.92	85.94±2.51	81.30±3.04	81.32±3.07	85.69±2.91
hypothyroid	99.34±0.27	99.31±0.28	97.67±0.41	97.73±0.39	99.28±0.31
ionosphere	92.24±2.34	92.04±2.55	89.67±2.91	89.95±2.70	90.21±2.96
kr-vs-kp	98.96±0.39	98.34±0.34	94.07±0.62	94.06±0.65	99.02±0.32
lymphography	80.48±4.81	79.73±5.04	74.52±5.72	73.04±5.56	76.21±5.89
sonar	82.71±4.47	79.68±4.77	78.39±5.15	77.66±4.75	78.83±5.09
spectf	84.16±4.15	82.65±3.93	81.24±3.69	83.58±4.27	84.53±4.27
splice	94.06±0.69	93.95±0.64	93.72±0.76	93.70±0.76	94.00±0.70
vehicle	70.76±2.51	70.38±2.68	69.05±2.90	56.22±2.64	70.13±2.63
wine	95.72±2.75	95.42±2.41	93.21±3.30	93.23±3.30	95.59±2.52
Avg.	89.4	88.75	86.37	85.25	88.39

Table 3 Average accuracy (%) on PART

Dataset	WJMI	JMI	CFS	FCBF	mRMR
anneal	98.40±0.97	98.45±0.97	96.80±1.06	96.80±1.08	98.67±0.67
colic	86.21±2.57	85.97±2.64	80.95±3.16	80.81±3.16	85.81±2.72
hypothyroid	99.41±0.30	99.43±0.26	97.70±0.36	97.73±0.36	99.40±0.28
ionosphere	92.47±2.27	91.22±3.15	90.11±3.08	90.06±2.98	90.71±2.52
kr-vs-kp	98.69±0.39	98.15±0.45	94.15±0.63	94.13±0.63	98.56±0.40
lymphography	80.18±4.56	78.43±5.89	76.95±6.18	76.13±5.81	77.73±6.04
sonar	84.35±3.94	81.44±4.54	79.73±4.63	79.36±4.58	80.58±4.56
spectf	84.84±3.20	83.10±3.63	82.01±3.72	84.64±3.67	84.97±3.64
splice	93.83±0.81	93.87±0.84	92.47±0.88	92.45±0.88	93.85±0.77
vehicle	71.19±2.98	70.45±2.86	69.00±2.65	55.99±2.50	69.88±2.72
wine	95.45±3.28	94.79±2.42	93.30±4.45	93.52±4.18	95.22±3.20
Avg.	89.55	88.66	86.65	85.6	88.67

Table 4 Average accuracy (%) on IB1

Dataset	WJMI	JMI	CFS	FCBF	mRMR
anneal	99.59±0.45	99.04±0.70	97.47±0.97	97.54±0.91	99.25±0.55
colic	85.59±2.84	85.73±2.36	82.23±2.97	82.14±3.16	85.74±2.89
hypothyroid	99.29±0.24	98.88±0.29	97.67±0.40	97.78±0.39	99.04±0.25
ionosphere	93.10±2.23	93.59±1.98	91.68±2.37	90.00±2.47	91.53±2.25
kr-vs-kp	97.78±0.45	96.92±0.48	94.17±0.62	94.17±0.62	97.36±0.52
lymphography	85.02±5.13	84.43±4.64	82.52±5.33	78.17±5.50	84.37±5.00
sonar	84.66±3.76	82.71±4.12	79.49±4.27	79.14±4.08	82.01±4.27
spectf	84.88±3.28	82.51±3.74	78.30±4.06	85.33±3.37	85.72±3.33
splice	89.59±0.95	89.51±1.05	80.28±1.18	80.25±1.20	89.60±0.88
vehicle	71.36±2.66	71.22±2.24	68.94±2.43	56.68±2.59	70.40±2.37
wine	98.55±1.46	98.28±1.49	97.55±1.78	98.20±1.56	98.31±1.62
Avg.	89.95	89.35	86.39	85.4	89.39

From the result in Table 2-4, it can be found that WJMI outperforms the other four algorithms on almost all datasets and classifiers. All of the results above prove the effectiveness of our algorithm.

Conclusion

This paper proposes a new filter feature selection algorithm based on weighted joint mutual information-WJMI. WJMI takes feature interaction into consideration and avoids overestimating some features. Experiments carried out on UCI real world dataset against other algorithms prove the effectiveness of WJMI. In future research, we would focus on exploring new methods of measuring the relevance between whole feature subset and the class.

References

[1] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." *Journal of Machine Learning Research* (2003): 1157-1182.

- [2] Tsanas, Athanasios, Max A. Little, and Patrick E. McSharry. "A simple filter benchmark for feature selection." *Journal of Machine Learning Research*(2010): 1-24.
- [3] Battiti, Roberto. "Using mutual information for selecting features in supervised neural net learning." *Neural Networks, IEEE Transactions on* 5.4 (1994): 537-550.
- [4] Yang, Howard Hua, and John E. Moody. "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data." *NIPS*. Vol. 99. 1999.
- [5] Fleuret, François. "Fast binary feature selection with conditional mutual information." *The Journal of Machine Learning Research* 5 (2004): 1531-1555.
- [6] Brown, Gavin. "A new perspective for information theoretic feature selection." *International conference on artificial intelligence and statistics*. 2009.
- [7] Bennasar, Mohamed, Yulia Hicks, and Rossitza Setchi. "Feature selection using Joint Mutual Information Maximisation." *Expert Systems with Applications* 42.22 (2015): 8520-8532.
- [8] Kwak, Nojun, and Chong-Ho Choi. "Input feature selection for classification problems." *Neural Networks, IEEE Transactions on* 13.1 (2002): 143-159.
- [9] Peng Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27.8 (2005): 1226-1238.
- [10] Zeng, Zilin, et al. "A novel feature selection method considering feature interaction." *Pattern Recognition* 48.8 (2015): 2656-2666.