Feature selection based on weighted conditional mutual information

Hongfang Zhou
School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China and
Shaanxi Key Laboratory of Network Computing and Security Technology, Xi'an, China, and
Xiqian Wang and Yao Zhang
School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China

Received 9 August 2019
Revised 21 November 2019
Accepted 24 December 2019

Abstract

Feature selection is an essential step in data mining. The core of it is to analyze and quantize the relevancy and redundancy between the features and the classes. In CFR feature selection method, they rarely consider which feature to choose if two or more features have the same value using evaluation criterion. In order to address this problem, the standard deviation is employed to adjust the importance between relevancy and redundancy. Based on this idea, a novel feature selection method named as Feature Selection Based on Weighted Conditional Mutual Information (WCFR) is introduced. Experimental results on ten datasets show that our proposed method has higher classification accuracy.

Keywords Feature selection, Conditional mutual information, Standard deviation

Paper type Original Article

This work was supported by the National Key Research and Development Plan, China [grant number 2017YFB1402103]; the National Natural Science Foundation of China [grant numbers 61402363, 61771387]; the Education Department of Shaanxi Province Key Laboratory Project [grant number 15JS079]; Xi'an Science Program Project [grant number 2017080CG/RC043 (XALG017)]; the Ministry of Education of Shaanxi Province Research Project [grant number 17JK0534]; and Beilin district of Xi'an Science and Technology Project [grant number GX1625]. The corresponding author would like to thank such financial aids.

Declaration of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publishers note: The publisher wishes to inform readers that the article "Feature selection based on weighted conditional mutual information" was originally published by the previous publisher of *Applied Computing and Informatics* and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use Zhou, H., Wang, X., Zhang, Y. (2019), "Feature selection based on weighted conditional mutual information", *Applied Computing and Informatics*. Vol. ahead-of-print No. ahead-of-print. https://10.1016/j.aci.2019.12.003. The original publication date for this paper was 07/01/2020.

Applied Computing and Informatics
Vol. 20 No. 1/2, 2024
pp. 55-68
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1016/j.aci.2019.12.003

1. Introduction

Feature selection is an important step in pattern recognition and data mining. The target features contain fewer ones, which can reduce the training time and improve the interpretability of the model [1]. Different from other dimension reduction technique, feature selection doesn't produce the new combinations of features. In other words, it only selects features [2].

Feature selection methods can be broadly classified into three types such as filter method [3,4], wrapper method [5–7] and embedded method [8]. Feature selection based on mutual information belongs to the first one. Different from the latter two ones, the filter method is independent of the classifier [9]. Accordingly, the filter method is usually faster than other two ones [10].

Mutual information is usually used to measure the relation between two variables [11]. Feature selection method based on it regards mutual information as feature selection criterion. The definition of MI can be shown in Eq. (1).

$$I(X_m; C|S) \tag{1}$$

where X_m is the candidate feature, S denotes the selected feature set and C is the class. Obviously, the higher the score is, the more important the feature is. However, it involves massive calculation of high dimensional joint probability. In some literatures, the high-order mutual information is decomposed into the sum of some multiple low-order mutual information under some independence assumptions [12]. But in real world, such assumption is unrealistic. In face of this, this paper presents a feature selection method-WCFR (Weight Composition of Feature Relevancy). In our proposed algorithm, standard deviation is applied to weigh the relations between the features and the feature sets.

This paper is organized as follows. Section 2 introduces the related work and some classical feature selection methods based on mutual information. In Section 3, we describe the proposed approach-WCFR. The experimental results on large amounts of data are given in Section 4. Section 5 gives the conclusion and future problem.

2. Related work

Feature selection based on MI belongs to the filter method. It uses the mutual information to select and evaluate the features [9]. Forward search is a greedy algorithm that selects one feature in per iteration. In this way, the feature set can be obtained after some iterations. Recently, the feature selection methods based on MI have already been proposed.

The earliest method using mutual information is MIM [13], in which MI is used to evaluate the relation between the feature and the class. The evaluation criterion is shown as follow.

$$J(X_m) = I(X_m; C) \tag{2}$$

where X_m is the candidate feature and C denotes the class. MIM is simple. It ignores the relations between the features and the selected feature sets, which could lead to a situation in which the feature subset involves too many redundant features. Mutual Information based on Feature Selection (MIFS) [14] is proposed by Battiti, and it is shown as Eq. (3).

$$J(X_m) = I(X_m; C) - \beta \sum_{X_s \in S} I(X_m; X_s) \tag{3}$$

Where X_s denotes the selected feature from the selected feature set S. MIFS considers the redundancy between the candidate feature and the selected features based on MIM. mRMR [15] is the variant of MIFS. However, there is a potential problem about mutual information. It tends to select some features containing more values. Let's take an example. When some features of sequential data are to be faced, such method will be incapable. To avoid such situation, some researches normalized the mutual information by scaling the value of mutual information to the interval from 0 to 1. La The Vinh has proposed NMIFS (Normalized Mutual Information Feature Selection) [16], which is shown in Eq. (4).

$$J_{NMIFS}(X_m) = \frac{I(X_m;C)}{\log_2(|\Omega_C|)} - \frac{1}{|S_m - 1|}\sum_{X_s \in S_{m-1}} \frac{I(X_s;X_m)}{\log_2(|\Omega_{X_m}|)} \tag{4}$$

where $|\Omega_k|$ denotes the size of the sample space of the variable k. MIFS or NMIFS represents a general idea that the candidate feature should be high-relevancy with C and be low-redundancy with S. Some evaluation criteria of feature selection are developed in this aspect. Conditional Informative Feature Extraction (CIFE) [17] is proposed by Lin and Tang, and the corresponding criterion is shown as follow.

$$J_{CIFE}(X_m) = I(X_m;C) - \sum_{X_s \in S}\{I(X_m;X_s) - I(X_m;X_s|C)\} \tag{5}$$

where $I(X_m;X_s|C)$ denotes the redundancy between X_m and X_s when giving C. The description about redundancy in CIFE is more specific than that in MIFS. $I(X_m;X_s) - I(X_m;X_s|C)$ is named as intra-class redundancy [1]. H.Y.Yang proposed JMI [18], which is shown in Eq. (6).

$$J_{JMI}(X_m) = I(X_m;C) - \frac{1}{|S|}\sum_{X_s \in S}[I(X_m;C) - I(X_m;C|X_s)] \tag{6}$$

By analyzing JMI, we can get Eq. (7).

$$J_{JMI}(X_m) = I(X_m;C) - \frac{1}{|S|}\sum_{X_s \in S}[I(X_m;X_s) - I(X_m;X_s|C)] \tag{7}$$

Therefore, JMI can be regarded as CIFE that adds the weight $\frac{1}{|S|}$. The second item in Eq. (7) uses the average to reflect the centralized tendency of the redundancy. RelaxFS [12], as shown in Eq. (8), is proposed by Vinh* and Zhou, which introduces the new redundancy containing more redundant information.

$$\begin{aligned} J_{RelaxFS}(X_m) = I(X_m;C) &- \frac{1}{|S|}\sum_{X_j \in S} I(X_m;X_j) + \frac{1}{|S|}\sum_{X_j \in S} I(X_m;X_j|C) \\ &- \frac{1}{|S||S-1|}\sum_{X_j \in S}\sum_{\substack{X_i \in S \\ i \neq j}} I(X_m;X_i|X_j) \end{aligned} \tag{8}$$

where $\frac{1}{|S||S-1|}\sum_{X_j \in S}\sum_{\substack{X_i \in S \\ i \neq j}} I(X_m;X_i|X_j)$ denotes the redundancy between X_m and X_i when giving X_j.

It should be noted that X_i and X_j are from the selected feature set S. Therefore, $\frac{1}{|S||S-1|}\sum_{X_j \in S}\sum_{\substack{X_i \in S \\ i \neq j}} I(X_m;X_i|X_j)$ contains more redundancy between X_m and S than $\frac{1}{|S|}\sum_{X_s \in S}[I(X_m;X_s) - I(X_m;X_s|C)]$. There is also a feature selection method, named as CFR [19], which is proposed by Wangfu Gao. Feature relevancy is composed of two parts in CFR, and it is shown in Eq. (9).

$$I(X_m;C) = I(X_m;C|X_s) + I(X_m;C;X_s) \tag{9}$$

where $I(X_m;C|X_s)$ denotes the information related to the class, $I(X_m;C;X_s)$ denotes the redundant information. The criterion of WCFR maximizes the correlation and minimizes the redundancy, which is shown in Eq. (10).

$$J_{CFR}(X_m) = \sum_{X_s \in S} I(X_m;C|X_s) - \sum_{X_s \in S} I(X_m;C;X_s) \tag{10}$$

In above mentioned methods, we can see there is a trend that relation between features is more concrete from mutual information to conditional mutual information. However, it involves lots of computation. For example, $\frac{1}{|S||S-1|} \sum_{X_j \in S} \sum_{\substack{X_i \in S \\ i \neq j}} I(X_m;X_i|X_j)$ can contain more redundant information than $\frac{1}{|S|} \sum_{X_s \in S} [I(X_m;X_s) - I(X_m;X_s|C)]$. However, the computational complexity of the former is higher than that of the latter. As a matter of fact, X_i and X_j are from the selected feature set S. When the redudant term $\frac{1}{|S||S-1|} \sum_{X_j \in S} \sum_{\substack{X_i \in S \\ i \neq j}} I(X_m;X_i|X_j)$ is calculated, it is required to search S twice. Therefore, how to improve the efficiency of feature selection without increasing the amount of computation is a problem.

3. Proposed method

3.1 Problem and method

The new proposed method, WCFR (Weighted Composition of Feature Relevancy), is the improvement of CFR. Eq. (11) can be gotten from the Eq. (9), which indicates the mutual information between X_m, C and X_s.

$$I(X_m;C;X_s) = I(X_m;C) - I(X_m;C|X_s) = I(X_m;X_s) - I(X_m;X_s|C) \tag{11}$$

When we use the r.h.s of Eq. (11) instead of the redundancy term in Eq. (10), we can get Eq. (12).

$$\begin{aligned} J_{CFR}(X_m) &= \sum_{X_s \in S} I(X_m;C|X_s) - \sum_{X_s \in S} I(X_m;C) - I(X_m;C|X_s) \\ &= \sum_{X_s \in S} I(X_m;C|X_s) - \sum_{X_s \in S} I(X_m;X_s) - I(X_m;X_s|C) \end{aligned} \tag{12}$$

The criteria of CFR is similar to those of MIM, mRMR, JMI, CIFE and Relaxmrmr, because all of them try to search the features that are high relevant with the class and low redundant with the selected feature set. Another common point is that relevancy and redundancy are expressed by the summation. In fact, there exists a new problem. Suppose there are two features which are X_1 and X_2. The relevance of the two features using Eq. (12) can get the same value. How to distinguish X_1 and X_2 is a problem. $I(X_m;C|X_s)$ denotes the information that X_m can provide while X_s cannot. The value of $I(X_m;C|X_s)$ is different for each X_s. However, the summation on $I(X_m;C|X_s)$ ignores the difference. Therefore, the difference measured by standard deviation is introduced in the proposed method. And it is shown in Eq. (13).

$$J_{WCFR}(X_m) = (1-\delta_1)\times\sum_{X_s\in S} I(X_m;C|X_s) - (1-\delta_2)\times\sum_{X_s\in S} I(X_m;X_s) - I(X_m;X_s|C) \quad (13)$$

In Eq. (13), the expressions of standard deviation δ_1 and δ_2 are shown as Eqs. (14) and (16).

$$\delta_1 = \sqrt{\frac{1}{|S|}\sum_{i=1}^{|S|}\left(I(X_m;C|X_i) - \mu_1\right)^2} \quad (14)$$

$$\mu_1 = \frac{1}{|S|}\sum_{i=1}^{|S|} I(X_m;C|X_i) \quad (15)$$

$$\delta_2 = \sqrt{\frac{1}{|S|}\sum_{i=1}^{|S|}\left(I(X_m;X_i) - I(X_m;X_i|C) - \mu_2\right)^2} \quad (16)$$

$$\mu_2 = \frac{1}{|S|}\sum_{i=1}^{|S|} I(X_m;X_i) - I(X_m;X_i|C) \quad (17)$$

Standard deviation is usually used to measure the degree of dispersion. Hence, we use it to adjust the importance degrees of relevant items or redundant items in WCFR. The higher the value of the standard deviation is, the higher the degree of dispersion is. In this way, WCFR can tackle above problem that how to select X_1 and X_2 when the summation of X_1 and X_2 on $I(X_m;C|X_s)$ is equal.

The Pseudo-code of WCFR is shown in Table 1. It contains two parts. The first part is the initialization of the selected feature set S (Lines 1–8), and the second part is the process of iteration in which it selects one feature in each iteration by the Eq. (13).

```
Algorithm WCFR: Weighted Composition of Feature Relevancy

Input: D dataset, C class, n the number of features, k the number of selected
   feature.Output: S selected feature subset.
Begin:
1. S = Φ
2. for i = 1 to n
3.   $relevance(F(i)) \leftarrow MI(F(i);C)$
4. end for
5. $f_{new} \leftarrow f_m satisfying f_m = \text{argmax}_{f_m \in F} relevance(f_m)$
6. $S = S \cup f_{new}$
7. $F = F - f_{new}$
8. count = 1;
9. while count < k
10.   l = n-count;
11.   for m = 1 to l
12.     for i = 1 to count
13.       Calculate $J_{WCFR}(X_m)$ according to Eq. (13)
14.     end for
15.   end for
16.   $f_{new} \leftarrow f_m satisfying f_m = \text{argmax}_{f_m \in F} J_{WCFR}(f_m)$
17.   $S = S \cup f_{new}$
18.   $F = F - f_{new}$
19.   count = count + 1
20. end while
End
```

Table 1.
Pseudo-code of WCFR.

3.2 Complexity analysis

WCFR contains a ‘while’ loop and two ‘for’ loops, and its time complexity is $O(k^2nm)$ (k is the number of selected features, n is the number of all features, $m$is the number of samples). The Complexity of WCFR is same as that of CFR, CIFE and JMI. And it is higher than that of MIM and is lower than that of RelaxFS.

4. Experiments

4.1 Data sets

To verify the effectiveness of the proposed WCFR, ten data sets are used in the experiments. They are from different fields and can be found in UCI [20]. There are hand written digital data (Semeion, Mfeatfac), text data (CANE9), voice data (Isolet), image data (ORL, COIL20, WarpPIE10p) and biodata (TOX171). More detail descriptions can be found in Table 2. Each data will be normalized and discretized, which is similar to other literatures [12,19].

4.2 Experiment settings

In this experiment, we use Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) to evaluate different feature selection methods. They are two classical and widely used classifiers in relative references [12,19]. In such mentioned ones, K is set to be 3 for KNN and linear kernel is used for SVM [12]. Analogously, we do the same strategy.

The experiment consists of three parts. The first part is data preprocessing. For the validity of the calculation, the value of every feature is shrunk into −1 to1 and is categorized into five equal-size bins. The second part is feature subsets generation. If the number of features is less than 50, the size of the feature subset is equal to the size of the features. Otherwise, the size of the feature subset is set to be 50. Feature selection methods are used to generate feature subset. The third part is the feature subsets evaluation. In this experiment, we use average classification accuracy and Macro-F1 to evaluate the classifiers on feature subset. Classification accuracy means the proportion of the number of correctly classified samples to the total number of samples. F1 is defined as follows.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{18}$$

where P denotes the precision and R is the Recall. F1 can be used to measure the binary classification problem. If the number of categories is greater than two, Macro average F1 can be used to treat F value of multi-class classification problems as the average F value of n binary classification. Macro average F1 is defined in Eq. (19).

Table 2.
Descriptions of datasets.

No	data	# of feature	# of examples	# of class
1	Vehicle	18	946	4
2	Sonar	60	208	2
3	Mfeatfac	216	2000	10
4	Semeion	256	1593	10
5	Isolet	617	1560	26
6	CANE9	857	1080	9
7	ORL	1024	400	40
8	COIL20	1024	1440	20
9	WarpPIE10p	2420	210	10
10	TOX 171	5749	171	4

$$Macro\_F1 = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{19}$$

4.3 Experiment results and discussion

4.3.1 Comparisons on classification accuracy. Tables 3 and 4 are average classification accuracy using KNN and SVM on ten datasets. If m is made as the size of feature subset, its variation range is from 1 to 50. We calculate the classification accuracy with 10-fold cross-validation for each m, the value in cell can be gotten by averaging accuracy corresponding to different m. The maximum value of each row in the table is identified by bold fonts. The row named as 'Average' means the average classification accuracy on all the datasets.

WCFR used a Kolmogorov-Smirnov test with other existing MI-based methods. The Kolmogorov-Smirnov test is a non-parametric test method [21] that does not require knowledge of the data distribution. The default significance level of the K-S test is 5% and we use it in our experiment. If the P value is less than 5%, the two algorithms are considered to have significant difference while if the P value is greater than 5%, there is no significant difference. Tables 3 and 4 show the results of the experiment. We employ '+', '=' and '-' to indicate that WCFR performs 'better than', 'equal to' and 'worse than' other methods. The last row of Tables 3 and 4 named as "W/T/L" implies the statistical results that WCFR win/tie/loss compared to other methods. The statistical results are summarized in Figure 1.

It can be seen from Table 3 that the highest result is obtained by RelaxFS, CFR and WCFR. The possible reason is that these three methods describe the relations between the features and the classes more precisely than other methods. In Eq. (8), Eqs. (10) and (13), relevancy and redundancy in criterion of RelaxFS, CFR and WCFR are measured by conditional mutual information that is different from that of MIM, JMI, mRMR and CIFE. The average classification accuracy of RelaxFS is higher slightly than that of CFR, which is because RelaxFS can eliminate more redundant information by the second item of Eq. (8). Our proposed WCFR outperforms RelaxFS and CFR on seven data sets. The reason is that the weight is introduced in WCFR.

Table 4 shows the classification results for KNN. We can find that the different classifier have a different influence on the verify of feature subset. When KNN is used as the classifier, the result of RelaxFS is better slightly than that of CFR, but the new WCFR method is better than RelaxFS. RelaxFS can get the highest accuracy on TOX171, which means the hypothesis in RelaxFS meets the pattern of data ToX171 on KNN. As can be seen from Tables 3 and 4, WCFR can improve classifier accuracy when compared with other methods.

In order to observe the influence of feature subset size on accuracy, performance on different dataset is given in Figures 2 and 3. In the Figure 2(a), Figures 2(b) and 3(a), the trends of WCFR and CFR are similar, especially when the dimension number of the data goes up. But WCFR is actually better slightly than CFR if we can combine them with Table 3. The accuracy of WCFR is lower than that of CFR on TOX17 while is higher than that of RelaxFS. It means the weight added in CFR cannot influence the performance of CFR to a large extent in the worst case. On the other datasets, the accuracy of CFR is higher than that of RelaxFS on Sonar and Isolet; the accuracy of the RelaxFS is higher than that of CFR on Semeion, ORL and COIL 20. However, it is obvious that the proposed method outperforms CFR, RelaxFS, MIM, JMI, mRMR and CIFE on these data sets.

Tables 5 and 6 show the highest classification accuracy of seven algorithms for SVM and KNN. In the Table 5, the results of WCFR are same as that of CFR on Vehicle, Sonar and CANE9, and are better than that of CFR on the rest datasets. The highest classification of WCFR on KNN is worse than on that of SVM. This situation is similar to the above experimental results. Therefore, WCFR is more suitable for SVM than KNN. In the Table 6,

Table 3. Classification Accuracy of Seven Algorithms using SVM.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	57.86 ± 0.14(=)	62.99 ± 0.07(=)	62.04 ± 0.09(=)	64.94 ± 0.07(=)	63.47 ± 0.07(=)	65.01 ± 0.07(=)	**65.15 ± 0.07**
Sonar	77.21 ± 0.03(+)	78.96 ± 0.03(=)	77.50 ± 0.02(+)	77.35 ± 0.02(+)	77.99 ± 0.02(+)	78.74 ± 0.03(=)	**79.34 ± 0.03**
Mfeatfac	85.89 ± 0.12(+)	89.06 ± 0.11(+)	89.72 ± 0.11(+)	88.76 ± 0.11(+)	90.24 ± 0.11(=)	90.39 ± 0.11(=)	**90.58 ± 0.11**
Semeion	59.53 ± 0.15(+)	64.00 ± 0.12(+)	66.00 ± 0.12(+)	63.63 ± 0.10(+)	**68.58 ± 0.12**(=)	67.27 ± 0.11(=)	68.41 ± 0.12
Isolet	41.52 ± 0.14(+)	61.31 ± 0.13(+)	62.49 ± 0.13(+)	58.11 ± 0.11(+)	68.96 ± 0.15(+)	69.94 ± 0.15(+)	**72.49 ± 0.16**
CANE9	71.41 ± 0.18(=)	72.68 ± 0.16(=)	72.60 ± 0.16(=)	67.98 ± 0.13(+)	73.29 ± 0.16(=)	73.17 ± 0.16(=)	**73.71 ± 0.16**
ORL	58.15 ± 0.23(+)	82.17 ± 0.21(+)	81.94 ± 0.20(+)	57.24 ± 0.14(+)	**82.21 ± 0.21**(=)	79.62 ± 0.21(+)	81.81 ± 0.22
COIL20	63.77 ± 0.20(+)	85.63 ± 0.14(+)	87.50 ± 0.14(+)	87.11 ± 0.14(+)	89.95 ± 0.14(+)	88.42 ± 0.15(+)	**90.23 ± 0.15**
WarpPIE10p	84.50 ± 0.15(+)	92.54 ± 0.12(+)	93.53 ± 0.12(+)	93.71 ± 0.12(+)	93.67 ± 0.12(+)	94.09 ± 0.12(+)	**94.92 ± 0.12**
TOX 171	68.49 ± 0.07(+)	76.54 ± 0.07(+)	77.40 ± 0.07(+)	76.72 ± 0.09(=)	79.63 ± 0.07(=)	**80.97 ± 0.12**(=)	80.52 ± 0.11
Average	66.83 ± 0.14	76.59 ± 0.12	77.07 ± 0.12	73.56 ± 0.10	78.80 ± 0.12	78.76 ± 0.12	**79.72 ± 0.13**
W/T/L	8/2/0	7/3/0	8/2/0	8/2/0	4/6/0	4/6/0	

Table 4. Classification Accuracy of Seven Algorithms using KNN.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	58.54 ± 0.13(+)	64.79 ± 0.08(=)	62.13 ± 0.09(+)	66.33 ± 0.08(+)	64.85 ± 0.07(+)	66.47 ± 0.08(=)	**67.18 ± 0.08**
Sonar	79.34 ± 0.10(+)	83.63 ± 0.08(=)	81.28 ± 0.08(=)	84.56 ± 0.07(=)	83.75 ± 0.08(=)	84.02 ± 0.07(=)	**84.50 ± 0.07**
Featfac	81.03 ± 0.19(+)	86.33 ± 0.17(+)	87.23 ± 0.17(+)	87.19 ± 0.17(+)	88.34 ± 0.17(=)	88.11 ± 0.17(+)	**88.64 ± 0.17**
Semeion	54.57 ± 0.18(+)	59.70 ± 0.15(+)	61.77 ± 0.15(=)	61.09 ± 0.12(+)	64.55 ± 0.14(=)	63.66 ± 0.14(=)	**64.76 ± 0.14**
Isolet	34.91 ± 0.14(+)	53.45 ± 0.13(+)	54.74 ± 0.13(+)	40.43 ± 0.07(+)	60.96 ± 0.15(+)	62.79 ± 0.16(+)	**65.10 ± 0.17**
CANE9	68.68 ± 0.18(=)	70.16 ± 0.17(=)	70.19 ± 0.17(=)	62.99 ± 0.11(+)	71.17 ± 0.16(=)	70.94 ± 0.17(=)	**71.40 ± 0.16**
ORL	52.83 ± 0.19(+)	72.55 ± 0.17(+)	73.79 ± 0.17(=)	47.16 ± 0.10(+)	**74.97 ± 0.19**(=)	70.66 ± 0.17(+)	73.50 ± 0.19
COIL20	62.43 ± 0.24(+)	87.90 ± 0.18(+)	89.05 ± 0.18(+)	90.63 ± 0.19(+)	90.77 ± 0.19(+)	90.66 ± 0.19(+)	**91.37 ± 0.19**
WarpPIE10p	80.50 ± 0.18(+)	92.20 ± 0.14(+)	92.76 ± 0.15(=)	83.87 ± 0.11(+)	92.88 ± 0.14(=)	93.26 ± 0.14(+)	**93.71 ± 0.14**
TOX 171	64.70 ± 0.10(+)	78.45 ± 0.12(+)	78.46 ± 0.13(+)	81.71 ± 0.13(=)	**90.94 ± 0.12**(=)	85.00 ± 0.13(=)	84.85 ± 0.12
Average	63.75 ± 0.16	74.92 ± 0.14	75.14 ± 0.14	70.60 ± 0.12	78.32 ± 0.14	77.56 ± 0.14	**78.50 ± 0.14**
W/T/L	9/1/0	7/3/0	5/5/0	8/2/0	3/7/0	5/5/0	

WCFR is also the best feature selection method except on TOX171. On the whole, the four tables and three figures show the same consequence that the weight used in WCFR has worked and can improve CFR.

4.3.2 Comparisons on Macro-F1. In order to measure the influence of weight in WCFR, Macro-F1 is used to evaluate the results of classifiers on different data subsets. Table 7 shows

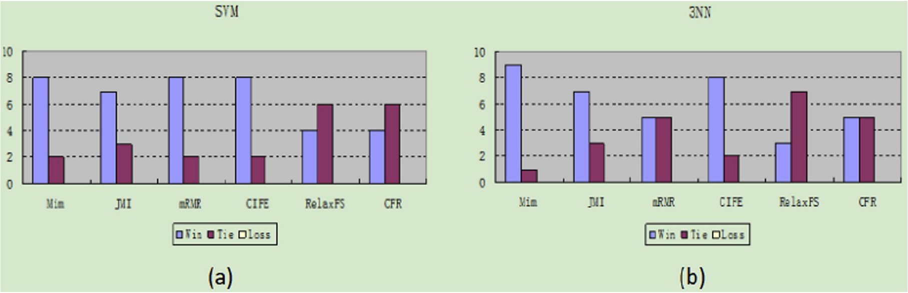
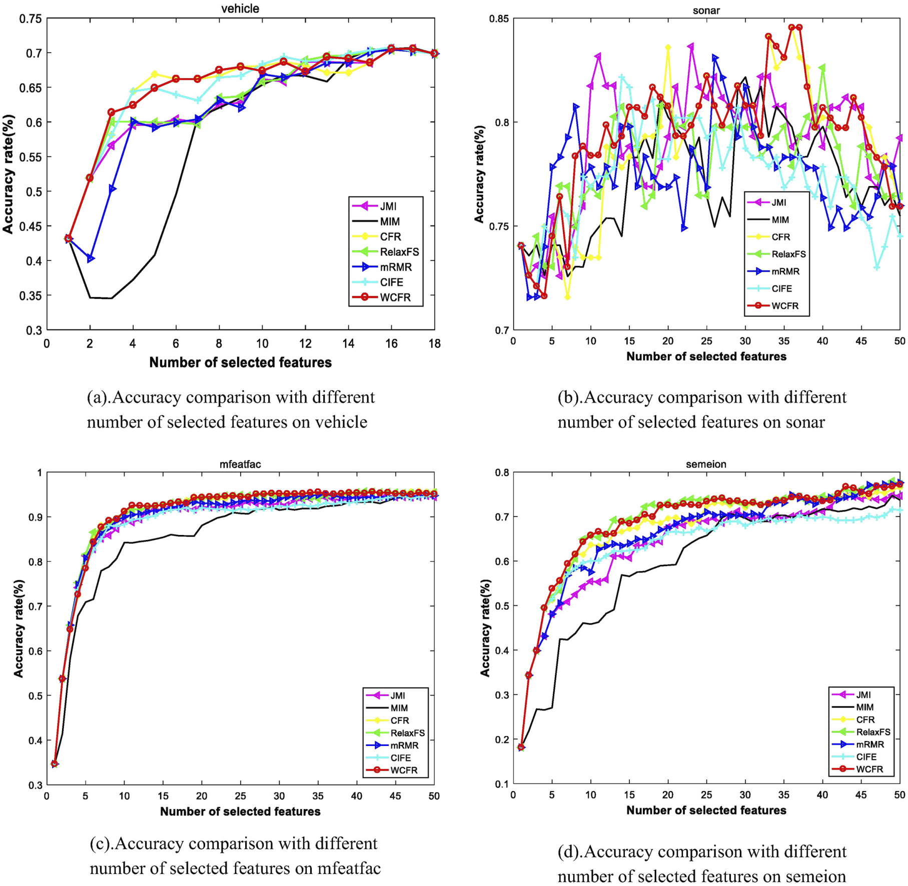


Figure 1.
Performance comparison of WCFR with other methods using K-S test. Win/tie/loss means 'better'/'equal to'/'worse' than other methods.



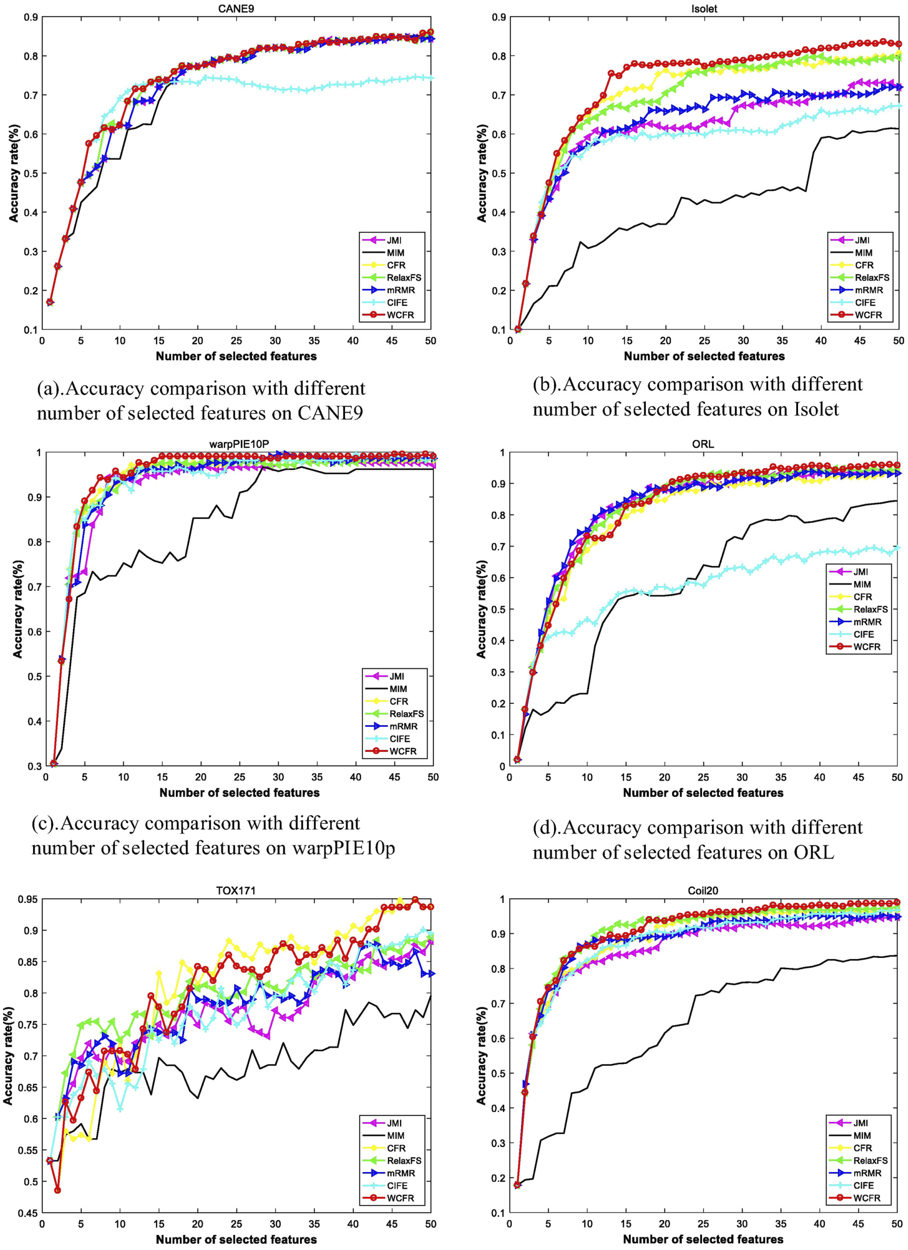
(a).Accuracy comparison with different number of selected features on vehicle

(b).Accuracy comparison with different number of selected features on sonar

(c).Accuracy comparison with different number of selected features on mfeatfac

(d).Accuracy comparison with different number of selected features on semeion

Figure 2.
Performance comparison on low-dimensional data sets with SVM.



(a).Accuracy comparison with different number of selected features on CANE9

(b).Accuracy comparison with different number of selected features on Isolet

(c).Accuracy comparison with different number of selected features on warpPIE10p

(d).Accuracy comparison with different number of selected features on ORL

(e).Accuracy comparison with different number of selected features on TOX171

(f).Accuracy comparison with different number of selected features on Coil20

Figure 3.
Performance comparison on high-dimensional data sets with SVM.

the result of Macro-F1 with SVM. It can be seen that F1 of WCFR is higher than that of CFR on all datasets, and is lower than that of RelaxFS only on Semeion and ORL. The evaluation criterion of RelaxFS can eliminate more redundant information while the weight in WCFR can adjust the importance between relevancy and redundancy. In the Table 8, the

Table 5.
Highest Classification Accuracy of Seven Algorithms using SVM.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	0.7068	0.7056	0.7044	0.7079	0.7044	**0.7056**	**0.7056**
Sonar	0.8217	0.8364	0.8310	0.8217	0.8262	**0.8455**	**0.8455**
Mfeatfac	0.9460	0.9455	0.9530	0.9440	0.9550	0.9540	**0.9580**
Semeion	0.7471	0.7483	0.7760	0.7163	**0.7797**	0.7653	0.7716
Isolet	0.6147	0.7327	0.7199	0.6718	0.7987	0.8077	**0.8359**
CANE9	0.8519	0.8481	0.8481	0.7463	**0.8602**	**0.8602**	**0.8602**
ORL	0.8350	0.9575	0.9375	0.7225	0.9575	0.9500	**0.9650**
COIL20	0.8417	0.9521	0.9583	0.9653	0.9771	0.9729	**0.9903**
WarpPIE10p	0.9667	0.9810	**0.9952**	**0.9952**	0.9905	0.9857	**0.9952**
TOX 171	0.7958	0.8827	0.8775	0.9007	0.8892	0.9477	**0.9484**

Table 6.
Highest Classification Accuracy of Seven Algorithms on KNN.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	0.7116	0.7104	0.6902	0.7020	0.6939	**0.7197**	**0.7197**
Sonar	0.8850	0.8948	0.8836	**0.9133**	**0.9133**	0.9033	0.9040
Mfeatfac	0.9350	0.9435	0.9504	0.9485	0.9575	0.9535	**0.9605**
Semeion	0.7226	0.7690	0.7652	0.7000	**0.7759**	0.7734	0.7734
Isolet	0.5385	0.6378	0.6538	0.4487	0.7288	0.7365	**0.7705**
CANE9	0.8454	0.8454	0.8454	0.7056	**0.8583**	**0.8583**	**0.8583**
ORL	0.7352	0.8325	0.8400	0.5475	**0.8675**	0.8350	0.8575
COIL20	0.8653	0.9715	0.9771	0.9847	0.9845	0.9861	**0.9958**
WarpPIE10p	0.9524	0.9810	0.9905	0.9095	0.9857	0.9905	**0.9952**
TOX 171	0.7611	0.8948	0.9000	0.9062	0.9294	**0.9415**	0.9301

Table 7.
Macro-F1 of seven algorithms using SVM on different datasets.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	0.6337	0.6094	0.6076	0.6348	0.6194	0.6337	**0.6381**
Sonar	0.7664	0.7839	0.7683	0.7645	0.7738	0.7805	**0.7862**
Mfeatfac	0.8533	0.8862	0.8932	0.8820	0.8982	0.8996	**0.9015**
Semeion	0.5769	0.6255	0.6455	0.6237	**0.6722**	0.6584	0.6708
Isolet	0.3915	0.5935	0.6047	0.5599	0.6691	0.6812	**0.7069**
CANE9	0.7027	0.7218	0.7210	0.6768	0.7284	0.7267	**0.7335**
ORL	0.3698	0.5268	0.5286	0.3582	**0.5300**	0.5092	0.5228
Coil20	0.6063	0.8430	0.8622	0.8611	0.8897	0.8749	**0.8943**
WarpPIE10p	0.7703	0.8466	0.8542	0.8561	0.8577	0.8591	**0.8698**
TOX 171	0.6404	0.7376	0.7504	0.7436	0.7764	**0.7988**	0.7985

Table 8.
Macro-F1 of seven algorithms using KNN on different datasets.

Dataset	Mim	JMI	mRMR	CIFE	RelaxFS	CFR	WCFR
Vehicle	0.5652	0.6364	0.6139	0.6539	0.6411	0.6532	**0.6616**
Sonar	0.7749	0.8220	0.7950	**0.8308**	0.8215	0.8262	**0.8308**
Mfeatfac	0.7984	0.8543	0.8647	0.8632	0.8759	0.8737	**0.8795**
Semeion	0.5279	0.5830	0.6034	0.5926	0.6328	0.6220	**0.6350**
Isolet	0.3226	0.5046	0.5150	0.3807	0.5849	0.6067	**0.6302**
CANE9	0.6699	0.6884	0.6885	0.6152	0.6995	0.6966	**0.7028**
ORL	0.3350	0.4684	0.4817	0.2874	**0.4847**	0.4552	0.4673
Coil	0.5982	0.8704	0.8821	0.8995	0.9014	0.8999	**0.9072**
WarpPIE10p	0.7340	0.8616	0.8683	0.7750	0.8690	0.8718	**0.8801**
TOX 171	0.6221	0.7699	0.7729	0.8083	0.8340	**0.8428**	0.8409

classification result of WCFR is lower than that of CFR on TOX 171, and is higher than other methods on Vehicle, Mfeatfac, Isolet, CANE9, COIL20 and WarpPE10p.

5. Conclusions and feature work

Mutual information is usually used to measure the relations between the feature and the class. Most of the feature selection methods based on low-order mutual information try to describe the relations more precisely. We introduce a new method to improve the quality of feature subset by using the standard deviations. The new method, WCFR, is an improvement on CFR without increasing the time complexity. And the experiment results show such improvement.

WCFR is more effective than other method, while the improvement doesn't solve the essential issue of feature selection based on mutual information. The feature selection methods mentioned above are all based on low-order mutual information, and this leads to lose a lot of information. In the future, we plan to describe the relations among features with high order mutual information.

References

[1] HongFang Zhou, Yao Zhang, YingJie Zhang, HongJiang Liu, Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy, Appl. Intell. 48 (7) (2018) 883–896.

[2] Abdulhamit Subasi. Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques. 2019, pp. 193-275.

[3] H.F. Zhou, J. Guo, Y.H. Wang, A feature selection approach based on interclass and intraclass relative contributions of terms, Comput. Intell. Neurosci. 2016 (17) (2016) 1–8.

[4] H.F. Zhou, J. Guo, Y.H. Wang, A feature selection approach based on term distributions, SpringerPlus. 5 (1) (2016) 1–14.

[5] S. Das, Filters, Wrappers and Boosting-Based Hybrid for Feature Selection, Proceedings of the International Conference on Machine Learning, 2016 (2016), pp. 74–81.

[6] Qi-Hai Zhu, Yu-Bin Yang, Discriminative embedded unsupervised feature selection, Pattern Recogn. Lett. 112 (1) (2018) 219–225.

[7] L. Jiang, G. Kong, C. Li, Wrapper framework for test-cost-sensitive feature selection, IEEE Trans. Syst., Man, Cybern.: Syst. (2019) 1–10.

[8] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1-2) (1997) 273–324, available at: http://dx.doi.org/10.1016/S0004-3702(97)00043-X.

[9] M. Francisco Macedo, Rosario Oliveira, *et al.*, Theoretical foundations of forward feature selection methods based on mutual information, Neurocomputing 2019 (325) (2019) 67–89.

[10] M.A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, Seventeenth International Conference on Machine Learning, 2000.

[11] C.M. Bishop, Pattern Recognition and Machine learning, Springer, 2006.

[12] N.X. Vinh, S. Zhou, J. Chan, J. Bailey, Can high-order dependencies improve mutual information based feature selection?, Pattern Recognition. 53 (C) (2015) 46–58.

[13] D.D. Lewis. Feature selection and feature extraction for text categorization, in: Proceedings of The Workshop on Speech and Natural Language, Association for Computation Linguistics Morristown, Nj, USA, pp 212-217.

[14] R.Battiti. Using mutual information for selecting features in supervised neural net learning, IEEE Transaction on Neural Networks, 5(4): 537-550.

[15] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Machine Intell. 27 (8) (2005) 1226–1238, available at: http://dx.doi.org/10.1109/TPAMI.2005.159.

[16] L.T. Vinh, S. Lee, A novel selection method based on normalized mutual information, Appl. Intell. 37 (1) (2011) 100–120.

[17] D. Lin, X. Tang. Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion. In: European Conference on computer version. pp 68-82.

[18] H.H.Yang, J.Moody, Feature Selection Based on Joint Mutual Information, in: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis, 1999, pp. 22–25.

[19] Wangfu Gao, Hu. Liang, Ping Zhang, Jialong He, Feature selection considerding the composition of feature relevancy, Pattern Recogn. Lett. 112 (2018) 70–74.

[20] available at: http://archive.ics.uci.edu/ml.

[21] Janez Demišar, D. Schuurmans, Statistical Comparisons of Classifiers over Multiple Data Sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30.

Corresponding author
Hongfang Zhou can be contacted at: zhouhf@xaut.edu.cn