Applied Intelligence (2022) 52:5457–5474
https://doi.org/10.1007/s10489-021-02524-x

Feature selection based on mutual information with correlation coefficient

Hongfang Zhou[1] · Xiqian Wang[1] · Rourou Zhu[1]

Accepted: 8 May 2021 / Published online: 12 August 2021

Abstract

Feature selection is an important preprocessing process in machine learning. It selects the crucial features by removing irrelevant features or redundant features from the original feature set. Most of feature selection algorithms focus on maximizing relevant information and minimizing redundant information. In order to remove more redundant information in the evaluation criteria, we propose a feature selection based on mutual information with correlation coefficient (CCMI) in this paper. We introduce the correlation coefficient in the paper, and combine the correlation coefficient and mutual information to measure the relationship between different features. We use the absolute value of the correlation coefficient between two different features as the weight of the redundant item denoted by the mutual information in the evaluation standard. In order to select low redundancy features effectively, we also use the principle of minimization in the evaluation criteria. By comparing with 7 popular contrast algorithms in 12 data sets, CCMI has achieved the highest average classification accuracy for two classifiers of SVM and KNN. Experimental results show that our proposed CCMI has better feature classification capability.

Keywords Feature selection · Mutual information · Correlation coefficient · Filter method

1 Introduction

Feature selection is an important work in data mining and pattern recognition, and it is one of the mainstream technologies for processing high-dimensional data. Too many features may cause dimension disaster in machine learning, so it is required to reduce the dimension number in feature spaces. Usually, there are two kinds of dimension reduction methods which are feature selection and feature extraction [1, 2]. Feature extraction uses mathematical methods to fuse some features to generate new features, which only have mathematical meaning. However, feature selection does not produce new features, it only evaluates features through evaluation functions, and selects important features. The feature subsets finally obtained from feature selection has practical significance. The important features mentioned above are also called as the relevant features, while the removed features include irrelevant features and redundant features. Irrelevant features refer to features irrelevant to the tasks, while redundant features refer to those features that contain redundant or unnecessary information.

According to the relationship with classifiers, feature selection methods can be divided into three categories which are filter methods [3, 4], wrapper methods [5–7] and embedded methods [8].

Wrapper feature selection depends on the performance of the classifier. This method directly uses the classification algorithm to evaluate the feature subset obtained by feature selection. The purpose of this method is to select the feature subset that is the most beneficial to the performance of a given classifier. Because the wrapper feature selection method is directly optimized for a given classifier, the performance of the wrapper feature selection method is better than that of the filter, but the computational cost of the wrapper is usually much higher than that of the filter.

✉ Hongfang Zhou
zhouhf@xaut.edu.cn

Xiqian Wang
Sophie_xq47139@163.com

Rourou Zhu
r145125@163.com

[1] School of Computer Science and Engineering, Xi'an University of Technology, NO.5 South Jinhua Road, Xi'an, Shaanxi, China

Embedded feature selection is also depending on the classifier. It integrates the feature selection process with the learner training process, both of which are completed in the same optimization process, i.e. feature selection is automatically performed in the classifier training process.

Filter feature selection is independent of the specific classifier. The method firstly selects the features of the data set, and then trains the classifier. Feature selection has nothing to do with the subsequent classifier. It is equivalent to "filter" the initial features with the feature selection process first, and then training the model with the filtered features. The calculation cost of this method is relatively low [9].

This paper uses filter feature selection method. Mutual information is a concept in information theory. It is a measure of the amount of information that a random variable contains another random variable. Mutual information can also be described as the reduction of the uncertainty of the original random variable given the knowledge of another random variable. This paper uses the concept of mutual information to measure the relationship between features and classes, features and features.

Feature selection based on mutual information is to select a feature subset of m features from the original data set X with M features, and this subset has the largest mutual information value with the class C , that is,

$$I(S;C) = \sum_{X_1,...,X_m,C} P(X_1,...,X_m,C)log \times \frac{P(X_1,...,X_m,C)}{P(X_1,...,X_m,C)P(C)} \tag{1}$$

Here, S is the finally selected feature subset and C is the class label. However, for high-dimensional joint mutual information, it is difficult to directly obtain the value of mutual information. The estimation of high-dimensional probability distribution has always been a challenge in statistics. Therefore, in many researches on feature selection [10, 11], some independent hypotheses are proposed to solve the problem of high-dimensional joint probability that is difficult to calculate, and the calculation of high-dimensional mutual information is converted to the sum of low-dimensional mutual information. Because these assumptions require some prerequisites for the data set, when some data sets do not meet these requirements, the final result of feature selection will be inaccurate. The main goal of this paper is to find an evaluation criterion, and use it to get a feature subset that can improve the classification accuracy. The target of feature selection is to select features that are highly correlated with the class label in the data set, and delete irrelevant features and redundant features. The weight relationship between relevant items and redundant items in the evaluation criteria will affect the quality of the final feature subset.

We introduce the correlation coefficient and combine the correlation coefficient and mutual information to measure the relationship between features in the paper. The correlation coefficient is used to study the degree of linear correlation between variables. The absolute value of the correlation coefficient is less than or equal to 1. The closer the absolute value of the correlation coefficient is to 1, the stronger the degree of linear correlation between the two variables is. Therefore, we use the absolute value of the correlation coefficient between features to measure the importance of redundant items. We use the absolute value of the correlation coefficient as the weight of the redundant item. We also use the principle of minimization in the evaluation criteria to select the feature subset that contains more classification information, thereby improving classification accuracy.

The rest of this article is organized as follows. Related concepts in information theory will be introduced in Section 2. In Section 3, some classic feature selection methods are introduced. In Section 4, our proposed CCMI is described in details. In Section 5, the performance of CCMI is tested on some popular data sets. We summarize the paper in Section 6.

2 Information theory

2.1 Related concepts

Definition 2.1 (Entropy) Entropy is a measure of uncertainty of a random variable. The higher the entropy is, the higher the uncertainty of random variable is. The entropy of a discrete random variable is defined as

$$H(X) = -\sum_{x\in X} p(x)log(p(x)) \tag{2}$$

Where X represents a random variable, $p(x)$ is a probability density function of X.

Definition 2.2 (Joint entropy) Joint entropy is a measure of the uncertainty of a joint distributed random system. For a pair of discrete random variables (X, Y) subject to joint distribution $p(x, y)$, the joint entropy $H(X, Y)$ is defined as

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)log(p(x,y)) \tag{3}$$

Where $p(x, y)$ is the joint probability distribution.

Definition 2.3 (Conditional entropy) If $(X, Y) \sim p(x, y)$, the entropy $H(Y|X)$ of Y under the given condition X is defined as

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)log(p(y|x)) \tag{4}$$

Where $p(x, y)$ is the joint probability distribution and $p(y|x)$ is the conditional probability distribution. Conditional entropy $H(Y|X)$ represents the uncertainty of random variables Y under the condition of known random variable X.

Definition 2.4 (Relative entropy) A measure of the distance between two random distributions. The relative entropy between $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)} \tag{5}$$

Definition 2.5 (Mutual Information) Mutual Information is used to measure the amount of information that a random variable contains in another random variable. Considering two random variables X and Y, their joint probability density function is $p(x, y)$, and their marginal probability density functions are $p(x)$ and $p(y)$ respectively. Mutual information $I(X; Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$, i.e.

$$\begin{aligned} I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \\ &= D(p(x, y)||p(x)p(y)) \end{aligned} \tag{6}$$

Definition 2.6 (Conditional mutual information) Conditional mutual information of random variables X and Y given random variable Z is defined as

$$\begin{aligned} I(X; Y|Z) &= \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p(x, y, z) log \frac{p(z)p(x, y, z)}{p(x, z)p(y, z)} \\ &= H(X|Z) - H(X|Y, Z) \end{aligned} \tag{7}$$

The relationship among entropy $H(X)$, conditional entropy $H(Y|X)$ and joint entropy $H(X, Y)$ is as follows.

$$H(Y|X) = H(X, Y) - H(X) \tag{8}$$

The relationship between mutual information $I(X; Y)$ and entropy is as follows.

$$I(X; Y) = H(X) - H(X|Y) \tag{9}$$

Figure 1 describes the relationship between mutual information and entropy. This relationship shows that mutual information represents the reduction of uncertainty of the original random variable given the knowledge of another random variable.

2.2 Relevance and redundancy

Feature selection based on mutual information is to select features in the original data set. The goal is to select features that are highly related to the class and have low redundancy between the selected features.

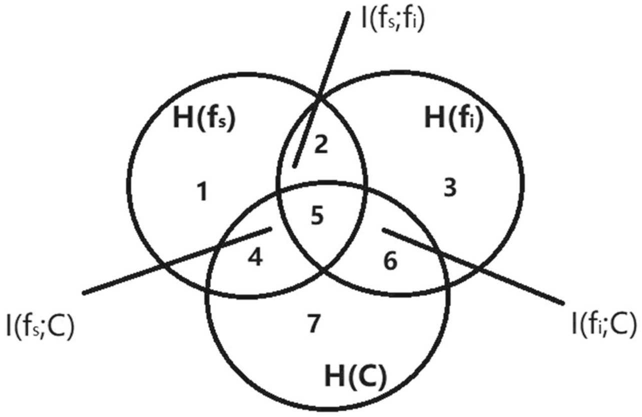

Fig. 1 Relationship between mutual information and entropy

Relevancy is measured by mutual information between the class label C and the candidate feature X_m, i.e. $I(C; X_m)$. If $I(C; X_m) = 0$, it indicates that the candidate feature X_m and the class C are independent to each other, and no classification information is provided. If $I(C; X_m) > 0$, it indicates that the candidate feature X_m is related to the class C, the larger the mutual information value is, the more information the candidate feature X_m can provide for classification.

Redundancy is measured by calculating mutual information between the candidate feature X_m and the selected feature X_s, i.e. $I(X_m; X_s)$. The larger the value of $I(X_m; X_s)$, the more redundant information in the classification information provided by this candidate feature. Therefore, a candidate feature with a small $I(X_m; X_s)$ value should be selected.

In the paper, $I(C; X_m|X_s)$ is used as a measure of conditional relevance. The following is the relationship between conditional mutual information and entropy

$$I(X_m; C|X_s) = H(C|X_s) - H(C|X_s, X_m) \tag{10}$$

Assuming $I(C; X_m) > 0$, when the value of the redundant term $I(X_m; X_s)$ is larger, the value of $I(C; X_m|X_s)$ will approach 0. At this time, the candidate feature X_m is the same as the classification information provided by the selected feature X_s. When the redundant item $I(C; X_m) = 0$, the value of $I(C; X_m|X_s)$ will be larger. At this time, the candidate feature X_m is different from the classification information provided by the selected feature X_s. In feature selection, our goal is to find candidate features with a large $I(C; X_m|X_s)$ value. It can also be seen from the above analysis that $I(C; X_m|X_s)$ has both relevant information and redundant information. Therefore, compared with $I(C; X_m)$, $I(C; X_m|X_s)$ describes the correlation between the candidate feature and the class label more accurately.

3 Related work

Feature selection is an important part of the pattern classification system. The feature selection algorithm based on mutual information is to select important features from the original data set, that is, to remove irrelevant features and redundant features. The feature selection algorithm based on mutual information is a filtering method. Currently, many classic feature selection algorithms based on mutual information have been proposed.

MIM algorithm [12] is the earliest feature selection method based on mutual information, which uses mutual information to evaluate the degree of correlation between features and the class label. The evaluation criteria of the algorithm is as follows.

$$J(X_m) = I(X_m; C) \tag{11}$$

Where X_m represents the candidate feature and C represents the class label. MIM algorithm only considers the degree of correlation between features and the class label, and can filter out irrelevant features, but does not consider the relationship between candidate features and the selected feature subset. If there are a lot of redundant features in the data set, the classification accuracy of this algorithm will be lower.

Mutual Information based on Feature Selection (MIFS) [13] proposed by Battiti considers the existence of redundant information between candidate features and selected features. The evaluation criteria of this algorithm is as follows.

$$J(X_m) = I(X_m; C) - \beta \sum_{X_s \in S} I(X_m; X_s) \tag{12}$$

Where S represents the selected feature set, X_s is the feature in the selected feature set S , and β is the balance factor between the relevant item and the redundant item, with a value between 0 and 1.

Minimal-redundancy-maximal-relevance (mRMR) algorithm [14] is a variant of MIFS algorithm. The evaluation criteria of the algorithm is as follows.

$$J(X_m) = I(X_m; C) - \frac{1}{\mid S \mid} \sum_{X_s \in S} I(X_m; X_s) \tag{13}$$

Where $|S|$ represents the number of features in the selected feature set S. It can be seen that the improvement of the MIFS algorithm by the mRMR algorithm is to set the value of β to $\frac{1}{|S|}$. It avoids the problem that the selection of redundant features becomes more and more difficult as the number of selected features increases.

Conditional Informative Feature Extraction (CIFE) [15] is proposed by Lin and Tang. The evaluation criterion of the algorithm is as follows.

$$J(X_m) = I(X_m; C) - \sum_{X_s \in S} [I(X_m; X_s) - I(X_m; X_s|C)] \tag{14}$$

Where $I(X_m; X_s|C)$ represents the degree of redundancy between the candidate feature X_m and the selected feature X_s given the class label C. CIFE algorithm uses intra-class redundancy $I(X_m; X_s) - I(X_m; X_s|C)$, which can also be written as $I(X_m; X_s; C)$. Using intra-class redundancy to calculate redundant information is more accurate.

H.Y.Yang proposed Joint Mutual Information (JMI) [16] to measure the importance of each feature by using the joint mutual information of candidate features, selected features and the class label. The specific evaluation criteria of the algorithm is as follows.

$$J(X_m) = I(X_m; C) - \frac{1}{\mid S \mid} \sum_{X_s \in S} [I(X_m; C) - I(X_m; C|X_s)] \tag{15}$$

Among them, $I(X_m; C) - I(X_m; C|X_s)$ is equal to $I(X_m; X_s; C)$, that is

$$\begin{aligned} I(X_m; X_s; C) &= I(X_m; C) - I(X_m; C|X_s) \\ &= I(X_m; X_s) - I(X_m; X_s|C) \end{aligned} \tag{16}$$

It can be seen that JMI algorithm is a modification of CIFE algorithm. JMI algorithm multiplies the redundancy item in CIFE algorithm with $\frac{1}{|S|}$ to obtain the redundancy item. JMI uses the average to reflect the central tendency of redundant items.

RelaxFS algorithm [10] was proposed by Vinh* and Zhou. This method proposes a new method of measuring redundancy that can contain more redundant information. The specific evaluation criteria of the algorithm is as follows.

$$\begin{aligned} J(X_m) = {} & I(X_m; C) - \frac{1}{\mid S \mid} \sum_{X_j \in S} I(X_m; X_j) \\ & + \frac{1}{\mid S \mid} \sum_{X_j \in S} I(X_m; X_j|C) \\ & - \frac{1}{\mid S \mid\mid S - 1 \mid} \sum_{X_j \in S} \sum_{\substack{X_i \in S \\ j \neq i}} I(X_m; X_i|X_j) \end{aligned} \tag{17}$$

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

$\sum_{X_j \in S} \sum_{X_i \in S j \neq i} I(X_m; X_i | X_j)$ contains more redundant information than $\frac{1}{|S|} \sum_{X_j \in S} I(X_m; X_j) - \frac{1}{|S|} \sum_{X_j \in S} I(X_m; X_j | C)$, which makes the measurement of redundancy between features more accurate.

Max-Relevance and Max-Independence (MRI) algorithm [17] proposed the concept of Independent Classfication Information (ICI). ICI combines the new classification information with the saved classification information, emphasizing the difference in the classification ability of each feature. Therefore, MRI will select features with low redundancy while tending to select features with differences. The evaluation criteria of the algorithm is as follows.

$$J(X_m) = I(X_m; C) + \sum_{X_s \in S} ICI(C; X_s, X_m) = I(X_m; C) + \sum_{X_s \in S} [I(C; X_m | X_s) + I(C; X_s | X_m) \quad (18)$$

Where $I(C; X_m | X_s)$ represents the new classification information provided by X_m to the selected feature set S. In $I(C; X_s | X_m) + I(C; X_s; X_m) = I(C; X_s)$, $I(C; X_s)$ is a constant. When $I(C; X_s | X_m)$ takes the largest value, $I(C; X_s; X_m)$ takes the smallest value. $I(C; X_s | X_m)$ is equivalent to a redundant term in ICI. Therefore, ICI contains both new classification information and redundant information.

Wanfu Gao et al. proposed Composition of Feature Relevancy(CFR) algorithm [18]. They conclude that feature relevancy consists of two parts: new classification information and redundant information.The target of feature selection methods is to maximize the new classification information while minimizing redundant information. So the evaluation criteria of CFR is as follows.

$$J(X_m) = \sum_{X_s \in S} I(C; X_m | X_s) - I(C; X_m; X_s) \quad (19)$$

Where $I(C; X_m | X_s)$ is the new classification information X_m provided to the selected feature set S. $I(C; X_m; X_s)$ indicates the redundant information of X_m, C and X_s.

4 Proposed method

4.1 Problem and Method

The goal of feature selection based on mutual information is to select a feature subset that maximizes the value of (1). However, it is difficult to directly calculate (1). Therefore, many studies have turned the goal to use low-dimensional mutual information to estimate high-dimensional mutual information, and finally get the value of (1). In these studies, Brown et al. proposed a framework [19] to realize the estimation of (1), which can be suitable for many feature selection algorithms based on mutual information that have been proposed. The specific framework is as follows.

$$J(X_m) = I(X_m; C) - \beta \sum_{X_s \in S} I(X_m; X_s) + \gamma \sum_{X_s \in S} I(X_m; X_s | C) \quad (20)$$

Where β is the balance factor of redundant item between the candidate feature X_m and the selected feature X_s, and γ is the balance factor of redundant item between the candidate feature X_m and the selected feature X_s given the class label C.

From the above framework, it can be seen that the value of β and γ as balance factors will affect the quality of the feature subset finally selected. Therefore, the focus of our research is how to better adjust the importance of the relevant item and the redundant item.

In order to more accurately describe the degree of correlation between features, we introduce the concept of correlation coefficient. The correlation coefficient ρ_{xy} is used to study the degree of linear correlation between variables. The specific calculation is as follows.

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{D(X)} \bullet \sqrt{D(Y)}} \quad (21)$$

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \quad (22)$$

Where $Cov(X, Y)$ is the covariance of random variables X and Y, $D(X)$ is the variance of random variable X, $D(Y)$ is the variance of random variable Y, and E is the mathematical expectation of the random variable. The absolute value of the correlation coefficient is less than or equal to 1. When the correlation coefficient value is equal to 0, it means that there is no correlation between the two variables. The closer the absolute value of the correlation coefficient is to 0, the weaker the linear correlation between the two variables. The closer the absolute value of the correlation coefficient is to 1, the stronger the degree of linear correlation between the two variables. When it is equal to 1, the linear correlation must exist. We use the absolute value of the correlation coefficient to judge the degree of correlation between the features. The larger the absolute value, the greater the redundancy between the features. At this time, the redundant items in the corresponding evaluation criteria are more important. Therefore, we use the absolute value of the correlation coefficient between features to measure the importance of redundant items. We use the absolute value of the correlation coefficient as the weight of the redundant item, so as to achieve the purpose of removing more redundant information.

The goal of feature selection based on mutual information is to make the selected feature subset highly correlated with the class label. According to the chain rule,

$$\begin{aligned} I(S;C) &= I(S_{m-1} \cup X_m; C) \\ &= I(S_{m-1};C) + I(X_m;C|S_{m-1}) \end{aligned} \tag{23}$$

Where S_{m-1} represents the feature subset excluding the candidate feature X_m. In the above formula, $I(S_{m-1};C)$ has nothing to do with X_m and can be regarded as a constant. Therefore, the following formula can be obtained,

$$\begin{aligned} \arg\max \mathrm{I(S;C)} &= \arg\max_{X_m \in F \setminus S} \mathrm{I(S_{m-1} \cup X_m; C)} \\ &\equiv \arg\max_{X_m \in F \setminus S} \mathrm{I(X_m; C|S_{m-1})} \end{aligned} \tag{24}$$

Therefore, the goal of feature selection algorithm based on mutual information is to maximize $I(C;X_m|S)$. In Section 2, we also analyzed that the conditional correlation term $I(C;X_m|X_s)$ is more accurate than the correlation term $I(C;X_m)$ in describing the correlation between candidate features and the class label.

Among the existing feature selection algorithms, CMIM algorithm and JMI algorithm also used conditional relevance. The specific evaluation criterias are

$$J_{CMIM}(X_m) = \underset{X_s \in S}{min} I(X_m;C|X_s) \tag{25}$$

$$J_{JMI}(X_m) = \frac{1}{|S|} \sum_{X_s \in S} I(X_m;C|X_s) \tag{26}$$

It can be seen that CMIM algorithm takes the minimum value of the conditional relevant term, that is, the principle of minimization is used. However, the JMI algorithm focuses on the average situation, that is, using the principle of averaging.

The principle of minimization [20] emphasizes the individual's influence on the whole. This principle can select the feature subset with lower redundancy and effectively improve the classification accuracy. Therefore, we use the principle of minimization for the proposed algorithm. The following is our proposed feature selection based on mutual information with correlation coefficient (CCMI) algorithm.

$$\begin{aligned} J_{CCMI}(X_m) = &\underset{X_s \in S}{min} I(X_m;C|X_s) \\ &- \underset{X_s \in S}{min} |\rho_{X_m X_s}| \cdot I(X_m;X_s) \end{aligned} \tag{27}$$

$$\rho_{X_m X_s} = \frac{Cov(X_m, X_s)}{\sqrt{D(X_m)} \bullet \sqrt{D(X_s)}} \tag{28}$$

Where $\rho_{X_m X_s}$ represents the correlation coefficient between the candidate feature X_m and the selected feature X_s, $Cov(X_m, X_s)$ represents the covariance between the candidate feature X_m and the selected feature X_s, $D(X_m)$ is the variance of X_m, and $D(X_s)$ is the variance of X_s.

The implementation steps of CCMI algorithm are shown in the pseudo code.

Algorithm 1 CCMI: Feature selection based on mutual information with correlation coefficient.

```
Step1:
Input: A training sample D with an entire feature set
       $F = \{X_1, X_2, ..., X_n\}$ , the class $C$ and the
       number of selected features $K$.
Output: The set $S$ with the selected features
$S = \emptyset$
For i=1 to n
    $relevance(F(i)) \leftarrow I(F(i);C)$
End for
$f_{new} \leftarrow f_m$ $satisfying$ $f_m$=
arg max$_{f_m \in F}$ relevance(f$_m$)
$S$=$S \cup f_{new}$
$F$=$F - f_{new}$
count=1
while $count < K$ do
    r=n-count
    For m=1 to r
        For i=1 to count
        Calculate condition mutual information
    $I(X_m;C|X_i)$
        Calculate correlation coefficient $\rho_{X_m X_i}$
    according to Eq.(28)
        $\rho_{X_m X_i} = \frac{Cov(X_m,X_i)}{\sqrt{D(X_m)}\sqrt{D(X_i)}}$
        Calculate mutual information $I(X_m;X_i)$
        End for
        Calculate $J_{CCMI}(X_m)$ according to Eq.(27)
        $J_{CCMI}(X_m) =$
    $\underset{X_s \in S}{min}\, I(X_m;C|X_s) - \underset{X_s \in S}{min}\, |\rho_{X_m X_s}| \cdot I(X_m;X_s)$
    End for
    $f_{new} \leftarrow f_m$ $satisfying$ $f_m$=
    arg max$_{f_m \in F}$ J$_{\mathrm{CCMI}}$(f$_m$)
    $S$=$S \cup f_{new}$
    $F$=$F - f_{new}$
    count=count+1
end
```

4.2 Complexity analysis

Suppose the number of features to be selected is k, N is the total number of features, and M represents the number of samples in the data set. The time complexity of mutual information, conditional mutual information and joint mutual information is $O(M)$, because all samples need to be examined for probability estimation. As a result, the time complexity of MIM is $O(MN)$ and RelaxFS is $O(k^3MN)$. CCMI algorithm includes a while loop, a for loop in the while loop, and a for loop nested in the for loop. From the pseudo code, we can analyze the time complexity of CCMI algorithm is $O(k^2MN)$. CFR, CIFE, JMI, mRMR and MRI algorithms have the same time complexity as CCMI, which is higher than MIM algorithm and lower than RelaxFS algorithm.

5 Experiments

5.1 Data Sets

In order to verify the efficiency of our proposed algorithm, we select 7 classic mutual information-based feature selection algorithms and 12 commonly used and classic data sets [21]. By training these data sets on these algorithms, we discover the superiority of our proposed algorithm. These 12 data sets belong to different data types. There are image data sets, such as USPS, COIL20, WarpPIE10P data sets. Digital data sets include Wine, Mfeatfac, and Semeion data sets. There are also some text data sets, such as CANE9, BASEHOCK data sets, and sound data sets, such as Isolet data set. The specific information about these data sets is listed in Table 1.

In the experiment, we need to do some preprocessing on these original data sets. First, standardize the data set so that the value range of each feature is compressed to 0 to 1. Then, the data set is discretized, and all the feature values in the data set are divided into five levels [9]. Finally, apply the obtained data set to different algorithms for feature selection

5.2 Experimental setup

For a set F with n features, it has $2^n - 1$ subsets. If you evaluate each subset, and then select the best feature subset, the amount of calculation is quite large. Therefore, in this paper, we use forward iterative search. Forward iterative search refers to starting from an empty set, each time using the evaluation function to select a feature from the original feature set, put it into the selected feature set, and delete this feature from the original feature set. After m times, a feature subset of size m can be obtained. If the number of features in the original data set is less than 50, the value of m is the number of features in the data set. Otherwise, the value of m is set to 50.

Table 1 Details of data sets

Data sets	#of features	#of instances	#of classes
Wine	13	178	3
LCX	55	303	2
Mfeatfac	216	2000	10
Semeion	256	1593	10
USPS	256	9298	10
Isolet	617	1560	26
CANE9	856	1080	9
COIL20	1024	1440	20
WarpPIE10P	2420	210	10
RELATHE	4322	1427	2
BASEHOCK	4862	1993	2
Orlraws10P	10304	100	10

In this paper, we use SVM classifier and KNN classifier to train the data set. At present, there are many classifiers derived from SVM [22–26], but SVM classifier is still the most commonly used and classic classifier in feature selection. SVM classifier [9] is a sparse kernel-based method. SVM maps data from a low-dimensional space to a high-dimensional space, and converts the original non-linearly separable problem in the low-dimensional space into a linearly separable problem in the feature space, thereby achieving classification. The linear kernel is used in this paper. The KNN classifier [9] is a representative of lazy learning, without an explicit learning process. The classifier calculates the distance between the test instance and each instance point in the training set according to the selected distance measurement (such as Euclidean distance, Manhattan distance), selects k nearest neighbors according to the k value, and finally classifies the test instance according to the classification decision rule . The value of k in this paper is set to 5.

In this paper, we use the average classification accuracy, K-S test and F1 value to evaluate the feature subset. To ensure the accuracy of the experiment, we use “10 times 10-fold cross-validation”. A 10-fold cross-validation is performed on each data set, and 10-fold cross-validation is repeated 10 times. The final classification accuracy is the average of 10 10-fold cross-validation. Kolmogorov-Smirnov test [27] is a non-parametric test method, it does not need to know the data distribution. We use the default significance level of the K-S test in our experiments, which is 5%. If the P value is less than 5%, the two algorithms are considered to have a significant difference, and if the P value is greater than 5%, there is no significant difference.

The value of F1 [9] is the harmonic average of precision rate and recall rate. It gives the same weight to precision rate and recall rate, and this is suitable for two classification problems. The specific calculation formula is

$$F_1 = \frac{2 * P * R}{P + R} \tag{29}$$

Where P is precision rate and R is recall rate. For multi-classification problem, the macro F1 value will be calculated through the macro precision rate and the macro recall rate

$$Macro\_P = \frac{1}{t}\sum_{i=1}^{t} P_i \tag{30}$$

Table 2 Classification Accuracy(mean±std.) using SVM

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	91.26±0.05(=)	85.59±0.03(+)	91.22±0.05(=)	91.01±0.05(=)	91.36±0.05(=)	**91.47±0.05(=)**	91.41±0.05(=)	91.41±0.05
CANE9	73.06±0.17(+)	67.18±0.13(+)	72.66±0.17(=)	71.39±0.18(+)	73.08±0.17(=)	72.68±0.17(=)	73.41±0.17(=)	**75.55±0.17**
COIL20	88.19±0.15(+)	87.25±0.14(+)	85.41±0.14(+)	63.74±0.19(+)	88.62±0.15(+)	87.47±0.14(+)	89.86±0.14(=)	**90.17±0.15**
Isolet	70.32±0.15(+)	57.58±0.10(+)	61.48±0.13(+)	41.56±0.13(+)	69.02±0.15(+)	62.54±0.13(+)	68.82±0.15(+)	**74.78±0.17**
LCX	66.93±0.02(=)	66.00±0.02(+)	66.75±0.02(=)	67.10±0.02(=)	66.41±0.02(=)	66.70±0.02(+)	66.27±0.02(+)	**67.25±0.02**
Mfeatfac	90.37±0.11(=)	88.61±0.11(+)	89.06±0.11(+)	85.51±0.12(+)	90.35±0.11(=)	89.69±0.11(+)	90.47±0.11(=)	**90.74±0.11**
Orlraws10P	88.18±0.11(+)	72.96±0.08(+)	95.08±0.12(+)	75.82±0.16(+)	94.42±0.12(+)	94.38±0.11(+)	95.54±0.12(+)	**96.48±0.12**
RELATHE	78.55±0.03(+)	78.88±0.03(+)	77.29±0.03(+)	76.28±0.03(+)	78.50±0.03(+)	78.80±0.04(+)	79.23±0.04(+)	**81.02±0.05**
Semeion	67.18±0.11(+)	63.84±0.10(+)	64.35±0.12(+)	59.61±0.15(+)	66.62±0.12(+)	65.97±0.12(+)	68.49±0.12(+)	**75.20±0.14**
USPS	83.36±0.10(+)	83.25±0.10(+)	82.56±0.10(+)	63.67±0.11(+)	83.51±0.10(+)	84.47±0.11(+)	85.26±0.10(+)	**87.04±0.11**
WarpPIE10P	93.72±0.12(+)	92.68±0.12(+)	92.50±0.13(+)	84.38±0.16(+)	92.69±0.13(+)	93.08±0.13(+)	93.74±0.12(+)	**94.50±0.13**
Wine	92.26±0.07(=)	91.28±0.06(=)	92.10±0.06(=)	**92.66±0.07(=)**	92.23±0.07(=)	92.49±0.07(=)	92.45±0.07(=)	92.24±0.07
Average	81.95±0.10	77.93±0.09	80.87±0.10	72.73±0.11	82.23±0.10	81.65±0.10	82.91±0.10	**84.70±0.11**
W/T/L	8/4/0	11/1/0	8/4/0	9/3/0	7/5/0	9/3/0	7/5/0/	

$$Macro\_R = \frac{1}{t}\sum_{i=1}^{t} R_i \tag{31}$$

$$Macro\_F1 = \frac{2 * Macro\_P * Macro\_R}{Macro\_P + Macro\_R} \tag{32}$$

Where $Macro\_P$ is the macro precision rate, $Macro\_R$ is the macro recall rate, P_i is the precision of the ith category, R_i is the recall rate of the ith category and t is used to represent the number of the categories.

5.3 Experimental results and discussion

5.3.1 Classification accuracy

In order to prove the classification performance of CCMI algorithm proposed in this paper, we compare the proposed CCMI algorithm with the existing CFR, CIFE, JMI, MIM, MRI, mRMR and RelaxFS algorithms. We use SVM and KNN classifiers to train the CCMI algorithm and seven

Table 3 Classification Accuracy(mean±std.) using KNN(k=5)

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	87.86±0.10(=)	83.04±0.07(+)	87.32±0.09(+)	86.93±0.09(+)	87.82±0.10(=)	87.40±0.09(=)	87.51±0.09(+)	**88.06±0.10**
CANE9	71.68±0.16(+)	64.12±0.11(+)	71.06±0.16(+)	69.64±0.18(+)	71.62±0.16(+)	71.19±0.16(+)	71.81±0.16(+)	**74.46±0.17**
COIL20	89.00±0.17(+)	88.95±0.17(+)	86.71±0.17(+)	61.86±0.23(+)	89.17±0.17(+)	87.53±0.17(+)	89.71±0.18(+)	**90.96±0.18**
Isolet	64.55±0.16(+)	43.80±0.08(+)	54.43±0.12(+)	36.56±0.14(+)	63.05±0.15(+)	56.02±0.13(+)	62.86±0.15(+)	**67.20±0.17**
LCX	64.26±0.03(=)	63.52±0.03(=)	**64.51±0.03(=)**	63.64±0.02(=)	63.88±0.03(=)	61.50±0.02(+)	63.68±0.03(=)	63.89±0.03
Mfeatfac	88.53±0.16(+)	87.66±0.17(+)	86.68±0.16(+)	81.85±0.18(+)	88.31±0.17(+)	87.67±0.17(+)	88.85±0.17(=)	**89.18±0.16**
Orlraws10P	78.48±0.09(+)	58.18±0.05(+)	93.80±0.10(-)	72.38±0.14(+)	90.18±0.10(+)	92.84±0.09(=)	**94.40±0.10(-)**	92.52±0.10
RELATHE	67.92±0.07(+)	65.72±0.10(+)	67.22±0.07(+)	65.46±0.07(+)	68.12±0.07(+)	69.48±0.06(+)	69.87±0.06(+)	**74.58±0.09**
Semeion	64.61±0.13(+)	61.95±0.12(+)	61.35±0.14(+)	56.33±0.17(+)	64.41±0.13(+)	63.25±0.13(+)	65.92±0.13(+)	**72.15±0.16**
USPS	83.97±0.13(+)	84.26±0.13(+)	82.98±0.13(+)	62.29±0.13(+)	84.22±0.13(+)	85.23±0.14(+)	86.11±0.13(+)	**88.15±0.14**
WarpPIE10P	93.06±0.13(+)	82.90±0.11(+)	91.54±0.13(+)	81.23±0.16(+)	92.34±0.13(+)	92.80±0.14(+)	92.98±0.14(=)	**93.21±0.14**
Wine	**93.02±0.09(=)**	90.65±0.08(+)	92.37±0.08(=)	91.42±0.08(=)	92.97±0.09(=)	92.54±0.08(=)	92.50±0.08(=)	92.45±0.08
Average	78.91±0.12	72.90±0.10	78.33±0.12	69.13±0.13	79.67±0.12	78.95±0.12	80.52±0.12	**82.23±0.13**
W/T/L	9/3/0	11/1/0	9/2/1	10/2/0	9/3/0	9/3/0	7/4/1/	

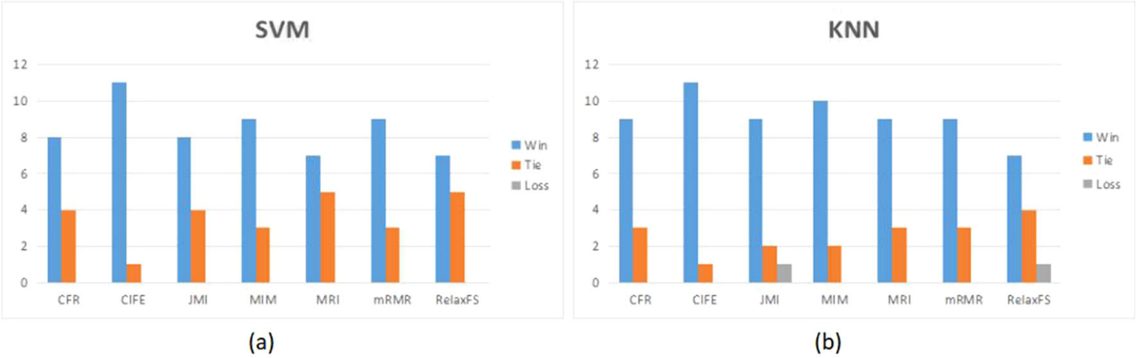

Fig. 2 Performance comparison using K-S test

other feature selection algorithms, and finally get the classification accuracy. We use classification accuracy to measure the classification performance of different feature selection algorithms.

Tables 2 and 3 record the average classification accuracy and standard deviation of CCMI algorithm and other 7 feature selection algorithms on 12 data sets. The classifier in Table 2 is SVM, and the classifier in Table 3 is KNN. In each row, we mark the maximum average classification accuracy in bold. The penultimate line is called “Average” is the average classification accuracy of each algorithm on all data sets. We use the K-S test to judge whether the classification performance of the newly proposed CCMI algorithm is significantly different from other algorithms in the average classification accuracy. We use “+”,“=” and “-”to indicate that the proposed CCMI algorithm is “superior”, “equal” and “inferior” to other feature selection methods. The last row “W/T/L” in Tables 2 and 3 indicates that compared with other methods, the statistical results are CCMI win/tie/loss. The statistical results are summarized in Fig. 2.

From the penultimate row in Tables 2 and 3, it can be seen that the average classification accuracy of CCMI is the highest on both the SVM classifier and the KNN classifier, which are 84.70% and 82.23%, respectively. It shows that in the overall situation, the proposed algorithm CCMI has better classification performance than other algorithms. The average classification accuracy of RelaxFS algorithm on SVM classifier and KNN classifier is second only to CCMI algorithm.

In Table 2, in addition to the BASEHOCK and Wine data sets, CCMI algorithm has achieved the highest classification performance on other data sets. Among them, on the data set Semeion, CCMI is 6.71% higher than RelaxFS, which has the second highest accuracy rate. On the Isolet data set, the accuracy of CCMI is 5.76% higher than that of MRI and 5.96% higher than RelaxFS. On the CANE9 data set, the

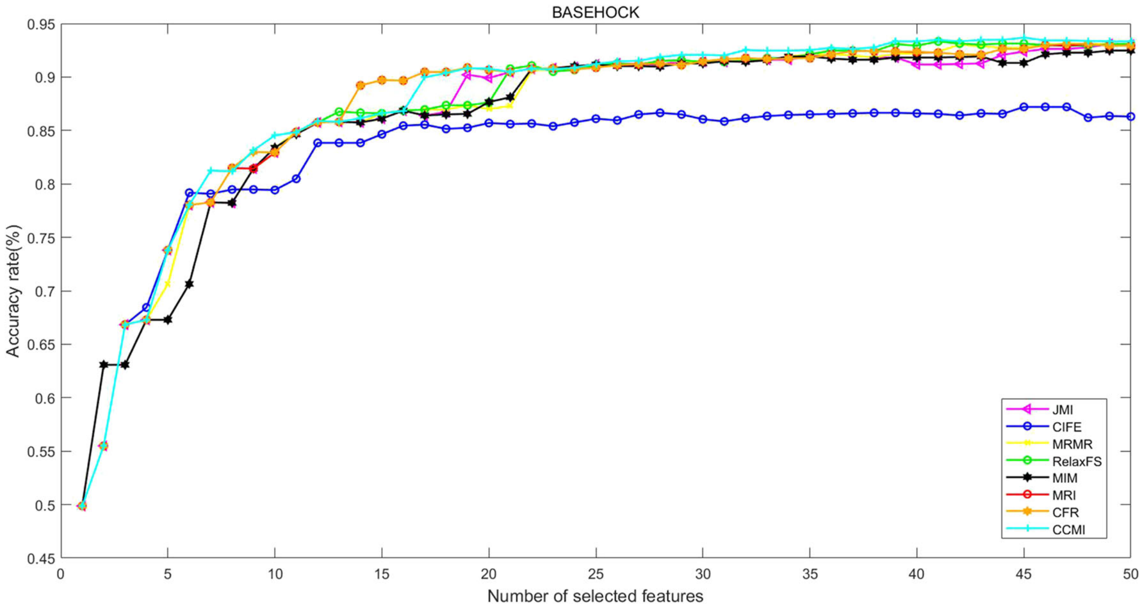

Fig. 3 Accuracy comparisons on BASEHOCK data set

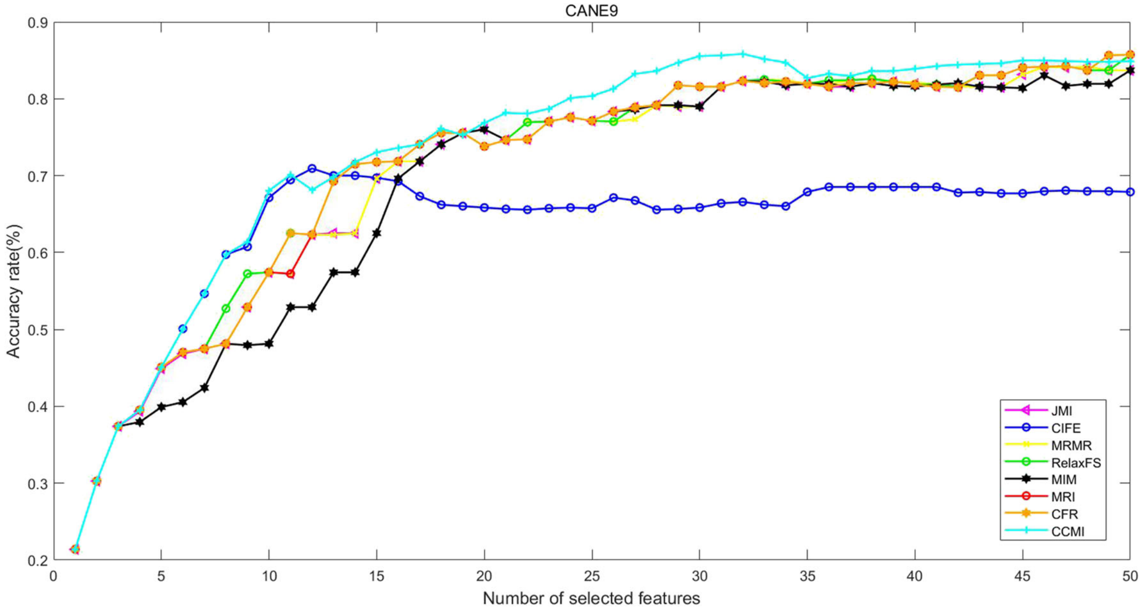

Fig. 4 Accuracy comparisons on CANE9 data set

accuracy of CCMI is 2.14% higher than that of RelaxFS. On the data sets RELATHE and USPS, the accuracy of CCMI is nearly 1.8% higher than that of RelaxFS. On the BASEHOCK data set, mRMR algorithm has the best performance. The accuracy of RelaxFS and CCMI is second only to mRMR, which is 0.06% lower. On the Wine data set, MIM algorithm has achieved the best classification accuracy. CCMI is 0.42% lower than it. It can be seen that the accuracy of all algorithms on the Wine data set is not much different. It can be seen in the last row of Table 2 that CCMI algorithm has a significant improvement compared to other algorithms.

Table 2 shows the accuracy of 8 different algorithms on the SVM classifier. By Comparing the average classification accuracy of CCMI and RelaxFS, the average classification accuracy of CCMI algorithm on the data set BASEHOCK is the same as that of RelaxFS on BASEHOCK. On the Wine data set, the average classification accuracy of CCMI is slightly lower than that of RelaxFS. While on other data sets, the classification accuracy of CCMI algorithm

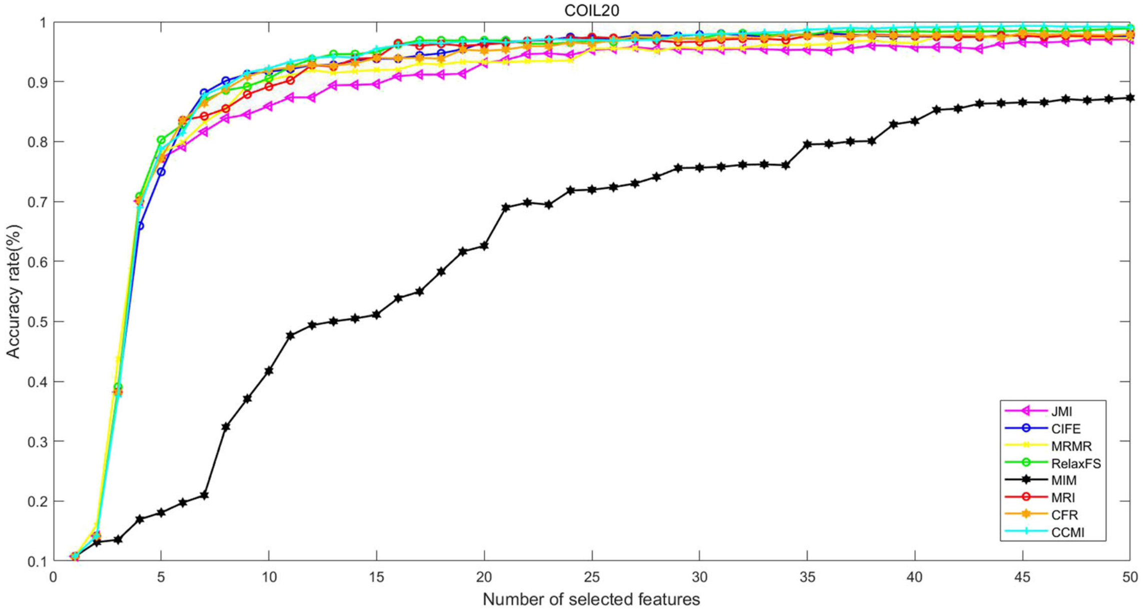

Fig. 5 Accuracy comparisons on COIL20 data set

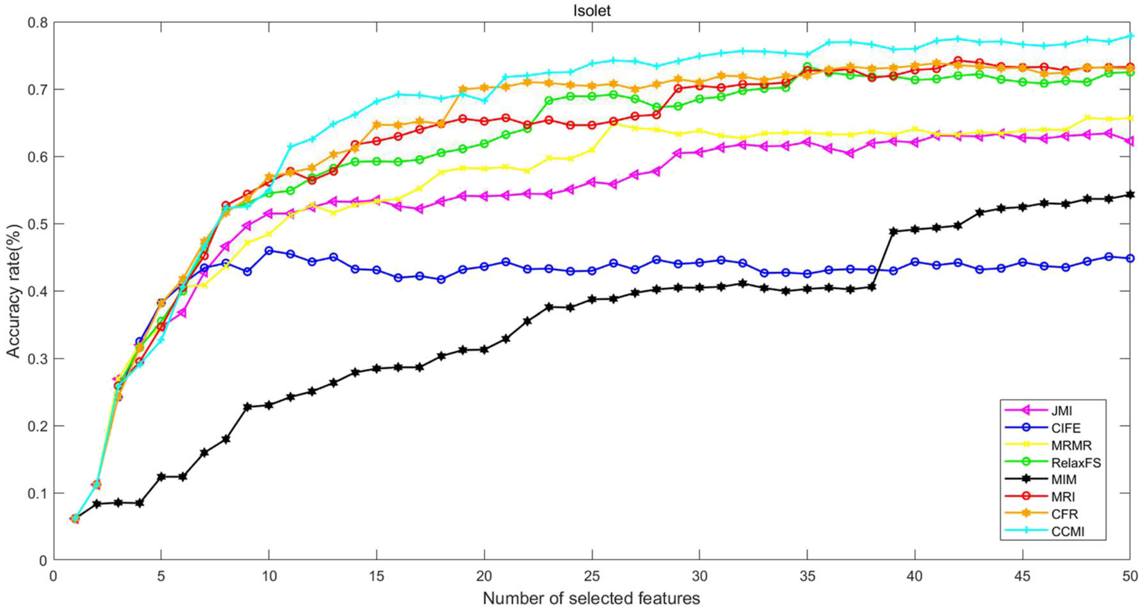

Fig. 6 Accuracy comparisons on Isolet data set

is higher than that of RelaxFS. The results of the experiment can be explained according to the evaluation criteria of CCMI and RelaxFS. RelaxFS uses the average value for processing which reflects the central tendency of related items and redundant items. But it also generates unnecessary redundant information which reduces the classification accuracy.CCMI evaluation standard uses minimization processing which reduces unnecessary redundant information compared with averaging processing. MIM algorithm only considers the relevancy between the class label and features, resulting in the smallest value of the average classification accuracy. Compared with MIM algorithm, CIFE algorithm considers both the relevant information and the redundant information. Compared with CIFE algorithm, JMI algorithm uses averaging processing in redundant items. So the average classification accuracy of JMI algorithm is higher than that of CIFE. MRI considers independent classification information which includes both independent classification information provided by candidate features and independent classification information retained by selected features.

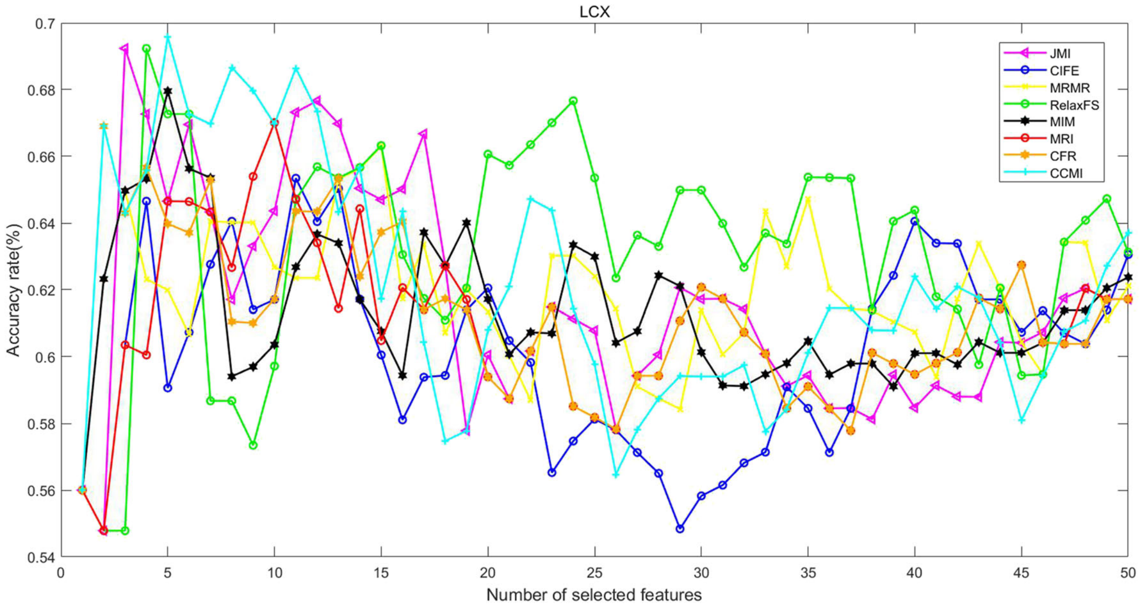

Fig. 7 Accuracy comparisons on LCX data set

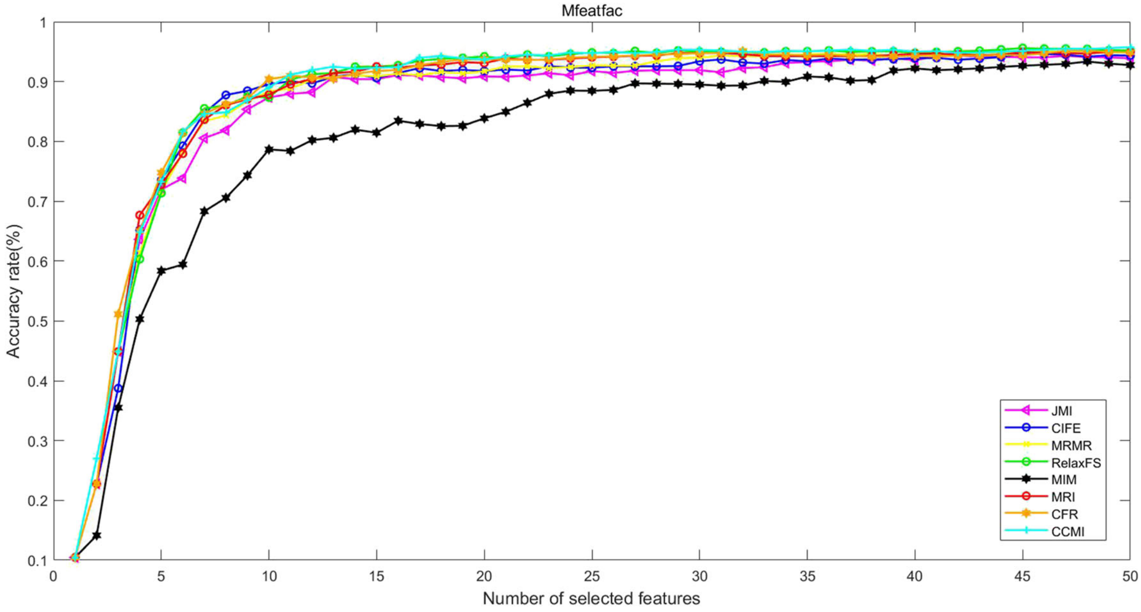

Fig. 8 Accuracy comparisons on Mfeatfac data set

So the average classification accuracy of MRI is relatively high.The redundant item of RelaxFS algorithm uses conditional redundancy to measure, resulting in an average classification accuracy higher than that of mRMR. It can be seen from the last row of the table that CCMI algorithm has a better average classification accuracy on the SVM classifier than the other 7 algorithms.

Table 3 shows the accuracy rates of 8 different algorithms on the KNN classifier. Comparing Table 2 with Table 3, it can be seen that the classification accuracy rates obtained by using different classifiers are different. On the Orlraws10P data set, CCMI has the highest classification accuracy when using the SVM classifier, and RelaxFS has the highest classification accuracy when using the KNN classifier.

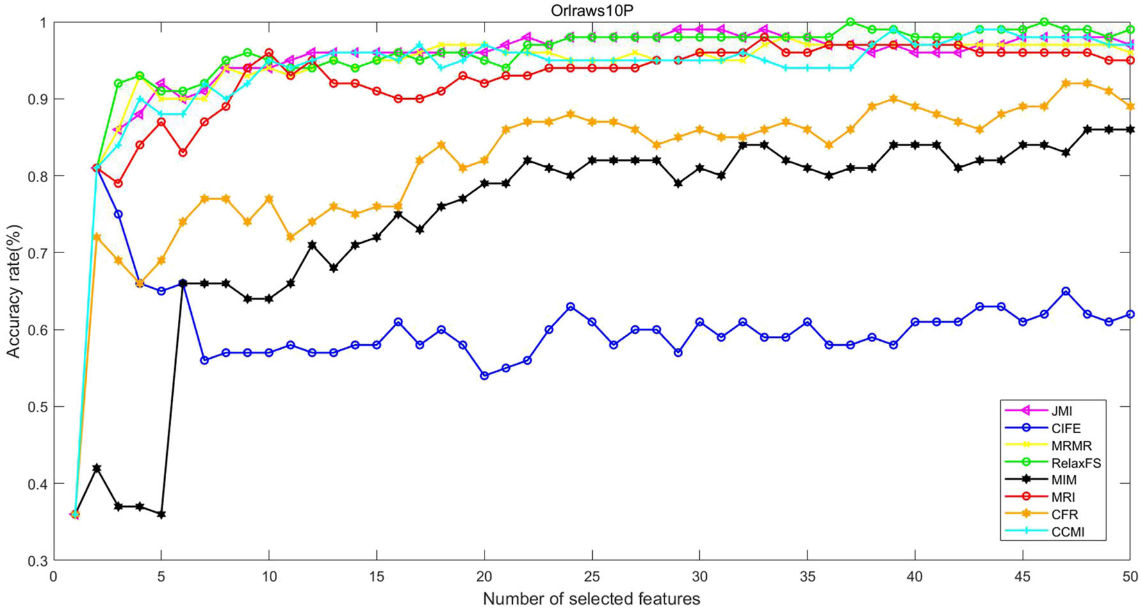

Fig. 9 Accuracy comparisons on Orlraws10P data set

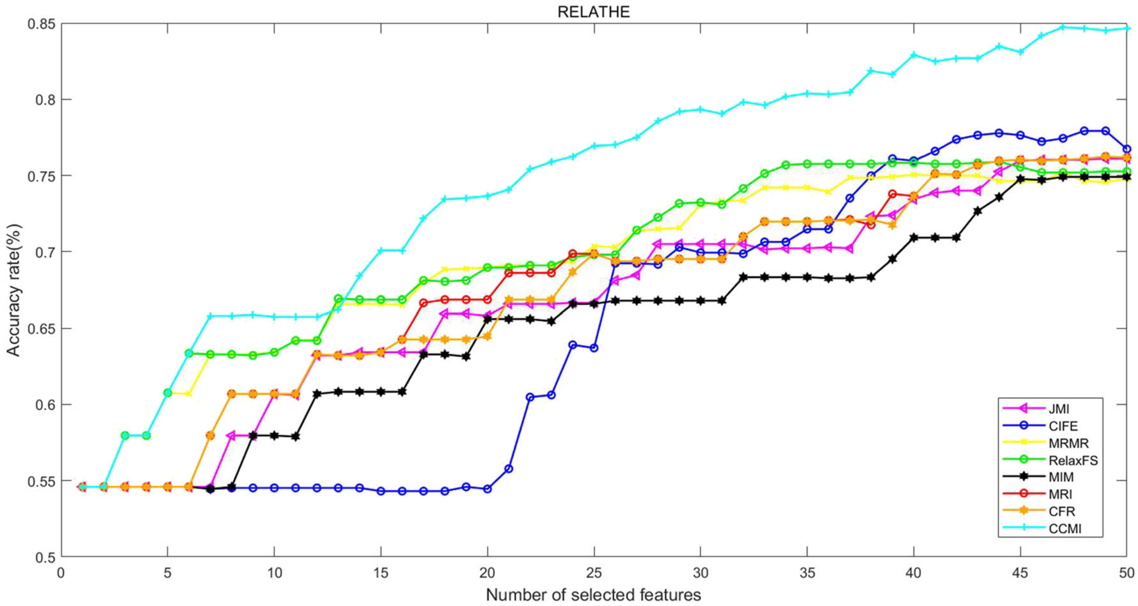

Fig. 10 Accuracy comparisons on RELATHE data set

Although different classifiers are used, the average classification accuracy of all algorithms has the same trend. That is CCMI>RelaxFS>MRI>CFR>mRMR>JMI>CIFE>MIM, where the classification performance of mRMR and CFR on the KNN classifier is almost the same.

In order to compare the classification accuracy of different feature selection algorithms under different number of features, Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 shows the classification accuracy graphs of CCMI algorithm and other 7 feature selection algorithms on the KNN classifier. These graphs are obtained by training 12 data sets. In these figures, the horizontal axis represents the number of features in the feature subset, and the vertical axis represents the classification accuracy. In the figure, different colors and shapes are used to distinguish different feature selection methods.

In order to compare the classification accuracy of different feature selection algorithms in different number of features, Figs. 3–14 shows the classification accuracy diagram generated by CCMI algorithm and other 7 feature selection algorithms respectively training 12 data sets on 3NN classifier. In these figures, the abscissa axis represents

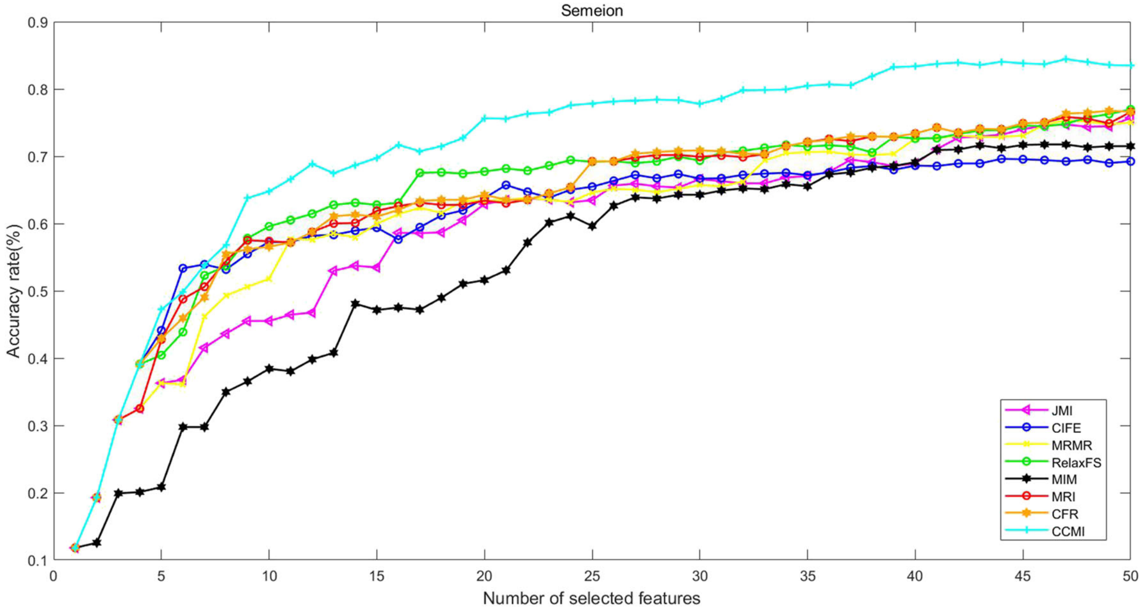

Fig. 11 Accuracy comparisons on Semeion data set

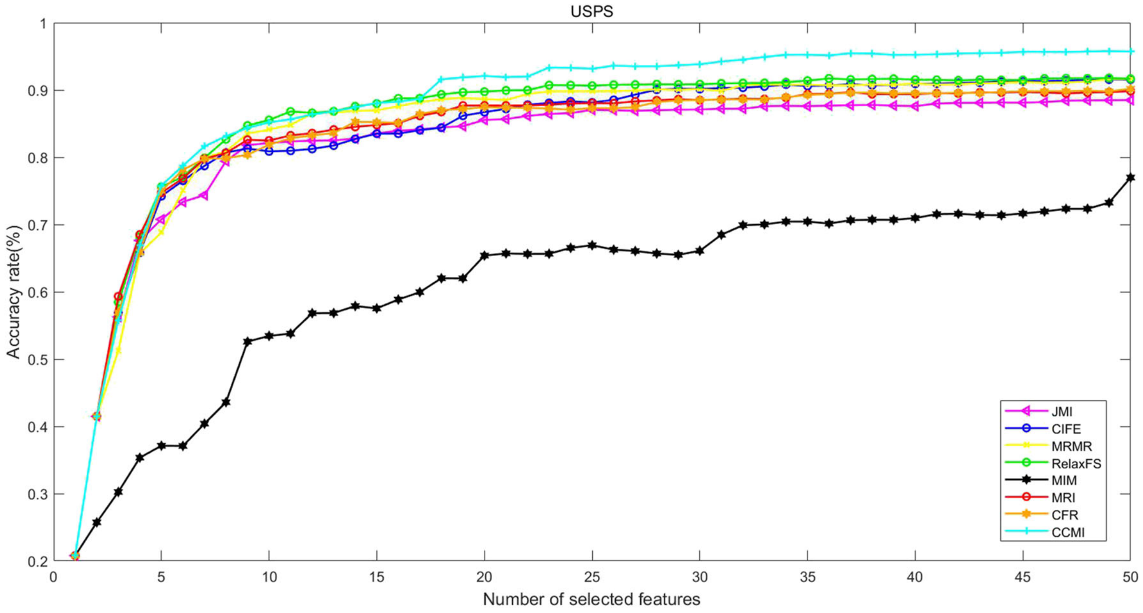

Fig. 12 Accuracy comparisons on USPS data set

the number of features in the feature subset and the vertical axis represents the classification accuracy. Different colors and shapes are used to distinguish different feature selection methods in the diagram.

According to the distribution of the accuracy charts in Figs. 3–14, we roughly divide them into three categories. The first category is the BASEHOCK, COIL20, Mfeatfac, Orlraws10P and WarpPIE10P data sets. The classification accuracy charts of these data sets have a common feature, that is, the accuracy of CCMI algorithm is within a small range compared with other algorithms. Among them, on the Orlraws10P data set, CCMI is slightly lower than RelaxFS, JMI and mRMR, which is consistent with the data in Table 3. On the other four data sets except the Orlraws10P data set, the accuracy of CCMI is higher than other algorithms. Moreover, with the increase in the number

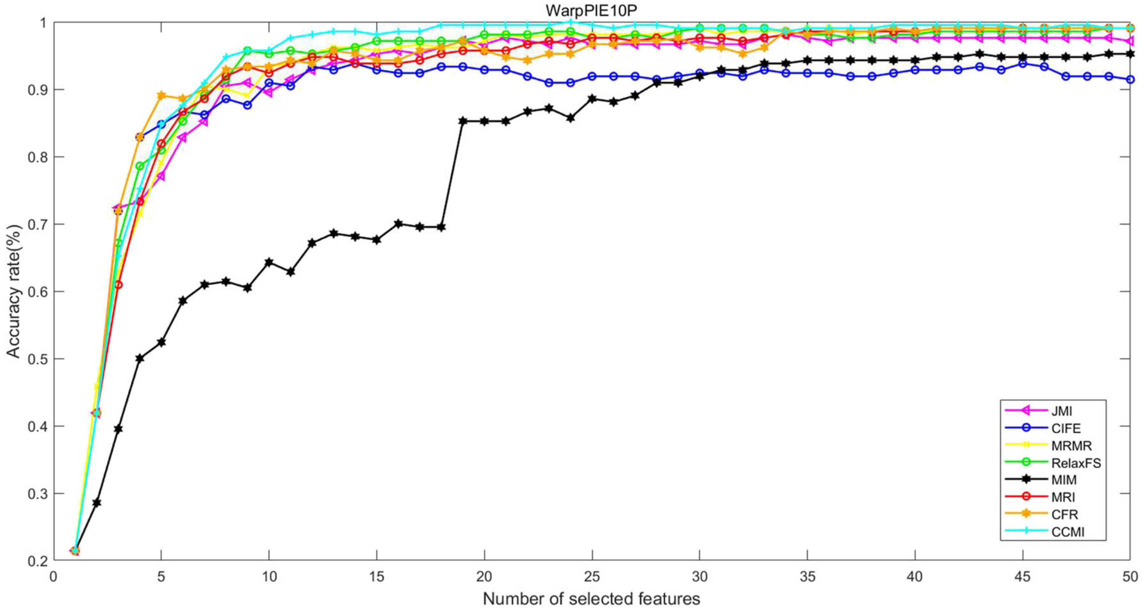

Fig. 13 Accuracy comparisons on WarpPIE10P data set

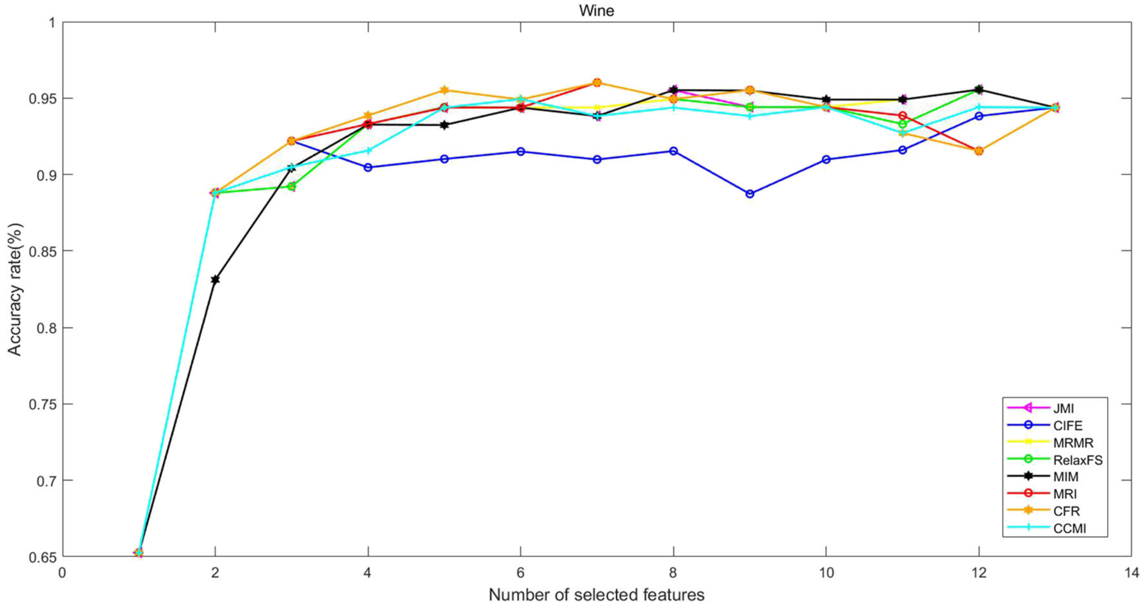

Fig. 14 Accuracy comparisons on Wine data set

of features, the classification accuracy rate also continues to increase. After the number of features is greater than 20, the increase in classification accuracy tends to be flat.

The second category is the CANE9, Isolet, RELATHE, Semeion and USPS data sets in Figs. 3–14. On these data sets, the classification accuracy of CCMI is significantly higher than other feature selection algorithms. The classification accuracy curve of CCMI on the RELATHE and Semeion data sets is much higher than all other algorithms. Especially after the number of features is greater than 15, the accuracy of the proposed algorithm has been increasing, and has been much higher than the accuracy of other algorithms. When the number of features of CANE9 and USPS data sets is less than 15, the classification accuracy of CCMI is not significantly different from other algorithms. When the number of features is greater than 15, the classification accuracy of CCMI has been significantly improved compared with other algorithms. On the USPS data set, after the number of features reaches 30, the accuracy of CCMI classification hardly changes. On the CANE9 data set, when the number of features reaches 30, the accuracy of CCMI decreases slightly and gradually tends to a fixed value. On the Isolet data set, the accuracy of CCMI is significantly improved compared with other algorithms after the feature number reaches 10.

The third category is the LCX and Wine data sets in Figs. 3–14. It can be seen in the figure that when the number of features of the LCX data set is less than 20, the accuracy

Table 4 Highest classification accuracy using SVM

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	0.9498	0.8745	0.9498	0.9488	0.9498	0.9508	0.9498	**0.9528**
CANE9	0.8630	0.7370	0.8574	0.8602	0.8685	0.8574	0.8685	**0.8704**
COIL20	0.9660	0.9688	0.9438	0.8306	0.9736	0.9535	**0.9806**	0.9799
Isolet	0.8090	0.6660	0.7276	0.6167	0.8045	0.7259	0.8006	**0.8686**
LCX	0.7025	0.6902	0.6996	0.7094	0.6865	0.6935	0.6997	**0.7099**
Mfeatfac	0.9580	0.9440	0.9470	0.9450	0.9595	0.9500	0.9585	**0.9610**
Orlraws10P	0.9500	0.8100	0.9900	0.8800	0.9900	0.9800	**1**	**1**
RELATHE	0.8129	0.8038	0.8157	0.8003	0.8157	0.8220	0.8325	**0.8689**
Semeion	0.7671	0.7106	0.7640	0.7320	0.7583	0.7621	0.7822	**0.8524**
USPS	0.8860	0.8928	0.8756	0.7863	0.8860	0.8992	0.9003	**0.9357**
WarpPIE10P	0.9857	0.9857	0.9762	0.9619	0.9857	0.9857	0.9905	**0.9952**
Wine	0.9660	**0.9663**	0.9608	0.9611	**0.9663**	**0.9663**	**0.9663**	**0.9663**

Table 5 Highest classification accuracy using KNN(k=5)

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	0.9318	0.8615	0.9303	0.9308	0.9318	0.9323	0.9333	**0.9348**
CANE9	0.8583	0.7000	0.8426	0.8417	0.8583	0.8426	0.8574	**0.8667**
COIL20	0.9660	0.9625	0.9514	0.8556	0.9681	0.9583	0.9715	**0.9903**
Isolet	0.7500	0.4795	0.6513	0.5615	0.7449	0.6513	0.7449	**0.7795**
LCX	0.6800	0.6892	0.7067	0.6867	0.6832	0.6635	0.7132	**0.7263**
Mfeatfac	0.9540	0.9495	0.9470	0.9275	0.9510	0.9485	0.9590	**0.9600**
Orlraws10P	0.8900	0.7500	0.9800	0.8300	0.9600	0.9700	**1**	**1**
RELATHE	0.7624	0.7919	0.7617	0.7526	0.7624	0.7448	0.7519	**0.8451**
Semeion	0.7784	0.7119	0.7634	0.7358	0.7784	0.7615	0.7835	**0.8393**
USPS	0.9045	0.9157	0.8922	0.7717	0.9021	0.9183	0.9193	**0.9550**
WarpPIE10P	0.9857	0.9143	0.9762	0.9524	0.9857	**0.9905**	**0.9905**	**0.9905**
Wine	**0.9722**	0.9608	0.9611	0.9608	**0.9722**	0.9667	0.9667	0.9663

of CCMI is significantly higher than other algorithms. When the number of features is greater than 20, there is no significant difference compared with other algorithms. On the whole, as the number of features increases, all algorithms have a downward trend in accuracy. On the Wine data set, CIFE algorithm is obviously less accurate than other algorithms, while the accuracy of CCMI is comparable to other algorithms.

This experiment selects up to 50 features. Within the range of these 50 features, by comparing the classification accuracy of different feature selection algorithms in different dimensions, we can conclude that the higher the dimension, the higher the classification accuracy.

Tables 4 and 5 respectively describe the maximum classification accuracy of the 8 feature selection algorithms on the SVM classifier and KNN classifier. The maximum value is marked in bold on each line.

As shown in Table 4, the maximum classification accuracy of CCMI on other data sets is higher than that of the other seven algorithms except for the COIL20 data set. On the COIL20 data set, the maximum classification accuracy of RelaxFS is higher than other algorithms and 0.07% higher than CCMI. On the Wine data set, the maximum classification accuracy of CCMI, RelaxFS, mRMR, MRI, and CIFE is the same, which is 96.63%.

In Table 5, the maximum classification accuracy of CCMI on other data sets is higher than other algorithms except for the Wine data set. Among them, on the Orlraws10P data set, the maximum value of CCMI and RelaxFS is both 100%, which is higher than other algorithms. On the WarpPIE10P data set, the maximum value of CCMI, RelaxFS and mRMR algorithms is the same and is 99.05%, which is higher than other algorithms.

5.3.2 Macro-F1

The value of Macro-F1 is the harmonic average of precision ratio and recall ratio, which is suitable for performance

Table 6 Macro-F1 comparison using SVM

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	0.9108	0.8529	0.9103	0.9080	0.9117	**0.9129**	0.9123	0.9122
CANE9	0.7248	0.6686	0.7214	0.7066	0.7255	0.7218	0.7297	**0.7525**
COIL20	0.8733	0.8630	0.8423	0.6026	0.8756	0.8651	0.8906	**0.8924**
Isolet	0.6826	0.5579	0.5981	0.3967	0.6748	0.6068	0.6712	**0.7350**
LCX	0.6390	0.6278	0.6388	0.6417	0.6327	0.6362	0.6322	**0.6430**
Mfeatfac	0.8998	0.8807	0.8841	0.8473	0.8974	0.8908	0.8985	**0.9019**
Orlraws10P	0.6054	0.4560	0.6097	0.4678	0.6083	0.6015	0.6148	**0.6255**
RELATHE	0.7834	0.7870	0.7710	0.7603	0.7833	0.7861	0.7905	**0.8084**
Semeion	0.6579	0.6257	0.6294	0.5775	0.6510	0.6459	0.6715	**0.7397**
USPS	0.8120	0.8136	0.8053	0.5895	0.8144	0.8264	0.8349	**0.8545**
WarpPIE10P	0.8404	0.8310	0.8279	0.7360	0.8307	0.8342	0.8354	**0.8490**
Wine	0.8803	0.8588	0.8788	**0.8876**	0.8797	0.8826	0.8821	0.8803

Table 7 Macro-F1 comparison using KNN(k=5)

Data sets	CFR	CIFE	JMI	MIM	MRI	mRMR	RelaxFS	CCMI
BASEHOCK	0.8698	0.8218	0.8641	0.8603	0.8694	0.8649	0.8661	**0.8718**
CANE9	0.7048	0.6274	0.6996	0.6841	0.7042	0.7008	0.7061	**0.7353**
COIL20	0.8809	0.8807	0.8555	0.5865	0.8838	0.8655	0.8894	**0.9024**
Isolet	0.6237	0.4171	0.5152	0.3433	0.6067	0.5293	0.6029	**0.6505**
LCX	0.6074	0.6000	**0.6120**	0.6049	0.6035	0.5842	0.6086	0.6062
Mfeatfac	0.8760	0.8685	0.8573	0.8064	0.8736	0.8675	0.8796	**0.8837**
Orlraws10P	0.5220	0.3843	0.6425	0.4803	0.6202	0.6372	**0.6458**	0.6243
RELATHE	0.6069	0.5680	0.5960	0.5668	0.6101	0.6326	0.6378	**0.7052**
Semeion	0.6275	0.5984	0.5957	0.5417	0.6259	0.6138	0.6402	**0.7026**
USPS	0.8204	0.8251	0.8097	0.5825	0.8233	0.8355	0.8448	**0.8674**
WarpPIE10P	0.8357	0.7406	0.8121	0.7056	0.8268	0.8279	0.8293	**0.8371**
Wine	**0.9253**	0.9026	0.9184	0.9095	0.9247	0.9197	0.9192	0.9196

evaluation of the class imbalance problem. In order to evaluate the performance of CCMI algorithm proposed in this paper, Macro-F1 value is used to evaluate the results of classifiers in different data sets. Table 6 shows the Macro-F1 values on SVM. It can be seen that on all data sets, the F1 value of CCMI is only slightly lower than mRMR algorithm on BASEHOCK data set, slightly lower than MIM algorithm on Wine data set, and higher than other algorithms on other data sets. In Table 7, it shows the Macro-F1 values on KNN. F1 value of CCMI is slightly lower on LCX, Wine and Orlraws10P data sets while the value of F1 is the maximum on other data sets compared with other algorithms.

6 Summary

This paper mainly studies the feature selection algorithm based on mutual information. The research goal is to select a feature subset that is highly relevant to the class and has low redundancy between features. Mutual information is used to measure the relationship between the class label and features and between features and features. In order to filter out more redundant information, a correlation coefficient is introduced to describe the degree of redundancy between features, and the absolute value of the correlation coefficient is used to measure the importance of redundant items in the evaluation function. The absolute value of the correlation coefficient is used as the weight of the redundancy item to balance the importance of the relevant item and the redundancy item in the evaluation function. We also use the principle of minimization to select a subset of features with lower redundancy, thereby improving the classification accuracy. Experimental results show that our proposed CCMI algorithm has better classification performance compared with the existing feature selection algorithms.

CCMI algorithm introduces the absolute value of the correlation coefficient as the weight of the redundant item. The correlation coefficient mainly describes the degree of linear correlation between variables. Therefore, the degree of nonlinear correlation is not considered, and the redundant information of nonlinearity has not been completely removed although the performance has been improved. In the future, we plan to use distance correlation coefficients to filter out more redundant information.

Acknowledgements The corresponding author would like to thank the support from the National Key Research and Development Plan under the Grant of 2017YFB1402103, the National Natural Science Foundation of China under the Grant of 61971347, the Key Research and Development Program of Shaanxi under the Grant of 2021JZ-49, the Education Department of Shaanxi Province Key Laboratory Project under the Grant of 15JS079, Xi'an Science Program Project under the Grant of 2020KJRC0094.

References

1. Wang Z, Li M, Li J (2015) A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure. Inf Sci 307:73–88
2. Bennasar M, Hicks Y, Setchi R (2015) Feature selection using joint mutual information maximisation. Expert Syst Appl 42:8520–8532
3. Hoque N, Bhattacharyya DK, Kalita JK (2014) MIFS-ND: A mutual information-based feature selection method. Expert Syst Appl 41:6371–6385
4. Zhou H, Guo J, Wang Y (2016) A feature selection approach based on term distributions. Springerplus 5(1):1–14
5. Das SK, Das SR (2001) Wrappers and a boosting-based hybrid for feature selection. In: Proceedings of the eighteenth international conference on machine learning, pp 74–81
6. Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. Inf Sci 179:2208–2217
7. Jiang L, Kong G, Li C (2019) Wrapper framework for test-cost-sensitive feature selection. IEEE Trans Sys Man Cybern Sys:1–10

8. Zhu QH, Bin YY (2018) Discriminative embedded unsupervised feature selection. Patt Recogn Lett 112:219–225
9. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann, Burlington
10. Vinh NX, Zhou S, Chan J, Bailey J (2016) Can high-order dependencies improve mutual information based feature selection? Patt Recogn 53:46–58
11. Peng H, Fan Y (2017) Feature selection by optimizing a lower bound of conditional mutual information. Inf Sci 418-419:652–667
12. Lewis DD (1992) Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on speech and natural language. Association for Computation Linguistics, pp 212–217
13. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5:537–550
14. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Patt Anal Mach Intell 27(8):1226–1238
15. Lin D, Tang X (2006) Conditional infomax learning: an integrated framework for feature extraction and fusion. In: European Conference on computer version, pp 68–82
16. Yang HH, Moody J (1999) Feature selection based on joint mutual information. In: Proceedings of international ICSC symposium on advances in intelligent data analysis, pp 22–25
17. Wang J, Wei JM, Yang Z, Wang SQ (2017) Feature selection by maximizing independent classification information. IEEE Trans Knowl Data Eng 29(4):828–841
18. Gao W, Hu L, Zhang P et al (2018) Feature selection considering the composition of feature relevancy. Patt Recogn Lett 112: 70–74
19. Brown G, Pocock A, Zhao MJ, Lujan M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13(1):27–66
20. Zhou HF, Zhang Y, Zhang YJ, Liu HJ (2018) Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy. Appl Intell 49:883–896
21. Asuncion A, Newman DJ (2007) UCI machine learning repository, University of California, Irvine, School of Information and Computer Science. http://www.ics.uci.edu/mlearn/MLRepository.html
22. Borah P, Gupta D (2020) Unconstrained convex minimization based implicit Lagrangian twin extreme learning machine for classification (ULTELMC). Appl Intell 50:1327–1344
23. Gupta D, Sarma HJ, Mishra K, Prasad M (2019) Regularized Universum twin support vector machine for classification of EEG signal. In: IEEE international conference on systems, man and cybernetics (SMC), pp 2298–2304
24. Adhikary D, Das GD (2020) Applying over 100 classifiers for churn prediction in telecom companies. Multimedia Tools and Applications
25. Borah P, Gupta D (2020) Functional iterative approaches for solving support vector classification problems based on generalized Huber loss. Neural Comput Appl 32:9245–9265
26. Gupta D, Borah P, Prasad M (2017) A fuzzy based Lagrangian twin parametric-margin support vector machine (FLTPMSVM). In: IEEE symposium series on computational intelligence (SSCI), pp 1–7
27. Demišar J, Schuurmans D (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7(1):1–30

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Hongfang Zhou received the B.S. degree in Computer Application from Xi'an University of Technology, Xi'an, China, in 1999. She received M.S. degree in computer application technology from Xi'an University of Technology, Xi'an, China, in 2002. And she received the Ph.D. degree in computer software and theory from Xi'an Jiaotong University, Xi'an, China, in 2006. She is currently an associated professor in the School of Computer Science and Engineering, Xi'an University of Technology, China. Her research interests include Data Mining and Knowledge Engineering.

Xiqian Wang received the B.S. degree from the School of Computer Science and Engineering, Xi'an University of Technology in 2017 and is now studying for her Master degree in School of Computer Science and Engineering, Xi'an University of Technology. Her research interests are focused on data mining and feature selection.

Rourou Zhu received the B.S. degree from the School of Information and Enginering from Xi'an University of Technological Information in 2019 and is now studying for her Master degree in School of Computer Science and Engineering, Xi'an University of Technology. Her research interests are focused on data mining and feature selection.