IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2019, accepted October 12, 2019, date of publication October 17, 2019, date of current version October 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948095

Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization

XINZHENG WANG[1,2], BING GUO[1], YAN SHEN[3], CHIMIN ZHOU[1], AND XULIANG DUAN[1]

[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]School of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China
[3]School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China

Corresponding authors: Xinzheng Wang (47312582@qq.com) and Bing Guo (guobing@scu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772352, and in part by the Science and Technology Planning Project of Sichuan Province under Grant 2019YFG0400, Grant 2018GZDZX0031, Grant 2018GZDZX0004, Grant 2017GZDZX0003, Grant 2018JY0182, and Grant 19ZDYF1286.

ABSTRACT Feature selection is the first and essential step for dimension reduction in many application areas, such as data mining and machine learning, due to its computational efficiency and interpretability of the results. This paper focuses on feature selection methods based on information theory. By studying and analyzing the ideas and drawbacks of existing feature selection methods, it finds that in the process of feature selection separately focuses on a candidate feature its individual relationship with the predicted class vector may lead to some problems. And we believe that when the candidate feature is combined with the selected features, its comprehensive discriminative ability should be taken as the evaluation index of the candidate feature. Therefore, we propose a novel feature selection method in this paper. In the proposed method, we introduced the equivalent partition concept and adopted the mutual information gain maximize (MIGM) criterion to evaluate the candidate feature. In order to estimate the performance of MIGM, we conducted experiments on ten benchmark datasets and two different classifiers, k-Nearest Neighbor (KNN) and Naïve-Bayes (NB). Extensive experimental results demonstrate that our method can identify an effective feature subset that leads to better classification results than other methods.

INDEX TERMS Feature selection, mutual Information, equivalent partition.

I. INTRODUCTION

With the development of computer and network technology, especially the development of data sensors, data acquisition becomes more and more convenient. As a result, the size of the dataset has been increasing rapidly both in terms of the number of features and the number of observations. The growing size of data results in many research challenges to data analysis, for example, biomedical data analysis [1], pattern recognition, machine learning, etc. Actually, in practical applications, not all features of the data contribute equally to final results, for irrelevant and redundant features may be included on some occasions. And those unrelated features may affect many aspects of data mining, such as the accuracy of the classifiers and over-consumption on computational resources [2]. For machine learning systems or the expert systems, it is reasonable to identify the main contributing features and use them instead of the whole feature set [3]. For this reason, many techniques have been proposed to reduce the dimension of datasets [4]–[6]. The dimension reduction usually aims at two aspects: enhancing data representation and improving classification performance. The purpose of enhancing data representation is to preserve the topological structure of the data as much as possible while reducing the data dimension. And for the classification application is to find the feature subset that has more discriminative power. Besides, the underlying process of generating data can be better understood by dimension reduction [7]. Therefore, in the process of modern data analysis, the first and essential step is to reduce the dimension of data.

There are two types of strategies for dimension reduction: feature extraction and feature selection [8]. Feature extraction is a general term for methods of constructing a combination of original features to address the problems while still

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

describing the data with sufficient accuracy. So the feature extraction builds informative and non-redundant low dimensional representation by transforming the original features. There are many feature reduction methods belong to this type, such as the principal component analysis (PCA), Laplacian eigenmaps (LE) [9], local linear embedding (LLE) [10], and ISOMAP [11]. However, these techniques often involve the time-consuming calculation of eigenvalues and eigenvectors of the covariance matrix. Compared with the feature selection method, feature extraction is more time-consuming, and the result is inexplicable. And the feature selection strategy is to find the feature subset S from the original feature set F, where $S \subseteq F$. The subset S should preserve the characteristics of original features, and it can produce equal or better classification accuracy compared with feature set F. In this study, we concern feature selection rather than feature extraction.

There are two general approaches for feature selection: wrapper and filter [12]. The wrapper approach searches the feature space and tests all possible feature combination subsets by using the prediction accuracy of a specific classifier as the measurement of the quality of the selected feature subset [13]. Because the wrapper method chooses features related to a specific classifier, so it has high classification accuracy, but its generalization ability is low. Another problem with wrapper is that it is computation-intensive and time-consuming. In wrapper, each feature is indicated by one bit, 1 for the feature presence, and 0 for the feature absence. For the data with n features, there are n bits in each state, so the size of the search space is $O(2^n)$. It is impossible to search the whole feature space unless n is small. However, the filter method is different from the wrapper method that it is not associated with a specific classifier, and it evaluates the predictive ability of features according to some criteria, such as probability distance, Fisher score, and Pearson correlation. It is worth noting that many feature selection methods based on information theory were proposed in the past two decades [14], [15], such as the mutual information based method. And the mutual information based feature selection method belongs to the filter approach, and it is independent of the specific classifier. Mutual information [16], [17] theory is widely employed in feature selection methods, one of the important reasons is that mutual information and conditional mutual information can be used to measure the linear and non-linear correlations [18].

Using mutual information as the measure of feature selection mainly involves three concepts: feature relevance, redundancy, and interaction. The concepts of relevance and redundancy have been applied in extensive literature. One of the oldest feature selection methods based on information theory is the Mutual Information Maximization (MIM) [19], which adopts mutual information to measure the relevancy between each feature and the output class vector, while do not consider the relationship between features. As a result, the selected feature subset may contain many redundant features. Mutual Information Feature Selection (MIFS) [17] improves the performance of MIM by taking the relationship between features into consideration, and it reduces the redundancy of the selected feature subset. And many variants of MIFS have been developed, such as MIFS-U [20], MIFS-ND [21], and Minimum Redundancy Maximum Relevance (MRMR) [22]. Moreover, besides the relevance and redundancy of an individual feature, several methods also focus on the joint effects of multiple features. For instance, the Joint Mutual Information (JMI) [23] method employs the mutual information between the joint features and class vector as the measure criterion. The Joint Mutual Information Maximization (JMIM) and Normalized Joint Mutual Information Maximization (NJMIM) [13] also test the joint effect of features, and they utilize the ''maximize the minimum'' principle in the feature selection process. Many other mutual information based methods are also stated in the literature, such as the Conditional Mutual Information Maximization criterion (CMIM) [24] and Double Input Symmetrical Relevance (DISR) [25].

In this paper, we present a novel feature selection method based on mutual information. Our contributions are summarized as follows:

1) We introduced the feature equivalent partition concept for the feature subset.

2) A feature selection method based on mutual information gain maximization is proposed.

3) Experiments on ten benchmark data sets are carried out to verify the effectiveness of our method.

The rest of this paper is organized as follows: In section 2, the basic concepts of information theory are introduced. Related feature selection techniques based on mutual information are reviewed in section 3. In section 4, the drawbacks of existing methods are discussed, and a novel feature selection based on information theory is presented. The experimental results and comparisons with other approaches are given in section 5. Finally, the paper is concluded in section 6.

II. CONCEPTS OF INFORMATION THEORY

This section mainly introduces some basic concepts of information theory, including entropy, condition entropy, joint entropy, mutual information, conditional mutual information, joint mutual information, etc. [5], [18], [20]. In 1948, Shannon published his famous paper '' A mathematical theory of communication'', which proposed the concept of information entropy to quantify information [16]. The entropy of a random variable is a measurement of its uncertainty, as well as the average amount of information needed to describe the random variable. For a random discrete variable X has N different number of values, $X = \{x_1, x_2, \ldots, x_N\}$, Y has M different number of values, $Y = \{y_1, y_2, \ldots, y_M\}$, and the entropy of X is denoted as $H(X)$:

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

where

$$p(x_i) = \frac{number\ of\ instants\ with\ value\ x_i}{total\ number\ of\ instants\ of\ X} \tag{2}$$

For a continuous random variable, the definition of entropy is defined as follows:

$$H(X) = -\int p(x)\, logp(x)\, dx \tag{3}$$

The conditional entropy of variable Y is the amount of uncertainty left in variable Y after the variable X is introduced. It can be defined as follows:

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

The joint entropy of X and Y is the uncertainty that occur simultaneously with two variables, and it denotes as $H(Y,X)$:

$$H(Y, X) = -\sum_{x_i \in X} \sum_{y_j \in Y} p(y_j, x_i) log(p(y_j, x_i)) \tag{5}$$

where $p(y_j, x_i)$ is the joint probability of y_j and x_i.

The relationships among entropy, conditional entropy, and joint entropy are as follows [18]:

$$H(Y, X) = H(X, Y) = \begin{cases} H(X) + H(Y|X) \\ H(Y) + H(X|Y) \end{cases} \tag{6}$$

Mutual information (MI) is another import concept of information theory. It is the amount of information provided by the newly introduced variable, and it also can be regarded as the information shared by two variables. MI is used to quantify the mutual dependence between two variables. When the two variables are independent of each other, the MI is zero, and it increases with the increases of one variable's dependence on another. The mutual information between variable Y and variable X is denoted as $I(Y; X)$, and it can be formulized as follows:

$$I(Y; X) = \begin{cases} H(Y) - H(Y|X) \\ H(X) - H(X|Y) \\ H(Y) + H(X) - H(Y, X) \end{cases} \tag{7}$$

$$I(Y; X) = \sum_{x \in X} \sum_{y \in Y} p(y, x)\, log \frac{p(y, x)}{p(y)\, p(x)} \tag{8}$$

For continuous random variables, the mutual information is defined as follows:

$$I(X; Y) = \int p(x, y)\, log \frac{p(x, y)}{p(x)\, p(y)} dx\, dy \tag{9}$$

However, it's difficult to find the probability density functions ($p(x)$, $p(y)$, $p(x, y)$) exactly. Therefore, the continuous variables are usually discretized first, and then the entropy and mutual information are computed with discrete definitions.

Conditional mutual information $I(Y; X_i|X_j)$ quantifies the new discriminative information provided by X_i when X_j is selected, which is defined as follows:

$$I(Y; X_i \,|\, X_j) = \begin{cases} H(Y \,|\, X_j) - H(Y|X_i, X_j) \\ H(X_i \,|\, X_j) - H(X_i|Y, X_j) \\ I(X_i; Y|X_j) \end{cases} \tag{10}$$

Joint mutual information (JMI) quantifies the information shared by Y and the joint variables $(X_i, \ldots, X_k)$, which is defined as follows:

$$I(Y; X_i, \ldots, X_k) = H(Y) - -H(Y|X_i, \ldots, X_k) \tag{11}$$

Many literature define the amount of information shared among three variables as interaction information [26], [27], which is defined as follows:

$$I(Y; X_i; X_j) = I(Y; X_i) + I(Y; X_j) - I(Y; X_i, X_j) \tag{12}$$

The value of $I(Y; X_i; X_j)$ can be positive, negative, or zero, and it also can be regarded as the shared discriminative information of the two features. If $I(Y; X_i; X_j)$ is positive, it means X_i, X_j have redundancy. If $I(Y; X_i; X_j)$ is negative, it means X_i, X_j is complementary. And when $I(Y; X_i; X_j)$ is zero, it means that X_i and X_j are independent of each other [28]. When one of the features is selected, interaction information provided by another feature can be viewed as redundant for classification. So the conditional mutual information $I(Y; X_i|X_j)$ can be rewritten as $I(Y; X_i|X_j) = I(Y;X_i) - I(Y;X_i; X_j)$. It means that more new information a candidate feature provides, the less redundant information it corresponds.

Here we use a Venn diagram to demonstrate the relationships of the information concepts mentioned above.

III. RELATED WORK

In order to get an effective feature subset, many feature evaluation criteria were proposed. Such as the Fisher discriminant criterion, which uses the ratio of inter-class variance to intra-class variance of the feature as the feature score, then the features are ranked in descending order according to their scores, and the top k features will be selected. Although the feature ranking method is not optimal, it is widely used as a feature selection method in some cases because of its computational and statistical properties. Given the dataset $D(X_{i,j}, Y)$, where $i = 1 \ldots N$, $j = 1 \ldots M$, we can calculate the Correlation Coefficient $R(j)$ between j$^{\text{th}}$ feature and Y as follows:

$$R(j) = \frac{\sum_{i=1}^{N} (x_{i,j} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{13}$$

In fact, it is not necessary always to select the features with strong individual discriminative ability to form the feature subset. Because in real applications, the features of data are more or less correlated with each other, and the joint effects of them sometimes can lead to better classification accuracy [22]. Assuming there is a dataset $D(C, F)$, where C is the predicted class, $F = \{f_1, f_2, \ldots, f_n\}$ is the original feature set of data. The purpose of feature selection is to choose a feature subset $S \subset F$ from the original feature set, which can preserve the information of original features as much as possible. So the method based on mutual information is to maximize the joint mutual information between the selected feature subset S and predicted class vector C. Hence, it can be formulated as follows:

$$Max\ I(C; S) \tag{14}$$

where $S = \{f_1', \ldots, f_k'\}$, and k << n.

The mutual information (MI) based method is an important type of filtering method and Lewis etc. [19] first applied information theoretic metric to measure the correlation between

features and predicted class. These relationships are dependence, relevance, and redundancy. In addition, the interaction relationship is also defined in some literature [5], [25], [29]. The objective of feature selection methods based on mutual information is to identify the feature subset S, in which for any pair of features $(f_i, f_j) \in S$, the feature-feature mutual information is minimum and the feature-class mutual information is maximum.

In order to get the feature subset S, there are three strategies that can be considered. Firstly, all possible feature subsets are generated and then test the performance of each subset. In this way, the best subset can be found in theory, but it is impractical in applications. Because the number of combinations increases rapidly with the increases of feature number and it requires a lot of calculation. And choosing k features from n features has C_n^k combinations. Secondly, using an elimination strategy to eliminate the worst features from the full feature set F one-by-one until only k features remain. The third strategy is to start with an empty subset S, then select the best feature from F and add to the subset S one by one until the size of the subset S reaches k [20].

The Mutual Information Maximization (MIM), also known as Information Gain [19], is a feature ranking algorithm based on the feature-Class correlation and it ignores the feature-feature correlation. MIM has been widely used due to its simplicity and computational efficiency. This algorithm first computes the mutual information between each feature and the output class, $I(C; f_i)$, $i = 1,\ldots, n$. Then the features are ranked in descending order according to the value of $J_{MIM}(\cdot)$, and the top k features are selected. A larger value of $I(C; f_i)$ means greater relevance between f_i and C. The feature evaluation criterion of MIM is defined as follows:

$$J_{MIM}(f_i) = I(C; f_i) \tag{15}$$

MIM is based on the assumption that each feature of the data is independent, that is, $I(f_i; f_j) = 0, \forall f_i, f_j \in F$. But in practical application, it is not always the case. It may encounter the well-known trap that "the m best features are not the best m features" [30], [31]. Therefore, when features are interdependent, this approach will lead to sub-optimal results and may suffer from the limitation of feature redundancy. For this problem, the relationship between candidate features and the already selected features was also studied in MIFS [17]. It is defined as follows:

$$J_{MIFS}(f_i) = \mathrm{I(C}; f_i) - \beta \sum\nolimits_{f_j \in S} I(f_i; f_j) \tag{16}$$

where $f_i \in F - S$. In MIFS, the feature-feature and feature-class mutual information both are taken into consideration when evaluate the candidate feature to be selected. The parameter β in (16) is the redundancy coefficient. And different β values may have a great influence on the process of feature selection. When the value of β is low, the feature-class relation plays a major role, as β grows, the influence of feature-feature relation grows. When the value of β is extremely large, the feature-feature relation may be overestimated, and the feature-class relation would be ignored. Some really important features may not be selected, which results in low performance of MIFS. Therefore, we should be caution in choosing the value of β. MIFS-U [20] is designed to overcome the limitations of MIFS. It can better balance the correlation and redundancy of a feature. MIFS-U is defined as:

$$J_{MIFS\_U}(f_i) = I(C; f_i) - \beta \sum\nolimits_{f_j \in S} \frac{I\left(C; f_j\right)}{H\left(f_j\right)} I(f_i; f_j) \tag{17}$$

The minimizing redundancy and maximizing relevance criterion is used frequently in the process of feature selection. Base on such an idea, Peng et al. proposed a new feature selection approach called mRMR [22], which can be combined with other feature selectors such as wrappers to find a very compact subset effectively. MIFS and MIFS-U have the common defect that with the number of select features grows, the redundancy term grows faster than the relevancy term. As a result, some irrelevant features may be selected. The optimal solution depends on the value assigned to β, while optimal $\beta's$ being considered subject to the data structure. There are no specific guidelines on how to determine parameter β. In order to improve this situation, mRMR defines the value of β as the inverse of the number of selected features. And the feature evaluation criterion of mRMR can be defined as:

$$J_{mRMR}(f_i) = I(C; f_i) - \frac{1}{|S|} \sum\nolimits_{f_j \in S} I(f_i; f_j) \tag{18}$$

where |S| is the number of the already selected features. mRMR can prevent the cumulative sum of the redundancy has an excessive value at any size of feature subsets. In fact, mRMR is a specific form of MIFS when the value of β is chosen as 1/|S|.

Estévez et al. studied the feature selection theory based on mutual information and proposed NMIFS [32] algorithm. It uses the average normalized mutual information as a measure of redundancy among features, and it is an enhancement of MIFS, MIFS-U, and mRMR. It is defined as follows:

$$J_{NMIFS}(f_i) = \mathrm{I(C}; f_i) - \frac{1}{|S|} \sum\nolimits_{f_j \in S} NI(f_i; f_j) \tag{19}$$

where

$$NI\left(f_i; f_j\right) = \frac{I(f_i; f_j)}{\min\{H(f_i), H(f_j)\}} \tag{20}$$

MIFS-ND [21] method combines the feature-feature mutual information and the feature-class mutual information to find the optimal feature subset. Unlike MIFS and MIFS-U, the MIFS-ND method does not use feature-class mutual information and feature-feature mutual information directly. Instead, in order to select a strongly relevant but non-redundant feature, it calculates the domination count and dominated count. The first step of MIFS-ND is to sort the features according the feature-class mutual information and feature-feature mutual information, respectively. And naming the feature-class order domination count (Cd) and the feature-feature order dominated count (Fd). And then it selects the feature that has the largest value of (Cd - Fd).

If more than two features have the same (*Cd* - *Fd*) value, the algorithm selects the one that has the highest feature-class mutual information.

The joint mutual information (JMI) [23] method is another feature selection method based on mutual information, and it adopts a new criterion to evaluate the candidate features. JMI chooses the feature that has the maximum cumulative summation of joint mutual information with the selected features in each step and adds it to the subset S until the number of selected features reaches k. JMI is defined as follows:

$$J_{JMI}(f_i) = \sum_{f_j \in S} I(C; f_i, f_j) \tag{21}$$

Conditional mutual information maximization (CMIM) [24] employs the criterion that 'maximizes of the minimum.' By selecting the features that the minimum conditional mutual information is maximum, CMIM ensures the selected features are both individually informative and weakly dependent on each other. CMIM is defined as follows:

$$J_{CMIM}(f_i) = argmax_{f_i \in F-S}(min_{f_j \in S}(I(C; f_i|f_j))) \tag{22}$$

Brown et al. [33] seriously analyzed and studied the mutual information based feature selection approaches and proposed a unified feature selection framework. We can see that most of the existing methods based on information theory can be derived from this framework. Actually, this framework is a linear combination of mutual information. It is defined as follows:

$$J(f_i) = I(C; f_i) - \beta \sum_{f_j \in S} I(f_i; f_j) + \gamma \sum_{f_j \in S} I(f_i; f_j|C) \tag{23}$$

In this formula, it involves three kinds of information: relevance, redundancy, and complementarity. Different values of β and γ could lead to different feature selection methods. For example, $\gamma = 0$ leads to MIFS, $\beta = \gamma = \frac{1}{|S|}$ leads to JMI [33], $\beta = \frac{1}{|S|}$, and $\gamma = 0$ leads to mRMR, while $\beta = \gamma = 0$ leads to MIM. Relevance and redundancy are two contradictory aspects. High relevance usually means high redundancy, and how to balance these two aspects is still an open problem for feature selection methods [27].

Similar to CMIM, Joint Mutual Information Maximization (JMIM) and Normalized Joint Mutual Information Maximization (NJMIM) [13] also use mutual information and the 'maximizes of the minimum' criteria to choose features. They introduce a new objective function to overcome the limitations of some methods, such as overestimation of the feature significance. These two methods are defined as follows:

$$J_{JMIM}(f_i) = argmax_{f_i \in F-S}(min_{f_j \in S}(I(C; f_i, f_j))) \tag{24}$$

$$J_{NJMIM}(f_i) = argmax_{f_i \in F-S}(min_{f_j \in S}(\frac{I(C; f_i, f_j)}{H(C, f_i, f_j)})) \tag{25}$$

In fact, mutual information between (f_i, f_j) and C can be written as follows:

$$I(C; f_i, f_j) = I(C; f_j) + I(C; f_i|f_j) \tag{26}$$

where, $f_j \in S, f_i \in F - S$. When selecting a new feature, for all candidate features, $I(C; f_j)$ is the same. Thus, the CMIM algorithm is actually a variant of algorithm JMIM.

Generally speaking, methods that based on information theory can be divided into two categories according to the criteria of feature evaluation: one is minimizing feature redundancy, and the other is maximizing feature new classification ability. These two categories select the features just from different points of view. The minimizing feature redundancy methods mainly focus on minimizing feature redundancy and do not consider new classification information and vice versa. Thus, it will result in selecting features with high or low both classification information and redundancy. To solve this problem, Gao et al. [34] proposed a hybrid feature selection method that integrates the two types of feature selection methods.

IV. THE PROPOSED METHOD FOR FEATURE SELECTION

The difference between feature extraction and feature filter methods is that the feature filter methods do not generate new features but get an optimal feature subset from the original features according to certain criteria. By doing so, it can not only speed up the data processing but also improve the classification accuracy and reduce the complexity of the learning model. Many existing feature selection methods based on mutual information assume that each feature of the data influences the target variable independently. But in practice, this is not always the case. And in this section, we present a new input feature selection algorithm based on feature set equivalence and mutual information gain maximization.

A. DEFINITIONS OF SOME RELATED CONCEPTS

1) RELEVANCE

Feature f_i is more relevant to the output class C than feature f_j if

$$I(C; f_i) > I(C; f_j) \tag{27}$$

Kohavi and John formalized relevance in terms of an optimal Bayes classifier and categorize relevance into two types: strong relevance and weak relevance [12].

2) REDUNDANCY

The feature f_i is redundant if

$$I(C; f_i, F_S) = I(C; F_S) \tag{28}$$

since

$$I(C; f_i, F_S) = I(C; F_S) + I(C; f_i|F_S) \tag{29}$$

From (28) and (29) we can get

$$I(C; f_i|F_S) = 0 \tag{30}$$

This means that adding f_i to the feature subset F_S will not bring further discriminative information for output class C, therefore, f_i is redundant.

3) COMPLEMENTARY

Feature f_i and f_j are complementary for the output class C if

$$I(C; f_i, f_j) > I(C; f_i) + I(C; f_j) \tag{31}$$

The well-known example to illustrate this phenomenon is the XOR problem. This example is also stated in [5], [14], [25].

TABLE 1. Formula area.

Notation	Area
$H(X_i)$	1,2,4,5
$H(Y)$	2,3,4,6
$H(X_j)$	4,5,6,7
$H(Y\|X_i)$	3,6
$H(Y\|X_i,X_j)$	3
$I(Y;X_i)$	2,4
$I(Y;X_j)$	4,6
$I(Y;X_i,X_j)$	2,4,6
$I(Y;X_i;X_j)$	4

TABLE 2. XOR example data.

f_1	f_2	f_3	f_4	$C=f_2 \oplus f_3$
0	0	0	0	0
1	0	0	1	0
1	1	0	1	1
1	1	0	1	1
1	0	1	1	1
0	0	1	0	1
0	1	1	0	0
0	1	1	0	0

In Table 2, f_1, f_2, f_3, f_4 are binary random values, $f_4 = f_1$, and C is the XOR value of f_2 and f_3, the mutual information between each individual feature (f_2 or f_3) and C is 0,that is $I(C; f_2) = 0$, $I(C; f_3) = 0$. However, when f_2 and f_3 are combined together, the mutual information between (f_2, f_3) and C is 1, which is $I(C; f_2, f_3) = 1$. We can see that the individual mutual information of f_1 and f_4 with C are the top two, that is $I(C; f_1) = I(C; f_4) = 0.1887$, but when selected either of them, then the other one is redundant, that is $I(C; f_1, f_4) = 0.1887$. This example illustrates that, due to the complementarity property between features, when considering a feature alone, it is irrelevant to the target class, but when combined with other features, it may be highly relevant to the target class.

B. DRAWBACKS OF THE EXISTING METHODS

The idea of mutual information based feature selection method is to find the feature subset that has maximum mutual information between it and the output class, and it can be denoted as (14). However, when the size of the data feature is large, exhaustive search the feature space becomes impractical due to the computation complexity. In order to solve this problem, there are two strategies: one is ranking feature strategy, and the other is heuristic search feature strategy.

The ranking feature strategy is to sort the candidate features in descending order according to a certain criterion and selects the top k features. MIM [19] is a typical representative of such type. It ranks candidate features merely according to the mutual information between candidate features and the output class without considering the relationships between features, which would lead to redundancy in many cases. For instance, in Fig. 2, it's obvious that $I(C; f_1) > I(C; f_2) > I(C; f_3)$. Assuming f_1 has been selected, we can see that according to MIM rule the next feature to be selected is f_2. It definitely works well in the scenario I of Fig. 2, because f_1, f_2, f_3 are independent of each other. But for scenario II of Fig. 2, it is more appropriate to select f_3. Because, in addition to the relevant relationship between feature and output class, there is also a relationship between features. And we can see that f_2 is quite redundant when f_1 has been selected.

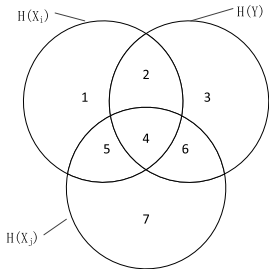

FIGURE 1. Variables relationship Venn diagram.

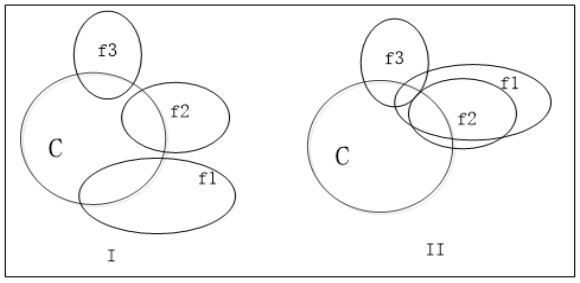

FIGURE 2. Feature redundancy.

The heuristic search strategy adopts the sequential forward selection method to select the candidate features. In each step, it selects the feature that can maximize the feature evaluation criterion. And most of the existing feature selection methods based on mutual information belong to this kind [35], such as JMIM, mRMR, and other methods mentioned in section III. The methods such as MIFS, JMI, mRMR, etc. take the mutual information between the candidate feature and features that have already been selected as the redundancy term of the candidate feature. In most cases, these methods are more effective than the feature ranking methods, but sometimes this will lead to the problem of overestimating candidate features [35], [36]. For instance, when the candidate feature

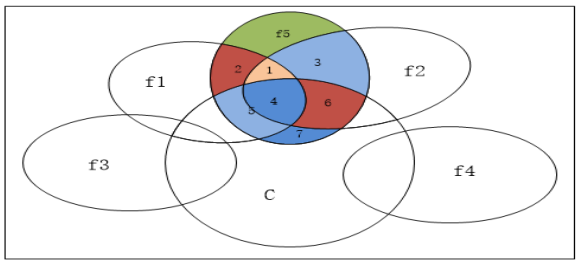

FIGURE 3. Feature overestimation.

is completely correlated to only one or a few selected features but almost independent of other selected features, the redundancy degree will be low by dividing the number of selected features. And it is obviously not in line with the actual situation that the redundancy of this feature is very high. A specific example is shown in Fig. 3. Assuming that f_1, f_2, f_3, and f_4 have been selected, and the importance of candidate feature f_5 will be overestimated. Because f_5 is only correlated with f_1 and f_2, and it is completely independent of f_3 and f_4. It can be seen in Fig. 3 that the mutual information between f_5 and f_1, that is $I(f_5; f_1)$, is the area {1,2,4,5}. And the mutual information between f_5 and f_2, that is, $I(f_5; f_2)$, is the area {1,3,4,6}. and $I(f_5; f_3) = 0$, $I(f_5; f_4) = 0$. So, if average the mutual information between f_5 and $\{f_1, f_2, f_3, f_4\}$ by dividing 4 will obtain a low result.

In addition to the average redundancy strategy, another category of heuristic feature selection approaches adopt the ''maximizes of the minimum'' extreme criterion to select features, such as CMIM [24], JMIM, NJMIM [13], and OLB_CMI [37]. CMIM selects the features that the minimum condition mutual information with the selected feature is max, and JMIM selects the feature that the minimum joint mutual information with the selected feature is max. Both CMIM and JMIM can be derived from JMI and can be viewed as two extreme forms of JMI [38]. Extreme criterion methods produce less redundancy than the methods using average principle and can obtain better results in some special cases. Besides, it's more certain and concrete than the methods based on average principle [36]. However, both average criterion and extreme criterion methods have the same drawback that they can only discover the low order correlations of feature-feature and features-class. Because they just calculate the mutual information between two features and the output class at most. And few studies have been conducted on features with higher-order correlation. In the case such as $I(C; f_1, f_2) = 0, I(C; f_1, f_3) = 0, I(C; f_2, f_3) = 0$, and $I(C; f_1, f_2, f_3) = 1$, when any two of the three features have been selected, to the best of our knowledge, none of the existing methods would assess the third feature as the best one.

In general, the existing methods may easily lead to three problems: 1) feature redundancy; 2) overestimation of features; 3) Inability to deal with the high-order relationship between features very well.

C. THE PROPOSED METHOD

In order to obtain the optimal feature subset $S_{optimal} = Max\ I(C;S)$, where C is the output class. In view of the problems arising from the previous methods, we propose a new method of feature selection based on mutual information, which adopts equivalent partition instead of feature subset and use the mutual information gain maximization (MIGM) criteria to evaluate the candidate feature.

According to (7), we know that $I(C;S) = H(C) - H(C|S)$, where $H(C)$ and $H(C|S)$ are entropy and conditional entropy of C, respectively. $H(C)$ denotes the uncertainty of variable C, and $H(C|S)$ denotes the remaining uncertainty after introducing S. Since $H(C)$ can be regarded as a constant for a specific problem, so in order to get maximum $I(C;S)$, it just needs to select the feature subset S that can makes $H(C|S)$ minimum.

Assuming F is the full feature subset of the data, and each feature $f_i \in F$ can be viewed as a partition of C, which divides C into several parts, and $H(C|f_i)$ is the average uncertainty of each part. If two features f_i and f_j are introduced, it can divide C into equal or more parts than any single feature. So $\{f_i, f_j\}$ can be viewed as a new partition of C, and C can be divided into many parts by the combinations of f_i and f_j, and $H(C|f_i, f_j)$ is the average uncertainty of each part. Obviously, the value of $H(C|\{f_1, f_2, \ldots\})$, where $f_1, f_2, \ldots \in S$, decreases monotonously with the increase of the select number in S. In other words, by introducing more features and using the combination of them to divide C can make the average purity of each part improved. MIGM is a heuristic sequential forward feature selection method, and if more than one feature to be selected, it selects those features that can form the partition to make C more purity, that is, $H(C|\{f_i, f_j, \ldots\})$ is minimum. It can be inferred from (7) that the minimum of $H(C|\{f_i, f_j, \ldots\})$ is equivalent to the maximum of $I(C;\{f_i, f_j, \ldots\})$. For the convenience of calculation, we use the latter form in the following discussion.

Assuming $S = \{f_1, f_2, \ldots, f_{k-1}\}$ is selected, and $\hat{s} = S \cup f_i$, where f_i is a candidate feature to be evaluated next. Instead of calculating mutual information between candidate feature f_i and each feature in the selected feature set S to obtain redundant information, we view S as a new partition SN and directly calculate the joint mutual information of $\{SN, f_i\}$ with output class C, and it is defined as follows:

$$J_{MIGM}(f_i) = argMax\ I(C; SN, f_i) \tag{32}$$

where $f_i \in F - S$. The problems of feature redundancy and complementarity encountered in the previous methods can be resolved very well by replacing the selected feature set with equivalent partition. For instance, assuming f_1 and f_2 have been selected, as shown in Fig. 4-I, then we try to determine whether f_3 or f_4 will be selected next. As shown in Fig. 4-II. We take $\{f_1, f_2\}$ as a whole and call it equivalent partition, which is denoted as SN. Then calculate $I(C;SN,f_3)$ and $I(C;SN,f_4)$ and choose the one that has bigger value. As can be seen from Fig. 4-II, f_3 will be chosen. Subsequently, using $\{SN,f_3\}$ to form new equivalent partition for the next iteration.

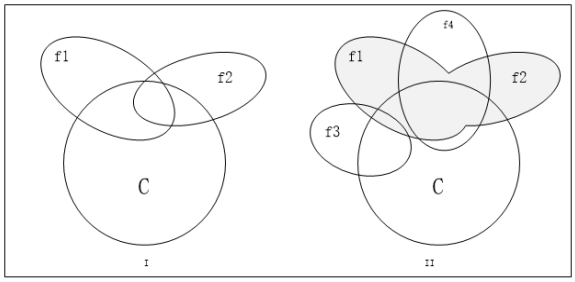

FIGURE 4. Equivalent partition.

In order to get the equivalent partition SN, we just need to give a unique label for each combination. For example, when features f_1 and f_2 in table 2 have been selected, in order to evaluate feature f_3 and f_4 we have to form the equivalent partition for $\{f_1, f_2\}$. Due to the combinations of f_1 and f_2 are {(0,0),(1,0),(1,1),(0,1)}, and we can give a unique label for each combination, such as 1 for (0,0), 2 for (1,0), 3 for (1,1) and 4 for (0,1), then we can get $SN = [1,2,3,3,2,1,4,4]'$. It is easy to verify that equation $I(C;SN) = I(C;f_1, f_2)$ holds. Here we call SN the equivalent partition of feature set $\{f_1, f_2\}$. Because $I(C;SN,f_3) > I(C;SN,f_4)$, so f_3 will be selected subsequently. Though $I(C; f_3) = 0$ *and* $I(C;f_4) = 0.1887$, which mean f_4 is more relevant to C than f_3, but f_4 is redundant for $\{f_1, f_2\}$, while f_3 and $\{f_1, f_2\}$ are complement to each other. When f_3 was selected, then using the combination of $\{SN,f_3\}$ to update SN for the next iteration, and so on.

The details of MIGM are presented in Algorithm 1.

The pseudo code presented above is used to describe the feature selection process of our proposed method. It consists of three key stages. The first stage is to initialize the parameters and calculate the mutual information between each feature and output class (line 1-8). The second stage is to select the first feature, which has the maximum mutual information with the output class C (line 9-13). The third stage is to select the features that have maximum joint mutual information with output class C after combining with equivalent partition and then forming the new equivalent partition. It repeats the third stage until the number of selected features reaches K.

Algorithm 1 MIGM

Input:
A dataset D with a full feature set $F = \{f_1, f_2, \ldots, f_N\}$ and the output class label C and the user-specified feature number threshold K;
Output:
The selected feature subset S
1: $S \leftarrow \emptyset$;
2: $k \leftarrow 0$;
3: defines SN to store the equivalent partition of the selected features;
4: define MI_F2C_array[] to store mutual information between features and class;
5: for $i = 1$ to N do
6: calculate the mutual information $I(C; f_i)$
7: Store $I(C; f_i)$ in MI_F2C_array[]
8: end for
9: select the feature f_i with max value in MI_F2C_array[]
10: $S = S \cup f_i$;
11: $F = F - f_i$;
12: $k = k + 1$;
13: $SN = f_i$;
14: while $k < K$
15 set CombMI_array[ ] to store joint mutual information
16: for each candidate feature $f_i \in F$ do
17: calculate joint mutual information $I(C;SN,f_i)$
18: store $I(C;SN,f_i)$ in CombMI_array[ ]
19: end for
20: select the feature f_i with max value in CombMI_array[ ];
21: $S = S \cup f_i$;
22: $F = F - f_i$;
23: $k = k + 1$;
24 update SN according to previous SN and f_i
25: end while
26: Output S

D. COMPLEXITY ANALYSIS

In this section, we analyze the time complexity and space complexity of the proposed algorithm. Given an input dataset D that has M instances and N features, and the number of features to be selected is k. Among the feature selection methods mentioned in section III, MIM only needs to calculate the mutual information between output class and each feature and then sorts them in descending order. So the time complexity of MIM is relatively small, that is $O(MN)$. The time complexity of MIFS_U, mRMR, JMIM, and other similar methods are $O(kMN)$. The process of feature selection in MIGM is similar to JMIM, for each candidate feature it has to calculate the mutual information between C and $\{SN,f_i\}$, where $f_i \in F$ and the max number of F is N, the calculate time complexity of selecting one candidate feature is $O(MN)$, and there are k features to be selected, so the total time complexity of MIGM is $O(kMN)$. In order to speed up the computation, MIGM uses an array to store the mutual information between features and class, which is the main space consumption of the algorithm, so its space complexity is $O(N)$.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, a series of experiments were conducted to test and analyze the performance of the proposed method.

A. EXPERIMENT SETUP

In order to illustrate the performance of the proposed method, we compare it with six baseline methods mRMR, MIFS_U, NMIFS, JMIM, NJMIM, and MIM on ten benchmark data sets. These data sets were picked from different application

domains, and the size of instances varies from 178 to 20000, and the number of features varies from 13 to 857, and the number of classes varies from 2 to 26. The characteristics of these data sets are depicted in Table 3.

TABLE 3. Experimental data sets description.

No.	Datasets	Instances	Features	Classes	Area
1	Wine	178	13	3	Physical
2	wdbc	569	30	2	Life
3	CTGs	2126	21	3	Life
4	Parkinsons	195	22	2	Life
5	biodeg	1055	41	2	chemical
6	letterRecognition	20000	16	26	Computer
7	CNAE9	1080	857	9	Business
8	Spam	4601	57	2	Computer
9	Semeion	1593	256	9	Computer
10	Vehicle	846	18	4	Computer

1) Wine database: It consists of the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The wines are divided into three types, and this data set records the quantities of 13 constituents found in each of them.

2) wdbc dataset: it consists of 569 cases of cell biopsies, and each of which has 30 feature values and a class label of the diagnosis result that use character M represents malignant, and character B represents benign. The values of 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, and they describe characteristics of the cell nuclei present in the image.

3) CTGs dataset: it consists of 2126 fetal cardiotocograms (CTGs), which were automatically processed and the respective 21 diagnostic features measured. The CTGs were classified into three types by the expert obstetricians, and one of the classification labels was assigned to each of them.

4) Parkinsons dataset: It is consists of 195 voice records from 31 people include 23 with Parkinson's disease (PD), and each record was described by 22 variables. The main aim of this data set is to discriminate healthy people from those with PD, according to the "status" column, which is set to 0 for healthy and 1 for PD.

5) biodeg dataset: it consists of 1055 chemicals, each was described by 41 molecular descriptors, and these chemicals were classified into 2 classes, ready biodegradable (RB) and not ready biodegradable (NRB).

6) letterrecognition dataset: it consists of 20,000 black-and-white rectangular pixel character images, and each image is one of the 26 capital letters in the English alphabet and was converted into 16 primitive numerical attributes.

7) CNAE9 dataset: it consists of 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories.

8) Spam dataset: it consists of 4061 e-mails classified into two categories: Spam or Non-Spam. The attributes of the dataset indicate whether a particular word or character was frequently occurring in the e-mail.

9) Semeion dataset: it consists of 1593 handwritten digits from around 80 persons, each digit was stretched in a rectangular box 16x16 in a gray scale of 256 values. Then each pixel of each image was scaled into a Boolean (1/0) value using a fixed threshold.

10) Vehicle dataset: it consists of 846 instances, and each instance is the silhouettes of the vehicle 3D objects within 2D images, and the number of features is 18. Its purpose is to classify a given silhouette as one of four types of vehicles.

All of the above ten data sets are available from the UCI Machine Learning Repository, and they are also used in other literature [32], [36], [39]. Because the mutual information based feature selection methods need to estimate the probability distribution for calculating the entropy and the joint entropy of variables, therefore, we discretize continuous features into ten bins using equal-width discretization, and then use the discretized data in feature selection process, and such method also was used in other literature [25], [38]. Since the proposed method is a filter method and its efficiency might differ from one classifier to another. In order to reduce the bias of a specific classifier and test the robustness of the proposed method, we adopt the average classification accuracy of two different and common used classifiers, k-Nearest Neighbor (KNN) and Naïve-Bayes (NB), as the measurement. Furthermore, for the purpose of better evaluate the performance of the proposed method, we perform KNN on the original data and perform NB on the discretized data. Both these two classifiers are provided by MATLAB R2016a, and the k value of KNN is set to 3 [39].

Experiments were carried out on a desktop PC with 4 GB main memory, 2.4GHz Intel(R) Core(TM) i5-6200U processor and 64-bit Windows 7 operating system. All algorithms in our experiments are implemented in MATLAB R2016a. The Five-fold cross-validation is employed in our experiments, and each data set was divided into five parts randomly, and run five times on the data. In each run, one part (20%) is used for the test, and the other four parts (80%) are used as training data. And the average classification accuracy of five runs is used as the accuracy of the correspondent method.

B. RESULTS AND ANALYSIS

The average classification accuracy of KNN and NB across the comparing feature subsets are recorded in Table 4. Besides, a paired two-tailed t-test is conducted between MIGM and other methods. The notations "+" / "−" / "= " indicate the statistically significant (at 5%) that our method's wins/losses/equals over other methods. The last row (W/T/L) indicates that the number of the data sets our method has higher (or equal, lower) accuracy than comparison methods. The bold font indicates the maximal value of the row.

Table 4 shows that the proposed method MIGM achieves the highest average accuracy on two different classifiers. It can be observed from Table 4 that the overall average accuracy of MIGM over the ten datasets is 75.37%, and it outperforms the other six feature selection methods obviously. Meanwhile, MIGM obtains the highest average accuracy on

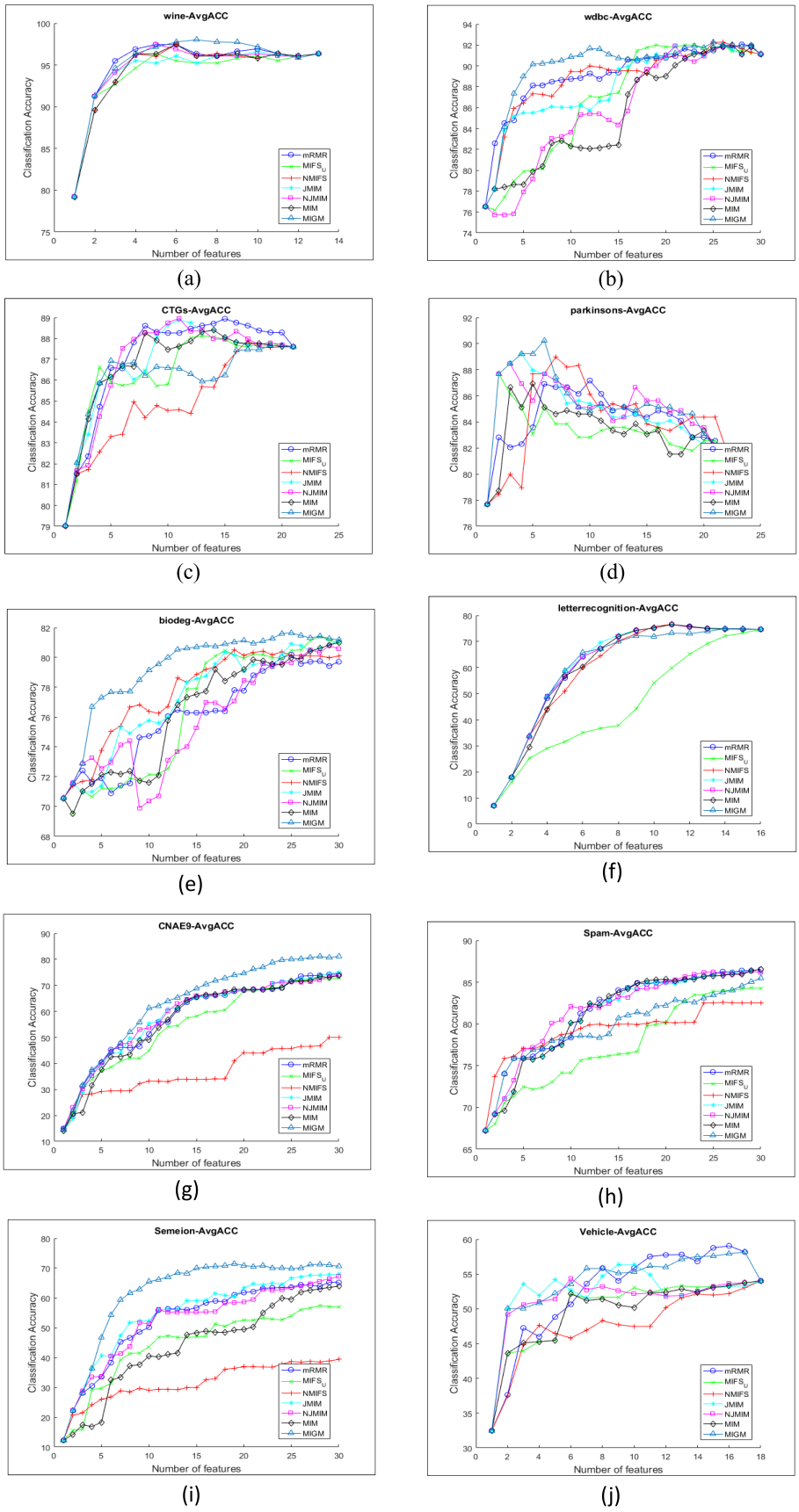

FIGURE 5. Average classification accuracy of KNN classifier and NB classifier on ten data sets.

TABLE 4. Average classification accuracy (mean ± std.) of two classifiers (KNN and NB) on ten data sets.

Data sets	mRMR	MIFS_U	NMIFS	JMIM	NJMIM	MIM	MIGM
wine	94.81±0.45(+)	93.84±0.6(+)	94.21±0.27(+)	94.15±0.07(+)	94.47±0.57(+)	94.23±0.06(+)	**95.05±0.59**
wdbc	89.06±10.07(+)	86.81±11.82(+)	88.76±9.59(+)	88.02±10.69(+)	85.89±12.23(+)	85.51±13.68(+)	**89.85±8.92**
CTGs	**86.99±4.12(−)**	86.32±5.18(=)	84.89±4.91(+)	86.7±5.35(−)	86.77±4.35(−)	86.68±4.86(−)	86.08±5.91
parkinsons	84.13±7.58(+)	83.26±9.26(+)	84.2±9.4(=)	84.77±8.82(+)	85.05±8.29(=)	83.31±8.9(+)	**85.37±9.81**
biodeg	76.07±9.6(+)	76.4±7.1(+)	77.65±6.5(+)	77.24±7.68(+)	75.81±9.05(+)	76.32±7.6(+)	**79.3±6.13**
letterrecognition	60.46±24.71(=)	45.7±23.65(+)	59.32±25.8(=)	**60.85±25.29(−)**	60.53±25(=)	59.75±24.78(=)	59.89±26.75
CNAE9	57.67±1.15(+)	54.75±1.96(+)	36.42±3.06(+)	58.06±0.39(+)	58.25±0.34(+)	56.71±2.12(+)	**63.15±0.86**
Spam	81.5±8.99(−)	77.51±11.19(+)	79.27±12.61(=)	81.12±9.12(−)	**81.6±8.88(−)**	81.18±8.7(−)	79.72±10.64
Semeion	52.5±1.71(+)	44.6±1.66(+)	31.72±3.54(+)	54.67±2.03(+)	52.06±2.62(+)	44.05±3.38(+)	**61.64±2.78**
Vehicle	52.33±10.99(+)	49.84±14.6(+)	47.65±13.52(+)	52.13±14.03(+)	51.27±14.19(+)	49.52±14.86(+)	**53.65±14.38**
Average	73.55	69.9	68.41	73.77	73.17	71.73	**75.37**
W/T/L	7/1/2	9/1/0	7/3/0	7/0/3	6/2/2	7/1/2	

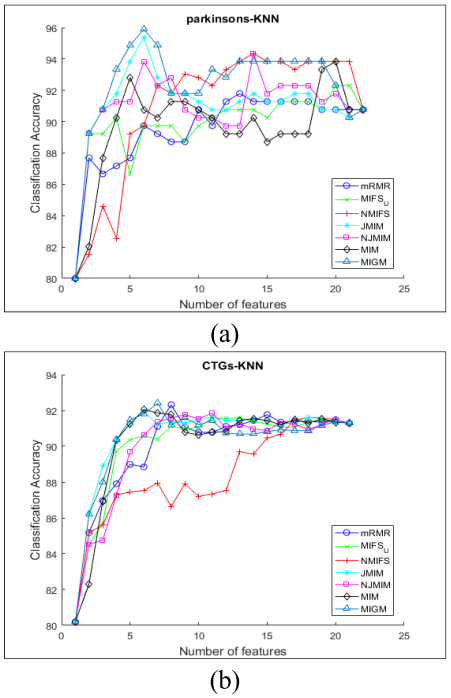

FIGURE 6. Average classification accuracy of KNN classifier on parkinsons and CTGs data sets.

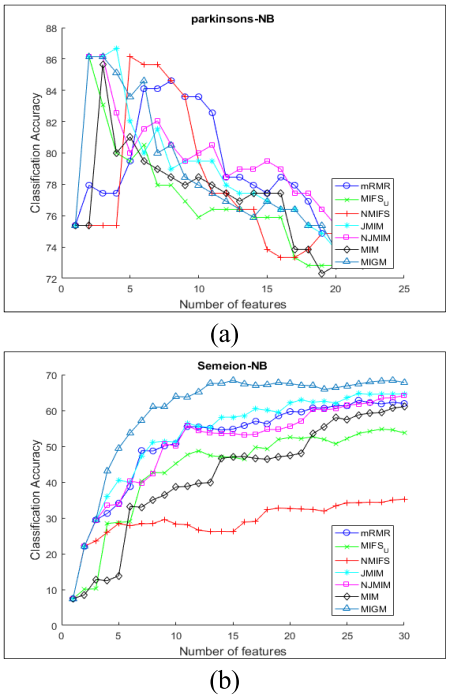

FIGURE 7. Average classification accuracy of NB classifier on parkinsons and semeion.

seven out of ten data sets. And the detailed results are shown in Fig. 5(a)-(j). Here, the X-axis represents the number of selected features, i.e., k, which increases from 1 to 30 at intervals of 1, and the Y-axis represents the average accuracy of the two classifiers in the corresponding k selected features. Note that, if the number of original features of the data set is less than 30, then k reaches up to the number of original features.

As can be seen from Fig. 5, classification accuracy exhibits different variations across different datasets. On some data sets, such as Parkinsons and CTGs, the classifiers achieve their best performance only with a few features, and it will not be improved by monotonously expanding the number of features, and sometimes it may be weakened, such as in the case of Parkinsons in Fig. 7(a). For this kind of data set, the convergence state can be satisfied within a small range of features and adopts the wrapper model to select features is also a good choice. On the contrary, in other data sets, such as CNAE9 and Spam in Fig. 5, with the increase of the selected features, the classification accuracy shows an obvious upward trend. Therefore, how to determine the optimal size of the feature subset is still an open problem for the filter selection method, which adopts the individual feature evaluation strategy.

In general, the proposed method MIGM outperforms or comparable to other mutual information based feature selection methods in its ability to select discriminative features in most data sets, as illustrated in Fig. 5. Among them, MIGM achieves the best average classification accuracy on data sets wdbc, biodeg, CNAE9, and Semeion, and secondly on data sets wine, Parkinsons, and Vehicle. On data set letterrecogni-

TABLE 5. Average classification accuracy (mean ± std.) of KNN on ten data sets.

Data sets	mRMR	MIFS_U	NMIFS	JMIM	NJMIM	MIM	MIGM
wine	94.49±5.11(=)	93.42±4.76(+)	94.02±5.1(=)	94.21±4.83(=)	94.07±4.76(=)	94.19±5.18(=)	**94.64±5.03**
wdbc	**96.19±1.26(=)**	95.17±1.72(+)	95.54±1.56(+)	95.58±1.38(+)	94.54±2.43(+)	95.19±1.47(+)	96.15±1.27
CTGs	89.9±2.86(=)	89.98±2.91(=)	88.37±2.79(+)	**90.48±2.67(−)**	89.85±3.07(=)	90.11±3.14(=)	90.25±2.66
parkinsons	89.49±2.62(+)	89.81±2.51(+)	90.85±4.46(+)	91±2.76(+)	90.91±2.76(+)	89.6±3.16(+)	**92.31±3.2**
biodeg	82.86±2.5(+)	81.42±3.99(+)	82.25±2.57(+)	82.67±2.32(+)	82.21±2.95(+)	81.69±3.37(+)	**83.64±2.22**
letterrecognition	77.93±27.27(+)	62.42±27.55(+)	77.56±26.85(+)	78.73±27.31(=)	78.21±27.2(=)	77.27±27.68(+)	**78.81±26.5**
CNAE9	58.48±14.63(+)	56.14±14.8(+)	38.59±8.71(+)	58.34±15.68(+)	58.01±15.13(+)	58.21±15.48(+)	**63.76±17.22**
Spam	87.85±3.93(−)	85.42±3.65(+)	**88.18±3.63(−)**	87.57±3.94(−)	87.88±3.76(−)	87.33±4.08(=)	87.24±3.56
Semeion	53.71±15.1(+)	45.77±12.09(+)	34.22±8.28(+)	56.11±14.72(+)	53.91±14.44(+)	46.44±15.25(+)	**63.6±17.25**
Vehicle	60.11±7.03(+)	60.16±6.79(+)	57.21±6.3(+)	62.06±5.78(+)	61.3±5.93(+)	60.04±6.65(+)	**63.81±6.57**
Average	79.1	75.97	74.68	79.67	79.09	78.01	**81.42**
W/T/L	6/3/1	9/1/0	8/1/1	6/2/2	6/3/1	7/3/0	

TABLE 6. Average classification accuracy (mean ± std.) of NB on ten data sets.

Data sets	mRMR	MIFS_U	NMIFS	JMIM	NJMIM	MIM	MIGM
wine	95.14±4.84(+)	94.27±4.57(+)	94.4±4.86(+)	94.1±4.67(+)	94.88±4.99(+)	94.27±4.81(+)	**95.46±5.24**
wdbc	81.94±5.56(+)	78.45±9.82(+)	81.98±6.15(+)	80.46±6.6(+)	77.24±8.99(+)	75.84±9.37(+)	**83.54±6.45**
CTGs	**84.07±2.82(−)**	82.65±2.16(−)	81.42±2.24(=)	82.92±2.68(−)	83.69±2.44(−)	83.24±1.99(−)	81.9±1.77
parkinsons	78.76±3.52(=)	76.71±3.5(+)	77.55±4.49(=)	78.53±4.05(=)	**79.18±3.42(=)**	77.02±3.18(+)	78.44±4.15
biodeg	69.28±4.92(+)	71.38±5.28(+)	73.05±3.96(+)	71.81±5.14(+)	69.41±5.21(+)	70.94±5(+)	**74.97±3.95**
letterrecognition	**42.98±17.36(−)**	28.98±16.06(+)	41.08±18.08(=)	42.97±17.41(−)	42.86±17.35(−)	42.23±17.92(−)	40.97±17.05
CNAE9	56.86±18.09(+)	53.36±18.33(+)	34.26±9.41(+)	57.78±17.96(+)	58.49±17.08(+)	55.2±19.61(+)	**62.54±20.4**
Spam	75.14±7.12(−)	69.59±7.52(+)	70.36±3.2(+)	74.67±7.65(−)	**75.32±7.37(−)**	75.03±7.85(−)	72.2±5.59
Semeion	51.29±13.7(+)	43.42±13.69(+)	29.22±5.44(+)	53.23±14.06(+)	50.21±13.59(+)	41.66±16.3(+)	**59.67±15.11**
Vehicle	**44.56±8.33(=)**	39.51±4.64(+)	38.08±4.81(+)	42.21±5.36(=)	41.24±4.85(+)	39.01±4.39(+)	43.48±5.53
Average	68	63.83	62.14	67.87	67.25	65.44	**69.32**
W/T/L	5/2/3	9/0/1	7/3/0	5/2/3	6/1/3	7/0/3	

tion, the performance of MIGM is very close to other methods, and only a little worse than the best method JMIM. It is worth noting that although the average classification effect of MIGM is not as good as that of other feature selection on data set CTGs, as shown in Figure 5(c), it achieves very good results in KNN classifier, as shown in Fig. 6(b). And it selects the most informative feature subset very soon, which is similar to dataset Parkinsons, as shown in Fig. 6(a) and Fig. 7(a).

Since the relationship between features is complicated, so it is impossible to design a general feature evaluation criterion that could find out the best k features in every step on all datasets. The feature that is selected according to a certain criterion may irrelevant and even redundant for the classifier at one moment, but when some other features are added later, they could become very informative. That is the reason why the accuracy curves do not increase or decrease steadily on some datasets, such as shown in fig 5(b) and fig 5(c).

VI. CONCLUSION

This paper presents a new feature selection method based on information theory named mutual information gain maximization, which adopts equivalent partition of the feature subset strategy to simplify and enhance the feature selection process. It views each feature of the data set as a discriminative partition for the classification task and uses a heuristic sequential forward search strategy to select the informative features. The main difference between the proposed method and existing ones is that excepts for the first selected feature, MIGM no longer considers the individual discriminative ability of a candidate feature but takes the joint discriminative ability of the candidate feature and the equivalent partition as the index. So it is not necessary to determine parameters to make the trade-off between feature relevance and feature redundancy when evaluating the candidate feature. And that is because the feature redundancy, relevance, as well as complementarity problem can be well resolved by forming the equivalent partition. It was compared with six classical mutual information based feature selection methods on ten well-known benchmark data sets from different application areas. The results demonstrate that our method could identify an effective feature subset that could lead to better classification results than other methods. And the experimental results also show that our method performs better in KNN classifiers than NB, especially in discrete and sparse data sets, such as CNAE9 and Semeion.

It also can be seen from the results that the characteristics of data sets have many impacts on the effectiveness of feature selection methods, so it is difficult to design a general method that can achieve better results than other methods on various data sets. For example, as shown in Fig. 5 (f) and Fig. 5 (g), even the simple feature ranking method MIM has better results than MIFS_U and NMIFS on dataset Letterrecognition and CNAE9, respectively. In future work, we will mainly focus on how to quickly detect the characteristics of the datasets so as to select the most suitable method for application.

ACKNOWLEDGMENTS

The authors would like to thank the editor and anonymous reviewers for their valuable comments and great effort in handling this paper.

REFERENCES

[1] A. K. Shukla, P. Singh, and M. Vardhan, "A two-stage gene selection method for biomarker discovery from microarray data for cancer classification," *Chemometrics Intell. Lab. Syst.*, vol. 183, pp. 47–58, Dec. 2018.

[2] Y. Yi, W. Zhou, Q. Liu, G. Luo, J. Wang, Y. Fang, and C. Zheng, "Ordinal preserving matrix factorization for unsupervised feature selection," *Signal Process., Image Commun.*, vol. 67, pp. 118–131, Sep. 2018.

[3] P. Marti-Puig, A. Blanco-M, J. J. Cárdenas, J. Cusidó, and J. Solé-Casals, "Feature selection algorithms for wind turbine failure prediction," *Energies*, vol. 12, no. 3, p. 453, Jan. 2019.

[4] D. You, X. Wu, L. Shen, S. Deng, Z. Chen, C. Ma, and Q. Lian, "Online feature selection for streaming features using self-adaption sliding-window sampling," *IEEE Access*, vol. 7, pp. 16088–16100, 2019.

[5] X. Tang, Y. Dai, P. Sun, and S. Meng, "Interaction-based feature selection using factorial design," *Neurocomputing*, vol. 281, pp. 47–54, Mar. 2018.

[6] H. Yuan, J. Li, L. L. Lai, and Y. Y. Tang, "Joint sparse matrix regression and nonnegative spectral analysis for two-dimensional unsupervised feature selection," *Pattern Recognit.*, vol. 89, pp. 119–133, May 2019.

[7] I. Guyon and A. Elissee, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[8] P. Pudil and P. Somol, "Identifying the most informative variables for decision-making problems—A survey of recent approaches and accompanying problems," *Acta Oeconomica Pragensia*, vol. 2008, no. 4, pp. 37–55, 2008.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[11] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[13] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Syst. Appl.*, vol. 42, pp. 8520–8532, Sep. 2015.

[14] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.

[15] M. Antonelli, P. Ducange, F. Marcelloni, and A. Segatori, "On the influence of feature selection in fuzzy rule-based regression model generation," *Inf. Sci.*, vol. 329, pp. 649–669, Feb. 2016.

[16] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[17] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[18] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature selection considering two types of feature relevancy and feature interdependency," *Expert Syst. Appl.*, vol. 93, pp. 423–434, Mar. 2018.

[19] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proc. Workshop Speech Natural Lang.*, Feb. 1992, pp. 212–217.

[20] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[21] N. Hoque, D. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6371–6385, 2014.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005. [Online]. Available: http://ieeexplore.ieee.org/document/1453511/

[23] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Proc. Adv. Neural Inf. Process. Syst.*, S. A. Solla, T. K. Leen, and K. R. Müller, Eds., vol. 12, 2000, pp. 687–693.

[24] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.

[25] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, Jun. 2008.

[26] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissestation, Comput. Inf. Sci., Univ. Ljubljana, Ljubljana, Slovenia, 2005, pp. 1–252.

[27] J. Wang, J. Wei, Z. Yang, and S. Wang, "Feature selection by maximizing independent classification information," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 828–841, Apr. 2017.

[28] S. Singha and P. P. Shenoy, "An adaptive heuristic for feature selection based on complementarity," *Mach. Learn.*, vol. 107, no. 12, pp. 2027–2071, Dec. 2018.

[29] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Applications of Evolutionary Computing*, F. Rothlauf, Ed. Berlin, Germany: Springer-Verlag, 2006, pp. 91–102.

[30] H.-Y. Lin, "Reduced gene subset selection based on discrimination power boosting for molecular classification," *Knowl.-Based Syst.*, vol. 142, pp. 181–191, Feb. 2018.

[31] C. Orsenigo and C. Vercellis, "Multivariate classification trees based on minimum features discrete support vector machines," *IMA J. Manage. Math.*, vol. 14, no. 3, pp. 221–234, Jul. 2003.

[32] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[33] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, Jun. 2012.

[34] W. Gao, L. Hu, P. Zhang, and F. Wang, "Feature selection by integrating two groups of feature evaluation criteria," *Expert Syst. Appl.*, vol. 110, pp. 11–19, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417418303294

[35] C. Yin, H. Zhang, R. Zhang, Z. Zeng, X. Qi, and Y. Feng, "Feature selection by computing mutual information based on partitions," *IEICE Trans. Inf. Syst.*, vol. E101D, no. 2, pp. 437–446, 2018.

[36] H. Zhou, Y. Zhang, Y. Zhang, and H. Liu, "Feature selection based on conditional mutual information: Minimum conditional relevance and minimum conditional redundancy," *Appl. Intell.*, vol. 49, no. 3, pp. 883–896, Mar. 2019.

[37] H. Peng and Y. Fan, "Feature selection by optimizing a lower bound of conditional mutual information," *Inf. Sci.*, vols. 418–419, pp. 652–667, Dec. 2017.

[38] W. Gao, L. Hu, P. Zhang, and J. He, "Feature selection considering the composition of feature relevancy," *Pattern Recognit. Lett.*, vol. 112, pp. 70–74, Sep. 2018.

[39] S. Lee, L. T. Vinh, Y.-T. Park, and B. J. d'Auriol, "A novel feature selection method based on normalized mutual information," *Appl. Intell.*, vol. 37, no. 1, pp. 100–120, Jul. 2012.

XINZHENG WANG received the B.S. and M.S. degrees from the Guilin University of Technology, China, in 2004 and 2007, respectively, and the Ph.D. degree with the College of Computer Science, Sichuan University. He is currently a Lecturer with the Guilin University of Technology. His current research interests include personal big data, data mining, and data visualization.

BING GUO received the B.S. degree in computer science from the Beijing Institute of Technology, China, in 1991, and the M.S. and Ph.D. degrees in computer science from the University of Electronic Science and Technology of China, China, in 1999 and 2002, respectively. He is currently a Professor with the College of Computer Science, Sichuan University, China. His current research interests include embedded real-time system and green computing.

YAN SHEN received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2004. She is currently a Professor with the Chengdu University of Information Technology. Her research interest includes big data analysis.

CHIMIN ZHOU received the B.S. degree from the University of Electronic Science and Technology of China, in 2004, and the M.S. degree from Xihua University, in 2006. He is currently pursuing the Ph.D. degree with Sichuan University. He is currently an Associate Professor with the Sichuan Radio and TV University. His research interests include formalization of block chains and data science.

XULIANG DUAN was born in 1982. He received the M.S. degree from Beijing Forestry University, in 2008, and the Ph.D. degree from the College of Computer Science, Sichuan University. He is currently an Associate Professor with the Sichuan University. His current research interests include personal big data, big data cleaning, and big data governance. He is also a Student Member of CCF.