# Conditional Mutual Information-Based Feature Selection Analyzing for Synergy and Redundancy

Hongrong Cheng, Zhiguang Qin, Chaosheng Feng, Yong Wang, and Fagen Li

**Battiti's mutual information feature selector (MIFS) and its variant algorithms are used for many classification applications. Since they ignore feature synergy, MIFS and its variants may cause a big bias when features are combined to cooperate together. Besides, MIFS and its variants estimate feature redundancy regardless of the corresponding classification task. In this paper, we propose an automated greedy feature selection algorithm called conditional mutual information-based feature selection (CMIFS). Based on the link between interaction information and conditional mutual information, CMIFS takes account of both redundancy and synergy interactions of features and identifies discriminative features. In addition, CMIFS combines feature redundancy evaluation with classification tasks. It can decrease the probability of mistaking important features as redundant features in searching process. The experimental results show that CMIFS can achieve higher best-classification-accuracy than MIFS and its variants, with the same or less (nearly 50%) number of features.**

**Keywords: Classification, feature selection, conditional mutual information, redundancy, interaction.**

Manuscript received Apr. 20, 2010; revised June 13, 2010; accepted June 28, 2010.

This work was supported by National High Technology Research and Development Program of China (863 Program) under Grant 2006AA01Z411 and 2009AA01Z422.

Hongrong Cheng (phone: +86 28 61831864, email: hrcheng@uestc.edu.cn), Zhiguang Qin (email: qinzg@uestc.edu.cn), Chaosheng Feng (email: csfenggy@126.com), Yong Wang (email: cla@uestc.edu.cn), and Fagen Li (email: fagenli@uestc.edu.cn) are with the Department of Computer Science, University of Electronic Science and Technology, Sichuan, China.

doi:10.4218/etrij.11.0110.0237

## I. Introduction

Feature selection plays an important role in improving accuracy, efficiency, and scalability of the classification process. Since relevant features are often unknown a priori in the real world, irrelevant and redundant features are introduced to represent the domain. However, more features will significantly slow down the learning process and lead to classification over-fitting. With a limited amount of sample data, irrelevant features may obscure the distributions of a small set of truly relevant features for the task and confuse the learning algorithms. It has been proven in both theoretical and empirical aspects that reducing the number of irrelevant or redundant features drastically increases the learning efficiency of algorithms and yields more general concepts for a better insight into the classification tasks.

In supervised classification learning, one is given a training set of labeled instances. An instance is typically described as an assignment of attribute values to a set of features $F$, and each instance is associated with one of $l$ possible classes in $C = \{c_1, \ldots, c_l\}$. The feature selection can be formalized by selecting a minimum subset $S$ from the original feature set $F$ such that $P(C|S)$ is as close as possible to $P(C|F)$, where $P(C|S)$ and $P(C|F)$ are the approximate conditional probability distribution given the training set [1]. The minimum subset $S$ is called an optimal subset. To find the best subset, the order of the search space is $O(2^n)$, where $n$ is the original number of features [2]. In practice, it is hard to search the feature subspace exhaustively because it is a huge number even for medium-sized $n$. A lot of problems related to feature selection are shown to be NP-hard [3]. Alternatively, many sequential-search-based approximation schemes have been proposed. In general, these methods can be grouped into two categories [4]: filters and

  

wrappers. Filter methods select subset of features based on a predefined measure which is independent of the subsequent learning algorithm. Wrappers utilize the learning machine performance to evaluate the goodness of feature subset. In spite of the better performance, wrappers can be computationally expensive and have a risk of over-fitting to their learning model. Therefore, our work in this paper focuses on the filter methods.

In filters, the evaluating measures take a crucial role in detecting feature relevance and redundancy. Information measures, such as entropy and mutual information (MI), have been widely used for feature selection [5]-[10] because of its natural advantage. Fano [11] has revealed that maximizing the MI between the features and the concept target can achieve a lower bound to the probability of error. However, directly computing Shannon's MI between high-dimensional vectors is impractical because of the limited number of samples and high computation cost. Alternatively, the evaluation of high-dimensional MI is simplified by evaluating several low-dimensional MI terms. The typical approximate criterion is Battiti's mutual information feature selector (MIFS) [5]. Instead of calculating the joint MI between the selected feature set and the class variables, MIFS evaluates MI between individual features and class labels, and selects those features that have maximum MI with class labels but less quantity proportional to the accumulated MI with the previously selected features. Kwak and Choi [6] improve MIFS in their mutual information feature selector under uniform information distribution algorithm (MIFS-U) on the assumption that the information of input features is distributed uniformly, that is, the ratio of $H(f_s)$ to $I(f_s;f_i)$ is equal to the ratio of $H(f_s|C)$ to $I(f_s; f_i|C)$, where $f_s$ and $f_i$ represent the selected feature and the candidate feature, respectively. Since MIFS-U as well as MIFS suffers from the difficulty of estimating appropriately the redundancy penalization parameter $\beta$ in the algorithms, the quadratic mutual information feature selector (QMIFS) [7] and the modified MIFS-U algorithm (mMIFS-U) [10] are presented to overcome this limitation. However, in many classification problems, the information of the features does not satisfy the assumption of uniform probability distribution. Peng and others [8] propose a variant of MIFS, the min-redundancy max-relevance (mRMR) criterion, for arbitrary feature distribution. The authors show that for first-order forward search, that is, when one feature is selected at a time, the mRMR criterion is equivalent to max-dependency. Considering the MI bias toward multi-valued features, the normalized mutual information feature selection (NMIFS) [9] is proposed to further enhance MIFS, MIFS-U, and mRMR criterions.

Although the variant algorithms described above effectively improve classification quality of MIFS, their performance may degrade because of the following two reasons. First, they all ignore the case of feature cooperation and suppose all features are competitive. Second, in their criterions, the feature redundancy is evaluated regardless of the classification problem at hand. However, if the redundant information between two important features is rarely relative to the corresponding target concept, even when they are highly redundant, neither of them should be ignored.

In this paper, we propose a novel criterion based on conditional mutual information to select promising features. This criterion considers not only the competition among features but also the cooperation. In addition, we propose feature classification redundancy (FCR) to indicate the part of feature redundancy (FR) which is really relative to the target classification task. By computing FCR information rather than FR information, it can decrease the probability of mistaking important features as redundant features in searching process. Based on our criterion and FCR, a fast sequential forward feature selection algorithm named conditional mutual information-based feature selection (CMIFS) is devised by using greedy optimization. With neither intensive matrix operation nor high-dimension MI evaluation, CMIFS provides a low-complexity solution for resource-constraint applications. The requirement of the memory storage of CMIFS is low because it need not calculate the accumulated MI between the candidate features and the selected ones. Experimental results show that CMIFS outperforms MIFS, mRMR, and NMIFS on the public benchmark data sets.

The reminder of this paper is organized as follows. Section II reviews the concepts of information theory and gives the formal definition of feature interaction. Section III presents our proposed CMIFS algorithm. Section IV reports experimental results on several public benchmark data sets. Finally, the conclusions are drawn in section V.

## II. Feature Selection Based on Conditional Mutual Information

### 1. Evaluation of Mutual Information

In classification tasks, the relevant features contain important discriminative information of the classes. Shannon's information theory provides us a way to quantify the feature information with entropy and MI [12], [13]. Evaluating the entropy and MI between two discrete feature variables is feasible and convenient through histograms. For continuous feature variables, some type of discretization methods such as the minimum description length (MDL) discretization method [14] will be applied before computing entropy or MI.

Let $F$, $S$, and $C$ denote the original feature set, the selected

feature subset, and output classes, respectively. We review the following three formulas which are used in our work.

**Formula 1.** Given an output set of classes $C = \{c_1, c_2, \ldots, c_l\}$, the initial uncertainty in $C$ is

$$H(C) = -\sum_{i=1}^{l} P(c_i)\log P(c_i), \quad (1)$$

where $P(c_i)$, $i = 1, \ldots, l$, is the probability for the specific classes [13].

**Formula 2.** Given a feature vector $F$ and an output class $C$, the average uncertainty in $C$ is

$$H(C \mid F) = -\sum_{f\in F} P(f)\sum_{c\in C} P(c \mid f)\log P(c \mid f), \quad (2)$$

where $P(f)$ is the probability for individual features in $F$, and $P(c|f)$ denotes the conditional probability for class $c$ given input feature $f$ [13].

**Formula 3.** Given a feature vector $F$ and an output class $C$, the amount of decreased class uncertainty is

$$\begin{aligned} I(C;F) &= I(F;C) = H(C) - H(C|F) \\ &= \sum_{c,f} P(c,f)\log\frac{P(c,f)}{P(c)P(f)}, \end{aligned} \quad (3)$$

where $P(c; f)$ is the joint probability of class $c$ and feature $f$ [13].

### 2. Optimal Feature Subset

**Definition 1.** The optimal feature subset is the feature subset $S$ iff $I(C; S)$ is maximized or $H(C|S)$ is minimized [5].

According to definition 1, in a greedy strategy of sequential forward searching, where one feature is selected at a time, the next feature $f_i \in F - S$ to be selected is the one that makes $I(C; \{S, f_i\})$ its maximum. Since $I(C; \{S, f_i\})$ satisfies the chain rule for information [13], it can be represented as

$$I(C;\{S,f_i\}) = I(C;S) + I(C;f_i \mid S). \quad (4)$$

For a given feature subset $S$, $I(C; S)$ is a constant. To maximize $I(C; \{S, f_i\})$, the conditional mutual information $I(C; f_i | S)$ should be maximized.

### 3. Interaction Information

Interaction information [15] among features can be understood as the amount of information (redundancy or synergy) bound up in a set of features, but not present in any subset. $I(C; f_i | S)$ in (4) can be represented as

$$I(C;f_i \mid S) = I(C;f_i) + I(C;f_i;S), \quad (5)$$

where $I(C; f_i; S)$ is called the interaction information among $C$, $f_i$, and $S$. For two features, $f_i$ and $f_j$, the interaction information $I(C; f_i; f_j)$ is a 3-way interaction. Let $\Theta$ denote a metric that measures the relevance of the class label with a feature or a feature subset.

**Definition 2.** Given a feature subset $S$ with $k$ features, the interaction of the $k$ features is said to form $k$-way feature interaction iff for any arbitrary partition of $S$, denoted as $S_1, S_2, S_3, \ldots, S_l$, where $l \geq 2$ and $S_i \neq \Phi$, we have $\forall i \in [1, l]$, $\Theta(S) > \Theta(S_i)$.

Unlike the mutual information, the interaction information can be negative, positive, or zero [16]. For simplicity, we use 3-way interaction to illustrate the three types of interaction. When $I(C; f_i; f_j) < 0$, the negative interaction suggests a redundancy between $f_i$ and $f_j$, meaning that $f_i$ and $f_j$ provide completely or partially the common information about $C$. When $I(C; f_i; f_j) > 0$, the positive interaction indicates a synergy between $f_i$ and $f_j$ instead. It means that $f_i$ and $f_j$ yield more information together than what could be expected from the two individual interactions with the class labels. When $I(C; f_i; f_j) = 0$, it implies that $f_j$ does not affect the relationship between $f_i$ and $C$. It is extremely important to note that the three types of interactions coexist in most real-life domains. However, directly identifying a $k$-way feature interaction requires exponential time. In order to avoid exponential time complexity, the interaction is approximately evaluated in MIFS and its variants described in section I. The estimation formula for $I(C; f_i|S)$ in MIFS and the variants can be uniformly expressed as

$$\hat{I}(C;f_i \mid S) = I(C;f_i) - \xi\sum_{f_j\in S} I(f_i;f_j), \quad (6)$$

where $f_i \in F - S$, $\xi$ is the feature redundancy penalization parameter and $\xi \geq 0$. By comparing (5) and (6), we can see that $I(C; f_i; S)$ is estimated by $-\xi\sum_{f_i\in S} I(f_i;f_j)$ in MIFS and its variants. Since mutual information is nonnegative and $\xi \geq 0$, $-\xi\sum_{f_j\in S} I(f_i;f_j) \leq 0$. It is obvious that (6) cannot deal with the case of positive interaction, which means (6) may cause a big bias when the relevant features are combined to cooperate together.

## III. CMIFS

In this section, we propose a conditional mutual information based feature selection criterion which considers synergy and redundancy interactions of features. In addition, we propose to use feature classification redundancy information rather than feature redundancy information to evaluate feature redundant information for classification. Based on our criterion and

feature classification redundancy, we devise the CMIFS algorithm.

## 1. Criterion of CMIFS

In fact, there is a link between interaction information and conditional mutual information: $I(C; f_i; S) = I(S; f_i | C) - I(S; f_i)$. Thus, $I(C; f_i | S)$ in (5) can also be represented as

$$I(C; f_i \mid S) = I(C; f_i) + [I(S; f_i \mid C) - I(S; f_i)]. \quad (7)$$

Let $f_k$ ($k$=1,...,$|S|$) denote the $k$-th feature previously selected into $S$ and $f_i \in F - S$ denote the candidate feature.

When $|S| = 1$, $I(C; f_i | f_1)$ can be calculated effectively. In the case of $|S| \geq 2$, it may be hard to calculate $I(C; f_i \mid S)$ exactly because of the limited number of samples and high computation effort. A substitute is to measure $I(C; f_i \mid S)$ approximately with low-complexity criterions.

Suppose we already have the subset $S$ with $n$ ($n \geq 2$) features. We propose that the next feature is selected by optimizing the following criterion:

$$f_{n+1} = \underset{f_i \in F-S}{\operatorname{argmax}} \{I(C; f_i \mid f_1) - [I(f_n; f_i \mid f_1) - I(f_n; f_i \mid C)]\}. \quad (8)$$

We use the inductive method to describe the derivation of our proposed criterion.

For $n = 2$, we reduce the dimension of measuring $I(C; f_i \mid \{f_1, f_2\})$ by evaluating $I(C; f_i | f_2)$ in the context of $f_1$. According to (7), $I(C; f_i | f_2) = I(C; f_i) + [I(f_2; f_i | C) - I(f_2; f_i)]$. Given $f_1$, we have

$$\hat{I}(C; f_i \mid \{f_1, f_2\}) = I(C; f_i | f_1) + [I(f_2; f_i | C) - I(f_2; f_i | f_1)]. \quad (9)$$

By using (9), we can approximately compute $I(C; f_i \mid \{f_1, f_2, f_3\})$ as

$$\begin{aligned}
&I(C; f_i \mid \{f_1, f_2, f_3\}) \\
&\approx \hat{I}(C; f_i \mid \{f_1, f_2\}) + [I(f_3; f_i | C) - \hat{I}(f_3; f_i \mid \{f_1, f_2\})] \\
&= I(C; f_i | f_1) + [I(f_2; f_i | C) - I(f_2; f_i | f_1)] \\
&\quad + I(f_3; f_i | C) - I(f_3; f_i | f_1) - [I(f_2; f_i | f_3) - I(f_2; f_i | f_1)] \\
&= I(C; f_i | f_1) + I(f_3; f_i | C) - I(f_3; f_i | f_1) \\
&\quad + [I(f_2; f_i | C) - I(f_2; f_i | f_1)] - [I(f_2; f_i | f_3) - I(f_2; f_i | f_1)] \\
&= I(C; f_i | f_1) + I(f_3; f_i | C) - I(f_3; f_i | f_1) + [I(f_2; f_i | C) - I(f_2; f_i | f_3)] \\
&\approx I(C; f_i | f_1) + I(f_3; f_i | C) - I(f_3; f_i | f_1).
\end{aligned} \quad (10)$$

We assume that $I(C; f_i | S_{k-1})$ ($k \geq 3$) can be approximately calculated by

$$\hat{I}(C; f_i | S_{k-1}) = I(C; f_i | f_1) + [I(f_{k-1}; f_i | C) - I(f_{k-1}; f_i | f_1)]. \quad (11)$$

For $n = k$, we can derive the approximate $I(C; f_i \mid \{S_{k-1}, f_k\})$ from (9) and (11) as

$$\begin{aligned}
&I(C; f_i \mid \{S_{k-1}, f_k\}) \\
&\approx \hat{I}(C; f_i | S_{k-1}) + [I(f_k; f_i | C) - \hat{I}(f_k; f_i | S_{k-1})] \\
&= I(C; f_i | f_1) + [I(f_{k-1}; f_i | C) - I(f_{k-1}; f_i | f_1)] + I(f_k; f_i | C) \\
&\quad - I(f_k; f_i | f_1) - [I(f_{k-1}; f_i | f_n) - I(f_{k-1}; f_i | f_1)] \\
&= I(C; f_i | f_1) + I(f_k; f_i | C) - I(f_k; f_i | f_1) \\
&\quad + [I(f_{k-1}; f_i | C) - I(f_{k-1}; f_i | f_k)] \\
&\approx I(C; f_i | f_1) + I(f_k; f_i | C) - I(f_k; f_i | f_1).
\end{aligned} \quad (12)$$

Thus, we get the approximate estimation formula for $I(C; f_i \mid S_n)$ ($n \geq 2$) as

$$\hat{I}(C; f_i \mid S_n) = I(C; f_i \mid f_1) + [I(f_n; f_i \mid C) - I(f_n; f_i \mid f_1)]. \quad (13)$$

According to (13), a $k$-way interaction ($k \geq 4$) can be eventually simplified to 3-dimensional conditional information computation. Unlike (6), (13) can measure positive interactions as well as negative and zero interactions. Although only two features, $f_1$ and $f_n$, rather than all the previously selected features, are used to decide the selection of $f_i$, the recurrence relation potentially involves all the selected feature subset in the searching process.

The computational complexity of (8) is $O(|F - S|)$. It is lower than that of the criterions used in MIFS and its variants with $O(|S| \cdot |F - S|)$. The linear computation complexity ensures that criterion (8) can be applied to many applications with high dimension data, such as image retrieval, object recognition, analysis of genomic microarrays, and text categorization.

## 2. Feature Classification Redundancy

In classification tasks, redundant features as a kind of noise data should be removed. MIFS and its variants recognize a redundant candidate feature $f_i$ based on its dependency with the selected features, that is, $\sum_{f_j \in S} I(f_i; f_j)$. The higher the sum is, the more likely $f_i$ will be identified as redundant regarding class discriminative power and not preferred. In fact, FR information, regardless of the classification task, such as $I(f_i; f_j)$, only says how overlapping these features are. It is possible that FR information does not subsume anything about $C$.

We define FCR information as the redundancy information which is really relative to the target concept. To distinguish FR and FCR, the comparison is illustrated in Fig. 1. We can see that in Figs. 1(a) or 1(c), FCR information ($I_{FCR}$) is part of FR information ($I_{FR}$) measured by $I(f_i; f_j)$. While in case Fig. 1(b), there is no $I_{FCR}$ between $f_i$ and $f_j$, even though $I_{FR}$ exists. Intuitively, if $I_{FCR}$ between $f_i$ and $f_j$ subsumes the total information of $f_i$ on the class discrimination, $f_i$ can be removed without compromising the learning of a classification rule. In

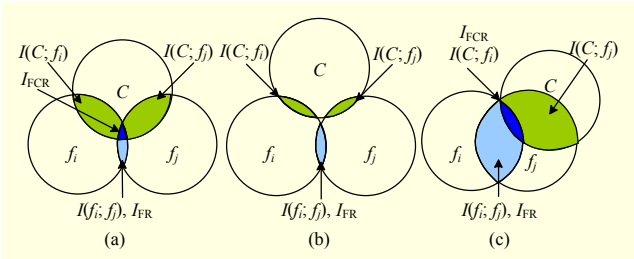


Fig. 1. FCR information vs. FR information.

this case, $f_i$ is called a classification redundancy feature of $f_j$. For example, in Fig. 1(c), given $f_j$, $f_i$ is a classification redundancy feature of $f_j$ and can be removed although $f_i$ is not a redundant feature of $f_j$. While for case Figs. 1(a) or 1(b), even when $I(f_i; f_j)$ is high, $f_i$ should not be treated as a classification redundant feature of $f_j$.

The comparison concludes that FCR and FR cannot be confused. If they are, important features may be mistaken as redundant features in the searching process and not be highly considered by the feature selectors.

We use FCR information rather than FR information to evaluate feature redundant information for classification. When $f_i$ is a classification redundancy feature of $f_j$, we have $I(C; f_i \mid f_j) = 0$. Extending $f_j$ to the selected subset $S$, we give the following definition of classification redundant feature.

**Definition 3.** A feature $f_i \in F - S$ is a classification redundant feature of $S$ if $I(C; f_i) > 0$ and $I(C; f_i \mid S) = 0$.

Obviously, according to the searching rule of maximizing $I(C; f_i|S)$, the existence of classification redundant features have little impact on the criterions' results of selected suboptimal subsets. However, in order to reduce unnecessary time costs, classification redundant features of $S$ should be removed before a new turn of looking through the candidate subset space.

## 3. CMIFS Algorithm

On the basis of the criterion proposed in (8), we devise the feature selection algorithm CMIFS. Although CMIFS is a greedy algorithm, it will not waste time on unnecessary features by removing classification redundancy features beforehand. The procedure description of CMIFS is illustrated in Fig. 2.

In CMIFS, the feature subset $S$ is built up step by step, by adding one feature at a step. For simplicity, we evaluate classification redundant features of individual features in the selected subset $S$. Since approximate classification redundant features are reasonable for general applications, we define the ratio of $I(C; f_i|s_n)$ to $I(C; f_i)$ as a candidate feature $f_i$'s classification information gain degree for the selected feature $s_n$. We set a threshold $\delta$ to quantify the likeness of $f_i$ being a classification redundancy feature of $s_n$. While the ratio of $I(C; f_i|s_n)$ to $I(C; f_i)$ is not more than $\delta$, $f_i$ is treated as the classification redundant feature of $s_n$ and removed from the candidate subset. This means that the contribution of $f_i$ decreasing the classification uncertainty is minor and can be ignored.

Algorithm CMIFS

Input: $F = \{f_1, f_2, \ldots, f_n\}, C, \delta$
Output: the features ranked in the selection order

$S \leftarrow [\,]$
$f^* \leftarrow \text{argmax}_{f_i \in F} I(C; f_i)$
$F \leftarrow F - f^*, S \leftarrow \{f^*\}, s_1 \leftarrow f^*, s_n \leftarrow f^*$
$f^* \leftarrow \text{argmax}_{f_i \in F} I(C; f_i \mid s_1)$
$F \leftarrow F - f^*, S \cup \{f^*\}, s_n \leftarrow f^*$
while $F \neq [\,]$ do
  for $i = 1$ to $|F|$ do
    if $I(C; f_i) > 0 \vee \frac{I(C; f_i|s_n)}{I(C; f_i)} \leq \delta$) then
      $F \leftarrow F - f_i$
    end if
  end for
  $f^* \leftarrow \text{argmax}_{f_i \in F} \{I(C; f_i|s_1) - I(f_i; s_n|s_1) + I(f_i; s_n|C)\}$
  $F \leftarrow F - f^*, S \cup \{f^*\}, s_n \leftarrow f^*$
end while

Fig. 2. Procedure description of CMIFS.

According to (5), we have $0 < \frac{I(C; f_i|s_n)}{I(C; f_i)} < 1$ when there is FCR information between $f_i$ and $s_n$. In order to detect approximate classification redundant features of $s_n$, we should set $\delta$ a value between 0 and 1. As $\delta$ grows, it excludes the classification redundant features more efficiently. However, the value of $\delta$ should be under a reasonable value range so as to prevent informative features from being removed unreasonably. Based on our observation, a value for $\delta$ between 0.05 and 0.2 is appropriate for many classification tasks to make good tradeoff between efficiency and reasonability.

If we set $\delta$ as zero, only the precise classification redundant features of $s_n$ will be removed from the candidate subset. If we set $\delta$ a value less than zero, the search space has the original size. The benefit of using threshold $\delta$ is that it provides the possibility to further speed up CMIFS. Which kind of $\delta$ can be determined based on the requirements of the classification task at hand.

# IV. Experiments

In this section, we describe how we conduct the experiments on the public data sets and present the experimental results.

## 1. Experiment Setup

CMIFS is compared with MIFS, mRMR, and NMIFS on

Table 1. Data sets used in experiments.

| Data set | #Feature | #Instance | #Class | Type |
|---|---|---|---|---|
| Zoo | 16 | 101 | 7 | Discrete |
| Testcolon | 2,000 | 62 | 2 | Discrete |
| Glass | 9 | 214 | 7 | Continuous |
| Sonar | 60 | 208 | 2 | Continuous |

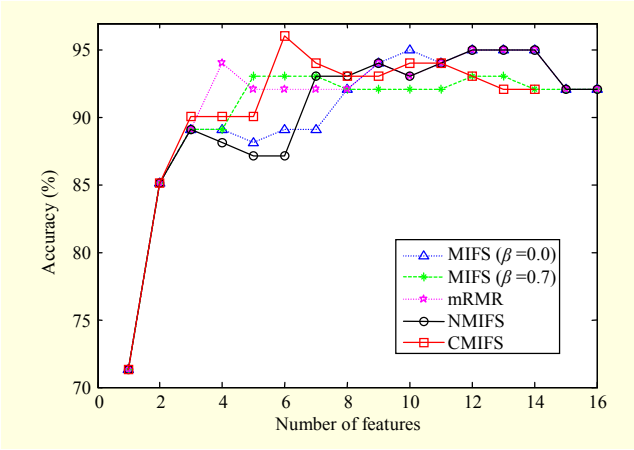


Fig. 3. Average classification accuracies on zoo data set.

four public benchmark data sets (see Table 1). Zoo, glass, and sonar data sets are available on the UC-Irvine machine learning database [17]. Testcolon is a microarray gene expression data set, available on the mRMR Web site [18]. The four data sets are very different in the number of instances, features, and label classes. In addition, the first two data sets are discrete, and the latter two are continuous. For continuous data, MDL discretization method is used before feature selection is taken.

In the experiments, we set the threshold $\delta$ used in CMIFS to 0.2. Because Battiti [5] finds that a value for $\beta$ (the feature redundancy penalization parameter) between 0.5 and 1 is appropriate for many classification tasks, we set $\beta$ used in MIFS to 0.7. In addition, we choose MIFS ($\beta = 0$) as the max-relevance criterion which only considers the relevance of individual features and the class labels.

In order to evaluate how good the selected features are, we apply a C4.5 classifier to evaluate the classification quality of the features selected by each of the four algorithms. All the classification experiments are conducted in the WEKA environment [19]. For every data set, the classification is performed by using 10-fold cross-validation.

## 2. Evaluation Metric

The metric of average classification accuracy $\overline{A}$ is used to evaluate C4.5 classifier and evaluated by the formula:

$$\overline{A} = \frac{1}{k}\sum_{i=1}^{k} A_i = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{t_i}{n}\times 100\%\right), \qquad (14)$$

where $t_i$ is the number of instances correctly classified in the $i$-th run, $n$ is the total number of instances, and $k$ is the number of repetitions. For a 10-fold cross-validation, $k = 10$.

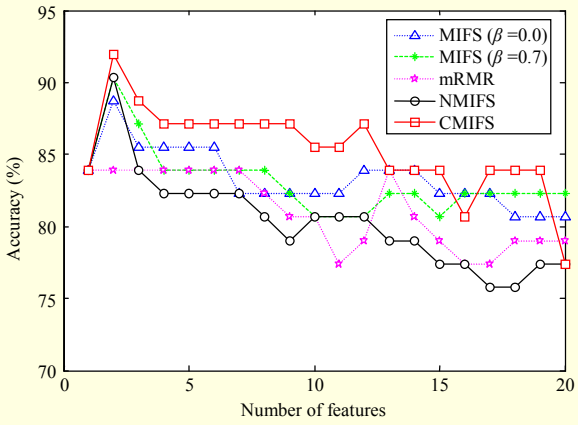


Fig. 4. Average classification accuracies on testcolon set.

## 3. Results on Benchmark Data

Figure 3 shows the average accuracies on the zoo data set. The highest average accuracy achieved by CMIFS is 96.04%, which is higher than 95.05% obtained by MIFS ($\beta$ = 0), mRMR, NMIFS, and the 93.07% gotten by MIFS ($\beta = 0.7$). In addition, the best classification result of CMIFS is achieved when only 6 features are selected rather than the 10 or more features required in MIFS ($\beta = 0$), MIFS ($\beta = 0.7$), mRMR, and NMIFS. On this data set, with the help of the threshold $\delta$, 2 classification redundancy features are removed beforehand, which makes CMIFS focus on informative features during searching process.

For the testcolon data set, 20 features are selected by using each feature selection algorithm. As we can see from the results shown in Fig. 4, the features selected by CMIFS have clearly higher discriminative ability than the features selected by MIFS ($\beta = 0$), MIFS ($\beta = 0.7$), mRMR, and NMIFS. Although all of them achieve their best average accuracy when 2 features are selected, CMIFS achieves the highest accuracy at 91.94%, while the best average accuracy achieved by MIFS ($\beta$ = 0.7) and NMIFS is 90.32%. MIFS ($\beta$ = 0) and mRMR obtains lower classification accuracy with 88.71% and 83.87%, respectively.

Figure 5 shows the average accuracies on the glass data set. We observe that the overall feature selection performance of CMIFS is better than that of other algorithms. The highest

Table 2. Comparison of algorithms on highest classification accuracy (%) and respective number of selected features.

| Data set | MIFS ($\beta = 0$) | | MIFS ($\beta = 0.7$) | | mRMR | | NMIFS | | CMIFS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | $N°$ | Accuracy | $N°$ | Accuracy | $N°$ | Accuracy | $N°$ | Accuracy | $N°$ |
| Zoo | 95.05 | 10 | 93.07 | 12 | 95.05 | 12 | 95.05 | 12 | 96.04 | 6 |
| Testcolon | 88.71 | 2 | 90.32 | 2 | 83.87 | 2 | 90.32 | 2 | 91.94 | 2 |
| Glass | 70.09 | 8 | 68.22 | 8 | 70.09 | 8 | 70.09 | 8 | 70.56 | 4 |
| Sonar | 80.77 | 17 | 74.04 | 41 | 79.81 | 9 | 81.73 | 16 | 82.69 | 4 |

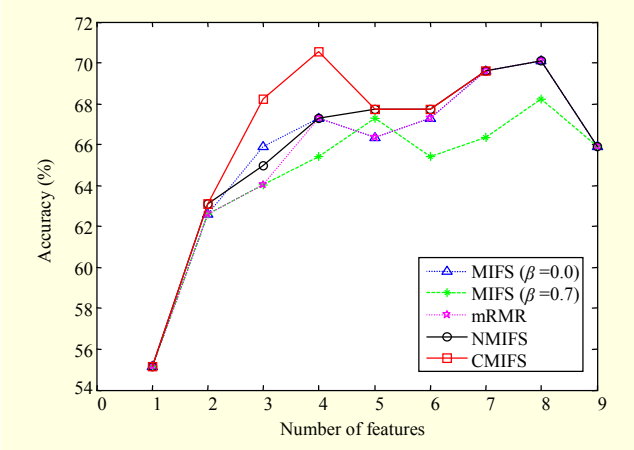


Fig. 5. Average classification accuracies on glass data set.

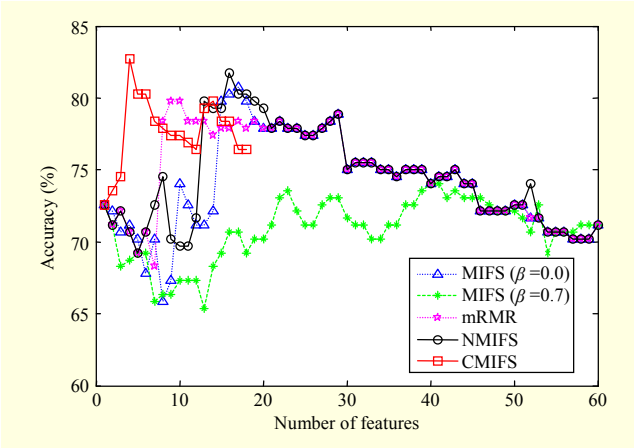


Fig. 6. Average classification accuracies on sonar data set.

accuracy achieved by CMIFS is 70.56% when 4 features are selected, better than the 70.09% obtained by MIFS ($\beta = 0$), mRMR, and NMIFS, and the 68.22% gotten by MIFS ($\beta = 0.7$) when 8 features are selected. On this data set, there are 2 classification redundancy features removed by CMIFS.

On the sonar data set, the highest accuracy obtained by CMIFS (shown in Fig. 6) is 82.69% when 4 features are selected. It is compared with the highest accuracy of 81.73%, 80.77%, 79.81%, and 74.04% obtained by NMIFS, MIFS ($\beta = 0$), mRMR, and MIFS ($\beta = 0.7$) when 16, 17, 9, and 41 features are selected, respectively. With the benefit of removing 42 classification redundant features, CMIFS finishes the searching process after selecting 18 features rather than the full feature set with 60 features.

Table 2 summarizes the best average classification accuracies obtained by the four algorithms and the respective number of selected features when the best accuracies are achieved. Obviously, CMIFS can more quickly find out the suboptimal feature subset than other three algorithms. In addition, the features selected by CMIFS have stronger classification discriminative power than the features selected by other algorithms. The outperformance of CMIFS can be attributed to the reason that CMIFS takes account of both positive and negative interactions. The feature dependency in CMIFS ensures the possibility to detect the relevant feature combinations in some degree.

Moreover, the removal of classification redundancy features makes CMIFS focus on informative features during searching process. Last but not least, the requirements of computation cost and memory storage of CMIFS are low. The calculation of CMIFS only needs $C, f_1, f_n$, and $F - S$ for evaluating the next feature to be selected. However, MIFS needs $C$, $S$, and $F - S$ to do the evaluation task. With the increase in the number of original features, the computation effort and memory storage are increased for considering the MI between the selected feature subset and the candidate feature subset.

## V. Conclusion

This paper presents a new greedy feature selection algorithm based on conditional mutual information named CMIFS. Unlike MIFS and its variants which ignore synergy of features, CMIFS can detect both cooperation and redundancy interactions of features. CMIFS computes FCR information rather than FR information. It can decrease the probability of mistaking important features as redundant features in searching process. With the benefit of removing unnecessary features, CMIFS focuses on informative features in searching process.

Besides, CMIFS imposes no limitation on feature information distribution. Experimental results, except those of the testcolon data set, show that CMIFS uses almost half the number of features required by other comparison methods (MIFS, mRMR, and NMIFS) to get higher best-classification-accuracy. For the testcolon data set, both CMIFS and the comparison methods use two features to obtain their own best-classification-accuracy, but CMIFS achieves 91.94% best-classification-accuracy, which is higher than the 88.71%, 90.32%, and 83.87% obtained by MIFS, NMIFS, and mRMR, respectively. Since the computation complexity of the criterion used in CMIFS is $O(|S - F|)$, CMIFS can be applied to the applications with high-dimensional data, such as image processing, biology, and multimedia data retrieval.

## References


[1] D. Koller and M. Sahami, "Toward Optimal Feature Selection," *Proc. 13th Int. Conf. Machine Learning*, 1996, pp. 284-292.

[2] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, 1997, pp. 131-156.

[3] E. Amaldi and V. Kann, "On the Approximation of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems," *Theoretical Computer Sci*., vol. 209, 1998, pp. 237-260.

[4] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intell.*, vol. 97, no. 1-2, 1997, pp. 273-324.

[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, 1994, pp. 537-550.

[6] N. Kwak and C.H. Choi, "Input Feature Selection for Classification Problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, 2002, pp. 143-159.

[7] J.J. HUANG et al., "Feature Selection for Classificatory Analysis Based on Information-Theoretic Criteria," *Acta Automatica Sinica*, vol. 34, no. 3, 2008, pp. 383-392.

[8] 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. Machine Intell*., vol. 27, no. 8, 2005, pp. 1226-1238.

[9] P.A. Estevez et al., "Normalized Mutual Information Feature Selection," *IEEE Trans. Neural Netw*., vol. 20, no. 2, 2009, pp. 189-201.

[10] J. Novovicova, "Conditional Mutual Information Based Feature Selection for Classification Task," *Progress Pattern Recog., Image Anal. Appl., LNCS*, Springer, vol. 4756, 2007, pp. 417-426.

[11] R.M. Fano, *Transmission of Information: A Statistical Theory of Communications,* New York, USA: Wiley Press, 1961.

[12] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana, Israel: University of Illinois Press, 1949.

[13] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York, USA:Wiley-Interscience Press, 1991.

[14] U.M. Fayyad and K.B. Irani, "Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. 13th Int. Joint Conf. Artificial Intell.*, 1993, pp. 1022-1027.

[15] W.J. McGill, "Multivariate Information Transmission," *Psychomeetrika*, vol. 19, no. 2, 1954, pp. 97-116.

[16] A. Jakulin and I. Bratko, "Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy." Available: http://arxiv.org/abs/cs.AI/0308002v3, 2004.

[17] C.J. Merz and P.M. Murphy, "UCI Repository of Machine Learning Databases [Online]." Available: http://www.ics.uci.edu/fmlearn/MLRepository.html.

[18] H. Peng, "mRMR Sample Data Sets [Online]." Available: http://penglab.janelia.org/proj/mRMR/test colon s3.csv.

[19] I.H. Witten and E. Frank, *Data Mining-Practical Machine Learning Tools and Techniques with JAVA Implementations*, Morgan Kaufmann Publishers, 2nd ed., 2005.



**Hongrong Cheng** received her BS and MS in computer science from the University of Electronic Science and Technology of China (UESTC), P.R. China, in 1998 and 2001, respectively. Since 2001, she has been a faculty member of the School of Computer Science and Engineering of UESTC. She is now pursuing her PhD at UESTC. Her research interests are information retrieval, pattern recognition, and machine learning.



**Zhiguang Qin** received his MS from Xiangtan University, P.R. China, in 1989. He received his PhD from UESTC, P.R. China, in 1996. He is now a professor and the dean of the School of Computer Science and Engineering, UESTC. His research interests include information retrieval, peer-to-peer systems, and mobile networks.



**Chaosheng Feng** received his MS in computer software and theory from Sichuan Normal University, P.R. China, in 2006. He received his PhD from UESTC, P.R. China, in 2010. He is now an associate professor at Sichuan Normal University. His research interests include peer-to-peer systems, mobile networks, and the simulation of computer networks.

**Yong Wang** received his MS in computer science from Sichuan University, P.R. China, in 2001. He received his PhD from the Institute of Computing Technology Chinese Academy of Sciences, in 2008. He is now an associate professor in the School of Computer Science and Engineering, UESTC. His research interests include information retrieval, peer-to-peer systems, and mobile networks.

**Fagen Li** received his MS from Hebei University of Technology, P.R. China, in 2004, and his PhD from Xidian University, P.R. China, in 2007. He was a postdoctoral fellow in Future University-Hakodate, Hokkaido, Japan, from 2008 to 2009. He is now an associate professor in the School of Computer Science and Engineering, UESTC. His recent research interests include cryptography and the design of protocols.