Feature Selection Based on Dependency Margin

Yong Liu, *Member, IEEE,* Feng Tang, *Member, IEEE*, and Zhiyong Zeng

***Abstract*—Feature selection tries to find a subset of feature from a larger feature pool and the selected subset can provide the same or even better performance compared with using the whole set. Feature selection is usually a critical preprocessing step for many machine-learning applications such as clustering and classification. In this paper, we focus on feature selection for supervised classification which targets at finding features that can best predict class labels. Traditional greedy search algorithms incrementally find features based on the relevance of candidate features and the class label. However, this may lead to suboptimal results when there are redundant features that may interfere with the selection. To solve this problem, we propose a subset selection algorithm that considers both the selected and remaining features' relevances with the label. The intuition is that features, which do not have better alternatives from the feature set, should be selected first. We formulate the selection problem as maximizing the dependency margin which is measured by the difference between the selected feature set performance and the remaining feature set performance. Extensive experiments on various data sets show the superiority of the proposed approach against traditional algorithms.**

***Index Terms*—Conditionally independent, dependency margin, feature selection, forward greedy search, redundant feature.**

I. Introduction

FEATURE selection tries to reduce the number of features while keeping the same or even better learning performance. It's a frequently used preprocessing step for many machine learning applications [1]. As it removes irrelevant, redundant and noisy features, the learning speed is usually significantly increased and in some cases the learning performance can even be improved because irrelevant and noisy features are excluded from the learning process. Feature selection is a broad research area and it can be divided into two categories, one is for supervised learning [2]–[8] where the target is to learn the prediction function between the features and the label, the other is for unsupervised learning [9]–[13] which tries to find the underlying structure of the data in some feature space. In this paper, we mainly focus on feature selection for supervised classification.

Manuscript received December 9, 2013; revised April 24, 2014 and August 4, 2014; accepted August 5, 2014. Date of publication September 26, 2014; date of current version May 13, 2015. This work was supported in part by the National Natural Science Foundation Project of China under Project 61173123, in part by the Natural Science Foundation Project of Zhejiang Province under Project LR13F030003, and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China, under Project ICT1315. This paper was recommended by Associate Editor J. Basak.

The authors are with the State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China (e-mail: yongliu@iipc.zju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2347372

Feature selection for supervised classification can be formulated as: given the original feature set $F = \{f_1, f_2, \ldots, f_i\}$ and label Y, find the proper feature subset $S \subset F$ which can "best" predict Y. That problem can be decomposed into four sub-problems [14].

1) Feature evaluation which evaluates the relevance between a feature and the class label.
2) Search strategies which quickly find the optimal subset from a large pool.
3) Stopping criterion which determines when to stop the search.
4) Validation strategies which validates the selected feature set.

In particular, the feature evaluation and search strategies play critical roles in a feature selection algorithm.

Feature evaluation functions measure the relevance between a feature and the label, the most straightforward and perhaps the most effective evaluation criteria is directly evaluating [15] the features by the classification performance. However, this process is usually very slow and not scalable to real problem when the feature set is very large or when the classifier is complicated. To solve this problem, researchers propose to use simpler indirect feature measures such as information entropy [16], [17], consistency [5], [18], dependency [19], divergence [20], and fuzzy-rough measurement [21], etc.

For search strategies, exhaustive search [22] may provide the optimal solution but it only works when the number of features is small. As the number of features increases, the computation increases exponentially, which makes it impractical for many learning algorithm with high dimensional feature. Some works [16], [23] propose to rank individual feature based on the relevance between each feature and the class label which drastically reduces the computation, making it linear to the number of features. However, this approach will not work well when there are redundant features in the feature set. Guyon and Elisseeff [24] also used sample examples to show that "perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them." To solve this problem, some researchers propose to use greedy strategies which iteratively maximize the gain of the feature importance evaluation functions and then generate the nested subset of features. It has been shown that the forward greedy search strategies are particularly computationally advantageous and robust against overfitting. However, the forward greedy strategies only considers the relevance between the feature and the label while ignoring the redundancy among the features. It suffers from the problem of generating nonoptimal subsets because the importance of features is not assessed in the context of remaining features [24]. In Fig. 1, we use Guyon and Elisseeff's [24] example to demonstrate the

2168-2267 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

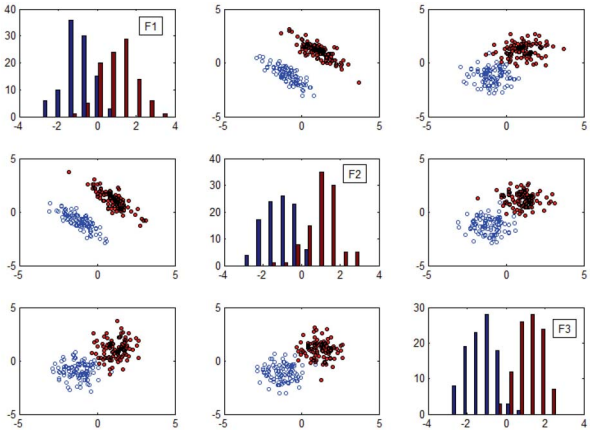


Fig. 1. Sample data set with three features and two labels from [24], we use circles and dots to represent samples with different labels. The sub figures in the diagonal show the histograms of the projections of the different samples on the axes, and the other sub figures show the scatter plots in the 2-D space of the corresponding features, to facilitate its reading, the scatter plot is shown twice with an axis exchange.

problem that the forward greedy search strategies may find a suboptimal subset. There are three features and the third one (third row in Fig. 1) separates the two classes best by itself (bottom right histogram in Fig. 1). It is therefore the best candidate in a forward greedy selection process. However, the two other features are better taken together than any subset of size two including it. In this example, the third feature is actually redundant to the other two features. If the evaluation function only considers the feature relevance, the third feature will always be chosen with a higher priority than the other two features in the beginning of forward greedy searching. This leads to the suboptimal solution and also suggests that we should pursue a better solution framework that can avoid the limitations of forward greedy search strategies.

In this paper, we present a new feature evaluation function, named dependency margin of subsets, which can overcome the problem of forward greedy searching strategies by considering both the feature relevance and redundancy simultaneously. We also present a feature selection algorithm based on Yu and Liu's [3] redundant feature selection theory.

The rest of this paper is organized as follows. In Section II, we will review the related works on feature selection and motivate the basic idea to solve the problems of current forward greedy search methods. In Section III, we briefly introduce some background knowledge on Yu and Liu's Bayesian conditionally independent theory on feature redundancy [3], which is used in our formulation. In Section IV, we present our dependency margin based feature selection method. Empirical comparisons of our method comparing with state-of-art feature selection methods are presented in Section V. Finally, we conclude this paper and discuss future directions in Section VI.

II. Related Works and Basic Idea

A. Related Works

Feature selection is one of the fundamental problems in machine learning. Earlier research mostly focuses on selecting "relevant" features that are highly relevant to the class label [5], [18], [25]. However, in many real world problems, such as genomic microarray analysis [26] and text categorization [27], the extracted features usually have high redundancy. This makes the "relevance" based algorithms selecting suboptimal features [4], [15], [25], [28]. Researchers proposed many algorithms for feature selection considering redundancy in the feature set. There are generally two types of redundancy considered in the work.

1) *Type I:* A feature is redundant with respect to another feature according to the capability to predict the class label.
2) *Type II:* A feature is redundant with respect to another subset of features according to the capability to predict the class label.

Based on different redundancy definitions, redundancy based feature selection algorithms can be roughly divided into two categories. The first category uses feature relevance to identify redundant features [17], [29]–[32]. These approaches tend to select features which provide maximal relevancy with respect to the label and at the same time minimize the relevance among each pair of the selected features. For example, Peng *et al*. [17] and Ding and Peng's [32] feature selection algorithm uses mutual information to measure the relevance among features. Biesiada and Duch's approach [31] uses the pairwise Pearson χ^2 test which uses the difference between the feature probability distribution of two random variables to measure the relevance. In Appice *et al*'s approach [29], their improved REDUCE algorithm operates by pairwise comparison of the feature redundancy. These approaches can handle the first type of redundant features well, however, they fail when the type II redundant features exist because they only consider pairwise redundancy. Furthermore, the relevance between the pairwise features may not be a proper evaluation measure for the feature redundancy. It has been proved by Guyon and Elisseeff's [24] example that "very high variable correlation (or anti-correlation) does not mean absence of variable complementarity."[1] Guyon and Elisseeff's example also demonstrates that correlated among features is not the same as redundancy among features.

The second category of redundant feature selection algorithms is based on the framework of Bayesian conditional independency [33]. It starts from the research works of John *et al.* [34], which classifies features into three disjoint categories, namely strong relevant, weakly relevant and irrelevant features [34]. Although they have realized that some features, which are totally conditional independent with the labels, can be viewed as redundant and eliminated, their theory still cannot explain why some weakly relevant features are not required by the optimal feature subset. As an extension, Yu and Liu [3] introduce the Markov blanket into John *et al.*'s [34] theoretical framework and give a more precise definition of the feature redundancy, which points out weakly relevant features with Markov blanket as also redundant. Although the second viewpoint and its theory can explain and define the feature redundancy very well, it is not really

[1] In Guyon and Elisseeff's example [24, Fig. 2(b)] they employ two highly correlated features which can achieve almost perfect separation results.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

practical as it is very difficult to find the Markov blanket efficiently. Some other works [3], [35], [36] tend to employ symmetrical uncertainty based approximate Markov blanket to remove the redundant features to improve the efficiency. Furthermore, all of those solutions consider individual feature relevancy and redundance, they cannot remove the redundant features of type II, e.g., the third feature in Fig. 1, which are identified by a subset of features.

B. Basic Idea

We use $D(X, Y)$ (abbreviated as $D(X)$) to denote the relevance based feature evaluation function which represents the relevance of subset X corresponding to the labels Y. Assuming P is the selected feature set in the forward greedy processing, F is the full feature set and x_i is the current candidate feature to be evaluated. The feature selection in each iteration of forward greedy searching can be formulated as follows:

$$x^* = \arg\max_{x_i} [D(P \cup \{x_i\}) - D(P)], \quad \forall x_i \in F - P.$$

Consider the example presented in Section I, suppose the three features in this example are denoted as $F = \{f_1, f_2, f_3\}$. Based on the setting in this example, we have $D(\{f_3\}) > D(\{f_1\})$, $D(\{f_3\}) > D(\{f_2\})$, $D(\{f_1, f_2\}) > D(\{f_3\})$, $D(\{f_1, f_2\}) > D(\{f_1, f_3\})$, and $D(\{f_1, f_2\}) > D(\{f_2, f_3\})$. In the forward greedy searching, if we use D as the evaluation function, it is obviously that f_3 will be selected firstly[2] and will lead to the weaker subset in feature selection. To solve this problem, we propose to reformulate the evaluation function as

$$E(X) = D(X) - \alpha D(F - X)$$

where α is a nonnegative balancing coefficient. The new evaluation function calculates the distance of the relevance evaluation between the selected feature subset and the remaining subset, we call the new evaluation function the margin of subsets. We evaluate the gain of the feature x_i with the new evaluation function as follows:

$$G(x_i) = E(P \cup \{x_i\}) - E(P) = [D(P \cup \{x_i\}) - D(P)] + \alpha [D(F - P) - D(F - P - \{x_i\})]. \quad (1)$$

According to the above definition, we can see that this new evaluation function considers the gain of the candidate feature x_i on both the selected feature subset P and remaining feature subset $F - P - \{x_i\}$.

In the new evaluation function, $\alpha D(F - X)$ measures the redundancy of current candidate feature. In the first iteration of the forward greedy processing when $P = \phi$, the gain on old evaluation function $D(X)$ will only evaluate the relevancy of each single feature with respect to the labels. While in our new evaluation function, the term $\alpha[D(F - P) - D(F - P - \{x_i\})]$ measures the redundancy of x_i with respect to the full feature set F. This means that although x_i is highly relevant with the label individually, if x_i is redundant, the value of $\alpha[D(F) - D(F - \{x_i\})]$ will penalize the total gain of the single feature x_i.

[2]In the first iteration of greedy searching, the gain of f_3 $(D(f_3) - D(\phi))$ is larger than the other two.

Thus, our new evaluation function measures the relevance and redundancy of candidate feature simultaneously. And α can be used to adjust the ratio of redundancy evaluation terms. In example 1, let us consider two gain functions of f_2 and f_3, respectively, in the first round of iteration with our new evaluation function.

The gain of f_3 with new function is

$$G(f_3) = E(f_3) - E(\phi) = D(\{f_3\}) - D(\phi) - \alpha \left[D(\{f_1, f_2\}) - D(\{f_1, f_2, f_3\})\right].$$

Gain of f_2 with new function

$$G(f_2) = E(f_2) - E(\phi) = D(\{f_2\}) - D(\phi) - \alpha \left[D(\{f_1, f_3\}) - D(\{f_1, f_2, f_3\})\right].$$

The difference of the gain between feature subset f_3 and f_2 is

$$G(f_3) - G(f_2) = \left[D(\{f_3\}) - D(\{f_2\})\right] - \alpha \left[D(\{f_1, f_2\}) - D(\{f_1, f_3\})\right].$$

In the above example, although $D(\{f_3\})$ is larger than $D(\{f_1\})$ and $D(\{f_2\})$, it is actually redundant with respect to $\{f_1, f_2\}$. The α is the penalty factor which controls the weight of the penalty term. It also means that a proper α can lead to $G(f_3) - G(f_2) < 0$, thus the forward greedy searching algorithm in the first iteration will select f_2 instead of f_3.

The basic idea of the new feature evaluation function is trying to maximize the relevance based gain margin of the candidate feature subset and remaining subset with respect to the labels.

III. Theoretical Framework for Feature Selection With Dependency Margin

In Yu and Liu's work [3], they build the optimal feature set by selecting relevant and nonredundant features. Their definition on optimal feature set is based on Koller and Sahami [4] and Kohavi and John's [15] works on feature redundancy, some related definitions are reviewed as follows.

Definition 1 (Conditional Independency [37]): Let U be a finite set of variables with discrete values, and X, Y, Z stand for any three subsets of variables in U. X and Y are conditionally independent given Z if

$P\left(X = \hat{x}|Y = \hat{y}, Z = \hat{z}\right) = P\left(X = \hat{x}|Z = \hat{z}\right)$ whenever $P\left(X = \hat{x}, Y = \hat{y}\right) > 0$.

We use notion $I(X, Y|Z)$ to denote the conditional independency of X and Y given Z. In the following sections, we also use $P\left(\hat{x}|\hat{z}\right)$ to represent $P\left(X = \hat{x}|Z = \hat{z}\right)$.

Based on the definition of conditional independency, Pearl [37] had defined several properties as follows.

Theorem 1: Let X, Y, Z be three disjoint subsets of variables from U, then

Symmetry

$$I(X, Y|Z) \Leftrightarrow I(Y, X|Z).$$

Decomposition

$$I(X, Y \cup W|Z) \Rightarrow I(X, Y|Z) \text{ and } I(X, W|Z).$$

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

Weak union

$$I(X, Y \cup W|Z) \Rightarrow I(X, Y|Z \cup W).$$

With the concept of conditional independency, several definitions on relevant features are listed below.

Let F be a full set of features, Y is the label, f is a feature and $S_i = F - \{f\}$. The concepts of feature relevance can be formalized as follows.

Definition 2 (Strong Relevance [3]): f is strong relevant iff there exists some $\hat{f}$, $\hat{y}$ and $\hat{s}_i$ for which $P(f = \hat{f}, S_i = \hat{s}_i) > 0$ such that $P(Y = \hat{y}|f = \hat{f}, S_i = \hat{s}_i) \neq P(Y = \hat{y}|S_i = \hat{s}_i)$.

Definition 3 (Weak Relevance [3]): f is weakly relevant iff it is not strong relevant, and there exists a subset of feature S'_i of S_i for which there exists some $\hat{f}$, $\hat{y}$, and $\hat{s}'_i$ for which $P(f = \hat{f}, S'_i = \hat{s}'_i) > 0$ such that $P(Y = \hat{y}|f = \hat{f}, S'_i = \hat{s}'_i) \neq P(Y = \hat{y}|S'_i = \hat{s}'_i)$.

According to the same notion, the irrelevance can be defined as follows.

Definition 4 (Irrelevance [3]): f is irrelevant iff $\forall S'_i \subseteq S_i, P(Y = \hat{y}|f = \hat{f}, S'_i = \hat{s}'_i) = P(Y = \hat{y}|S'_i = \hat{s}'_i)$.

According to the definitions, strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance in Definition 4 indicates that the feature is not necessary at all. As those weak relevant features are difficult to be determined whether they belongs to the optimal subset or not only from the viewpoint of relevance, Yu and Liu [3] tried to solve this problem from the viewpoint of redundancy.

In Yu and Liu's [3] redundant feature theories, they introduce the Markov blanket to describe the feature redundancy and the definition of Markov blanket is given as follows.

Definition 5 (Markov Blanket [4]): Given a feature f, let $M_i \subset F(f \notin M_i)$, M_i is said to be a Markov blanket for f iff

$$I(F - M_i - \{f\} \cup Y, f|M_i).$$

Definition 6 (Redundancy Feature [3]): Let F be the current set of features, a feature f is redundant and hence should be removed from F iff it is weakly relevant and has a Markov blanket M_i within F, that is $M_i \subset F - \{f\}$.[3]

According to the Definition 6, we can judge which of weakly relevant features should be selected and which of them removed to construct the optimal feature subset.

Thus, the optimal feature set [3] can be demonstrated with Fig. 2. There are four disjoint parts for an input feature set: 1) irrelevant features; 2) weakly relevant but redundant features; 3) weakly relevant but nonredundant features; and 4) strong relevant features. The optimal feature set is consisted of 3) and 4).

IV. Feature Selection With Dependency Margin

The proposed feature selection approach employs the forward greedy searching to generate a sequence of features.

[3]Here M_i should not equal to $F - \{f\}$, otherwise all the weakly relevant feature can find the Markov blanket $M_i = F - \{f\}$ based on Definition 3.

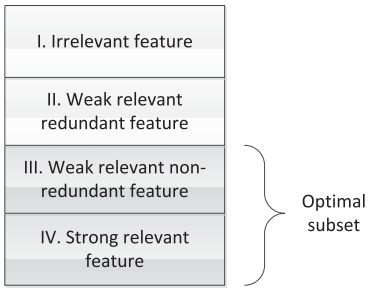


Fig. 2. View of feature relevance and redundancy [3].

The user can either choose top k features as the desired feature subset or use wrapper method to choose the best feature subset with corresponding classifier from the nested subsets constructed with the sequenced features. For example, suppose a sequence of features generated by our method is $F = \{f_1, f_2, \ldots, f_n\}$, then we can generate n nested candidate feature subsets from that sequence as $\{f_1\}, \{f_1, f_2\}, \ldots, \{f_1, f_2, \ldots, f_n\}$. Each candidate subset is evaluated with tenfold cross validation and the subset with highest classification accuracy is output.

A. Definitions on Dependency and Dependency Margin

In this section, we will present the definitions of dependency and dependency margin within the Bayesian framework, and several theorems based on those definitions are also presented.

Definition 7 (Dependency): Suppose F is the feature set, Y is the label, for a set $X \subseteq F$, the dependency of X is defined as

$$D(X) = \sum_{\forall \hat{x}_i, \hat{y}_j} P\left(X = \hat{x}_i, Y = \hat{y}_j\right)_{\{P(X=\hat{x}_i) = P(X=\hat{x}_i, Y=\hat{y}_j)\}}.$$

It is also can be denoted as

$$D(X) = \sum_{\forall \hat{x}_i, \hat{y}_j} P\left(X = \hat{x}_i, Y = \hat{y}_j\right)_{\{P(Y=\hat{y}_j|X=\hat{x}_i)=1\}}.$$

Here we let $D(\phi) = 0$.

Definition 8 (Dependency Margin): F is the feature set, Y is the label, for a set $X \subseteq F$, the dependency margin of X is defined as $E(X) = D(X) - \alpha D(\neg X)$.[4] Here, $\neg X = F - X$.

Theorem 2 (Monotonicity of Dependency): F is the feature set, Y is the label, considering two feature subsets $M \subseteq F$, $Q \subseteq F$. If $M \subseteq Q$, then $D(M) \leq D(Q)$.

Proof: If $M = Q$, obviously $D(M) = D(Q)$. Thus, we only need to consider the condition of $M \subset Q$, let us consider $M \cup \{f\} = Q$, $f \in F$. For any of $\hat{m}_i, \hat{y}_i$ that satisfies $P(\hat{y}_i, \hat{m}_i) = P(\hat{m}_i)$, $P(\hat{m}_i) = \sum_{\forall \hat{f}} P(\hat{m}_i, \hat{f}) = \sum_{\forall \hat{f}} P(\hat{m}_i, \hat{f}, \hat{y}_i)$, let $\hat{q}_i = (\hat{m}_i, \hat{f})$, then for each $\hat{q}_i$, we have $P(\hat{y}_i|\hat{q}_i) = 1$. This means $P(\hat{m}_i)$ will be always included in $D(Q)$ for any $\hat{m}_i, \hat{f}$ that satisfies $P(\hat{y}_i, \hat{m}_i) = P(\hat{m}_i)$. Thus, $D(M) \leq D(Q)$. ■

Based on Theorem 2, we can easily obtain the monotonicity of dependency margin as follows.

Theorem 3 (Monotonicity of Dependency Margin): F is the feature set, Y is the label, considering two feature subsets

[4]In this solution, we set $\alpha = 1$.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Strong Relevant Features Generation Algorithm

Input: F
Output: Strong relevant feature set A_1
1 $A_1 \leftarrow \phi$;
2 **for** *each* $a \in F$ **do**
3 **if** $D(F) > D(F - \{a\})$ **then**
4 $A_1 = A_1 \cup a$;
5 **end**
6 **end**

Algorithm 2: Approximate Weakly Relevant Nonredundant Feature Set Generation Algorithm

Input: F, A_1
Output: Approximate weakly relevant nonredundant feature set B_2
1 $B_2 \leftarrow \phi$;
2 **for** *each* $b \in F - A_1$ **do**
3 **if** $\forall m \in F - \{b\}$ *satisfied* $D(F - \{m\}) > D(F - \{m\} - \{b\})$ **then**
4 $B_2 = B_2 \cup b$;
5 **end**
6 **end**

$M \subseteq F$, $Q \subseteq F$ and $M \neq \phi$, $Q \neq \phi$. If $M \subseteq Q$, then $E(M) \leq E(Q)$.

Considering the definitions of dependency and conditionally independent, we can easily present the following corollary.

Corollary 1: F is the feature set, Y is the label, and $G = F - \{f\}$, if $D(F) > D(G)$, then $I(Y, f|G)$ will not be satisfied.

B. Feature Subsets Division

Based on the dependency margin, we develop the forward greedy search algorithm for feature selection as follows. Our approach has two stages, the first one is a preprocessing step to divide the whole feature set into three disjoint subsets, A_1, B_2 and C_3, where A_1 is the strong relevant feature set, B_2 is the approximate weakly relevant but nonredundant feature set and $C_3 = F - A_1 - B_2$ is the set of remaining features which may contain weakly relevant but redundant features and irrelevant features. In the second stage, the forward greedy based feature selection algorithm is applied to A_1, B_2, and C_3 sequentially to make sure those strong relevant features and weakly relevant but nonredundant features can be selected earlier than those weakly relevant but redundant features and irrelevant features.

The strong relevant features generation algorithm is detailed in Algorithm 1.

When considering Algorithm 1, we have the following theorem.

Theorem 4: F is the feature set, Y is the label, $f \in F$, if $D(F) > D(F - \{f\})$, then f is a strong relevant feature of F.

Proof: If $D(F) > D(F - \{f\})$, based on the Corollary 1, it is easy to see Y and f is not conditionally independent given $F - \{f\}$, and thus f is a strong relevant feature of F based on the Definition 2. ■

Theorem 4 can guarantee each feature in A_1 is a strong relevant feature, however, it cannot guarantee that all the strong relevant features are selected into A_1. We then use Algorithm 2 to select an approximate weakly relevant nonredundant feature set, B_2, which may contain strong relevant features. The approximate weakly relevant but nonredundant feature set generation algorithm is shown in Algorithm 2.

Considering Algorithm 2, we have the following theorem.

Theorem 5: F is the feature set, Y is the label, each element b in B_2 does not contain Markov blanket M_i within F ($M_i \subset F - \{b\}$).

Proof: Assuming $b \in B_2$ and there is a Markov blanket $M_i (M_i \subset F - \{b\})$ for b within F, that is $I(b, F - M_i - \{b\} \cup Y|M_i)$. For any $m \in F - M_i - \{b\}$, we have $I(b, \{m\} \cup Y|F - \{m\} - \{b\})$ based on the property of weak union. Thus, we have $I(b, Y|F - \{m\} - \{b\})$ based on the property of decomposition. Obviously, this will conflict with $D(F - \{m\}) > D(F - \{m\} - \{b\})$. Thus, b has no Markov blanket within F based on Corollary 1. ■

Obviously the features in B_2 satisfy the definition of weak relevance, thus, the feature in B_2 is either a strong relevant feature or a weakly relevant but nonredundant feature.

C. Forward Greedy Feature Selection With Dependency Margin

The feature selection algorithm with dependency margin is given in Algorithm 3. In this algorithm, we first select the features in A_1 according to the dependency margin function. Then a forward greedy searching is used to choose the feature that maximally increase the dependency margin function iteratively (step 5). After all the feature in A_1 are selected, B_2 and C_3 set are sequentially selected with the gain of dependency margin function.

D. Further Analysis on the Margin-Based Algorithm

Compared with traditional forward greedy based feature selection algorithms, our margin based feature selection algorithm will use three disjoint sequent subsets as its input and thus to select those strong relevant and weakly relevant but nonredundant features earlier than other features.

Based on the definitions of Markov blanket and dependency, we can obtain the following theorem.

Theorem 6: P is the feature subset of F, Y is the label, $f \in F$ and $f \notin P$, if $D(P \cup \{f\}) > D(P)$, then there does not exist M_i ($M_i \subset P$) to be a Markov blanket for f.

Proof: We first assume there is a subset M_i ($M_i \subset P$) to be a Markov blanket for f, then $I(P - M_i \cup Y, f|M_i)$. Based on the symmetry of Theorem 1, we obtain $I(f, P - M_i \cup Y|M_i)$, and then we can obtain $I(f, Y|P)$ based on the weak union of Theorem 1. Now we consider the condition that $D(P \cup \{f\}) > D(P)$, it is easy to obtain that $I(Y, f|P)$ (that is $I(f, Y|P)$)cannot be satisfied based on the Corollary 1. Thus, M_i does not exist. ■

Theorem 6 tell us that if $D(P \cup \{f\}) > D(P)$, f cannot be a redundant feature with respect to $P \cup \{f\}$ in the Bayesian feature redundancy theoretical framework. And let us reconsider the gain function presented in (1): when the current selected

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

Algorithm 3: Dependency Margin-Based Feature Selection Algorithm

Input: A_1, B_2, C_3
Output: Sequenced feature set P

$P \leftarrow \phi$;
$Q \leftarrow A_1$;
while $Q \neq \phi$ **do**
  for *each* a *in* $Q - P$ **do**
    $a^* = \arg\max_a [E(P \cup \{a\}) - E(P)]$;
  end
  $P \leftarrow P \cup \{a^*\}$;
  $Q \leftarrow Q - \{a^*\}$;
end
$Q \leftarrow B_2$;
while $Q \neq \phi$ **do**
  for *each* a *in* $Q \cup A_1 - P$ **do**
    $a^* = \arg\max_a [E(P \cup \{a\}) - E(P)]$;
  end
  $P \leftarrow P \cup \{a^*\}$;
  $Q \leftarrow Q - \{a^*\}$;
end
$Q \leftarrow C_3$;
while $Q \neq \phi$ **do**
  for *each* a *in* $Q \cup A_1 \cup B_2 - P$ **do**
    $a^* = \arg\max_a [E(P \cup \{a\}) - E(P)]$;
    if *there are multiple equal maximal* $a_i (i > 1)$ **then**
      $a^* = \arg\max_{a_i} D(P \cup \{a_i\})$;
    end
  end
  $P \leftarrow P \cup \{a^*\}$;
  $Q \leftarrow Q - \{a^*\}$;
end

feature set $P = \phi$, the gain function has calculate the term $D(F) - D(F - \{f\})$. Obviously $D(F) - D(F - \{f\}) \geq 0$ based on Theorem 2, and the condition that f is redundant with respect to F may occur only when $D(F) - D(F - \{f\}) = 0$, thus $D(F) - D(F - \{f\})$ in the gain function can be regarded as a penalty term to reduce the total gain when f is redundant with respect to F.

The main reason that traditional relevance based forward greedy searching approaches often selects weaker feature subset is they always choose those redundant but highly label-related features at the very beginning of the forward greedy searching. Compared with traditional approaches, our dependency margin based approach selects the feature that is not redundant with respect to both selected subset and remaining subset. This makes it more likely to converge to the global optimal subset.

As directly selecting optimal feature subset with the concept of Markov blanket is intractable, we introduce the dependency margin based feature selection algorithm, which first finds the strong relevant and weakly relevant but nonredundant feature subsets and then ranks the features by each feature's gain of dependency margin. According to the Definitions 2 and 3, Theorem 4 cannot guarantee that all the strong relevant features are selected into A_1, similarly, Theorem 5 also cannot guarantee that all the weakly relevant but nonredundant features are selected into B_2. Assuming P^* is the optimal feature subset presented in Fig. 2, then we have $(A_1 \cup B_2) \subseteq P^*$. It means C_3 may contain few weakly relevant but nonredundant features, which should also belong to the optimal feature subset P^*. That's why we also rank the features from C_3 in Algorithm 3.

V. Experiments

In this section, we empirically evaluate our approach [feature selection with dependency margin of subsets (FSDMS)] with other current state-of-art methods. In the following experiments, we use seven feature selection methods, which are ReliefF, consistency, dependency, information gain, correlation-based feature subset selection-sequential forward (CFS-FS), fast correlation based filter (FCBF), and IRelief, to carry out the comparable experiments. For the ReliefF, consistency, dependency, information gain, IRelief, and our FSDMS methods, we use forward greedy searching to generate the nested feature subsets, such as $\{f_1\}, \{f_1, f_2\}, \ldots, \{f_1, f_2, \ldots, f_n\}$. Then a wrapper based method is used to evaluate those feature subsets (to use desired classifier to test each candidate subset with tenfold cross validation and output the subset with best classification accuracy) and output the final feature subset. As the wrapper based method is implemented to our FSDMS, we also call it WFSDMS.

The detailed reviews on every comparable feature selection methods are given as follows.

1) *ReliefF [38]*: It searches for nearest neighbors of instances of each class and weights features according to how well they distinguish instances of different classes. In our experiments, we employ reliefF based evaluation function and forward greedy search to output the nested feature subsets and then use the same wrapper policy to output the optimal subset with highest classification accuracy among all the nested subsets.
2) *Consistency [5]:* The consistency based feature selection was originally presented by Dash and Liu [18] as a filter based approach. It evaluates the feature subsets based on the consistency with the labels. Hu *et al*. [5] improved the consistency based approach, which can enable the computation of consistency with numerical data. Hu *et al*. [5] also employed a forward greedy search policy to generate the sequence of features based on their gain of consistency to the selected features, then the wrapper policy is implemented in the nested subsets based on the feature sequence generated by the gain of consistency. In our experiments, we use Hu *et al*.'s [5] consistency method.
3) *Dependency [19]:* The dependency based feature selection defines the dependency as the discrimination of the subset on the labels and then uses a greedy forward searching policy to generate a sequence of features based on the dependency gain of each feature. The forward greedy searching is also used to generate the nested candidate feature subsets, and then the optimal subset is output via the wrapper method.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I
FEATURE SELECTED WITH DIFFERENT ALGORITHMS ON SYNTHETIC DATA, RELIEFF, CONSISTENCY, INFORMATION GAIN, IRELIEF, AND WFSDMS METHODS USING SVM, BAYES, AND CART AS WRAPPER CLASSIFIERS

	CorrAL	CorrAL-47	CorrAL-46
CFS-SF	A0, A1, B0, B1, **R**	A0, A1, B0, B1, **R**, **I1**	A0, A1, B0, B1
$FCBF_{(0)}$	**R**, A0, A1, B0, B1	**R** A0, A1, B0, B1, $\mathbf{I1_1}$	A0, A1, B0, B1
$FCBF_{(log)}$	**R**, A0, A1, B0, B1	**R**, A0, A1, B0, B1	A0, A1, B0, B1
Consistency-SVM	**R**, A0, A1, B0, B1	**R**, $\mathbf{A1_1}$, A0, $\mathbf{B0_1}$, B1, A1, B0	A0, A1
Consistency-Bayes	**R**, A0, A1, B0, B1	**R**, $\mathbf{A1_1}$, A0, $\mathbf{B0_1}$, B1, A1, B0	A0, A1
Consistency-CART	**R**, A0, A1, B0, B1	**R**, $\mathbf{A1_1}$, A0, $\mathbf{B0_1}$, B1, A1, B0	A0, A1
InfoGain-SVM	**R**, A0, A1, B0, B1	**R**, A0, A1, B0, B1	A0, A1, B0, B1
InfoGain-Bayes	**R**, A0, A1, B0, B1	**R**, A0, A1, B0, B1, $\mathbf{A0_0}$, $\mathbf{A1_0}$, $\mathbf{B0_0}$, $\mathbf{B1_0}$, $\mathbf{A1_1}$, $\mathbf{B0_1}$, $\mathbf{B1_1}$, $\mathbf{A0_1}$, $\mathbf{B1_2}$, $\mathbf{B1_3}$, $\mathbf{B0_3}$, $\mathbf{A0_3}$	A0, A1, B0, B1, $\mathbf{A0_0}$, $\mathbf{A1_0}$
InfoGain-CART	**R**, A0, A1, B0, B1	**R**, A0, A1, B0, B1	A0, A1, B0, B1
ReliefF-SVM	B0, A0, B1, A1	$\mathbf{A0_0}$, A1, $\mathbf{B1_0}$, A0, $\mathbf{A1_0}$, B1, $\mathbf{B0_0}$	$\mathbf{B0_0}$, B0, $\mathbf{A0_0}$, A0, $\mathbf{A1_0}$, A1, $\mathbf{B1_0}$
ReliefF-Bayes	B0, A0, B1, A1	$\mathbf{A0_0}$, A1, $\mathbf{B1_0}$, A0, $\mathbf{A1_0}$, B1, $\mathbf{B0_0}$	$\mathbf{B0_0}$, B0, $\mathbf{A0_0}$, A0, $\mathbf{A1_0}$, A1, $\mathbf{B1_0}$, B1, $\mathbf{A0_1}$, $\mathbf{A1_1}$
ReliefF-CART	B0, A0, B1, A1	$\mathbf{A0_0}$, A1, $\mathbf{B1_0}$, A0, $\mathbf{A1_0}$, B1, $\mathbf{B0_0}$	$\mathbf{B0_0}$, B0, $\mathbf{A0_0}$, A0, $\mathbf{A1_0}$, A1, $\mathbf{B1_0}$
IRelief-SVM	**R**,A0,B0,A1,B1	**R**,$\mathbf{B0_0}$,B0,$\mathbf{A1_0}$,A1,$\mathbf{A0_0}$,A0,$\mathbf{B1_0}$,B1,$\mathbf{A1_1}$,$\mathbf{B0_1}$,$\mathbf{A0_1}$	$\mathbf{B0_0}$,B0,$\mathbf{B0_5}$,B1,$\mathbf{A0_0}$,A0,$\mathbf{A1_0}$
IRelief-Bayes	**R**,A0,B0,A1,B1	**R**,$\mathbf{B0_0}$,B0,$\mathbf{A1_0}$,A1,$\mathbf{A0_0}$,A0,$\mathbf{B1_0}$,B1,$\mathbf{A1_1}$,$\mathbf{B0_1}$,$\mathbf{A0_1}$, $\mathbf{B1_1}$,$\mathbf{B1_2}$,$\mathbf{B0_2}$	$\mathbf{B0_0}$,B0,$\mathbf{B0_5}$,B1,$\mathbf{A0_0}$,A0,$\mathbf{A1_0}$,A1,$\mathbf{B1_1}$
IRelief-CART	**R**,A0,B0,A1,B1	**R**,$\mathbf{B0_0}$,B0,$\mathbf{A1_0}$,A1,$\mathbf{A0_0}$,A0,$\mathbf{B1_0}$,B1	$\mathbf{B0_0}$,B0,$\mathbf{B0_5}$,B1,$\mathbf{A0_0}$,A0,$\mathbf{A1_0}$
WFSDMS-SVM	A0, A1, B0, B1	A0, B0, B1 A1	A0, B0, B1, A1
WFSDMS-Bayes	A0, A1, B0, B1	A0, B0, B1 A1	A0, B0, B1, A1
WFSDMS-CART	A0, A1, B0, B1	A0, B0, B1 A1	A0, B0, B1, A1

4) *Information Gain:* This algorithm evaluates each attribute by measuring the information gain with respect to the class. This results in a sequence of features. We then apply the same wrapper method to output the optimal subset from those nested candidate subsets generated from the feature sequence. In our experiments, we use the information gain code from Weka [39].
5) *CFS-FS [25]:* It is a subset based evaluation algorithm, which exploits best-first search based on some correlation measure which evaluates the goodness of a subset by considering the individual predictive ability of each feature and the degree of correlation between them.
6) *FCBF [3]:* It is a filter based selection algorithm, which uses symmetrical uncertainty as the evaluation metric to find the relevant features and then removes those redundancy features with approximate Markov blanket.
7) *IRelief [40]:* Iterative Relief is an improved Relief algorithm. It addresses on Relief's weakness of lacking a mechanism to deal with outlier data and has been experimentally proved robust to the highly noisy data with a large amount of irrelevant features and/or mislabeling [40]. In our experiments, we use the IRelief code from machine learning python (MLPY).[5]

In order to fully evaluate the performance of the proposed feature selection algorithm, we also use three learning algorithms, naive Bayes classification (NBC), classification and regression tree (CART), SVM, to evaluate the predictive accuracy on the selected subset of features with a tenfold cross validation.

A. Experiments on Simulated Data

We use three synthetic datasets to evaluate the strength of WFSDMS and compare it with ReliefF, CFS-CF, consistency, dependency, information gain, FCBF, and IRelief. The fist dataset is CorrAL data which has been widely used in many feature selection evaluations [3], [34]. It contains six Boolean features ($A0, A1, B0, B1, I, R$) and a Boolean class Y defined by $Y = (A0 \wedge A1) \vee (B0 \wedge B1)$. Feature $A0$, $A1$, $B0$ and $B1$ are independent to each other, feature I is uniformly random, and feature R matches the class Y 75% of the time. In CorrAL dataset, the optimal subset includes $A0$, $A1$, $B0$, and $B1$. The Boolean features $A0$, $A1$, $B0$, and $B1$ have 16 combinations (instances), we then repeat each combination with the same probability to construct a CorrAL data with 1024 samples. The irrelevant feature I and redundant feature R are then generated using the dataset of 1024 instances.

We also use two other synthetic data, CorrAL-47 and CorrAL-46, which are generated by adding more irrelevant features and redundant features. The CorrAL-47 contains a total of 47 boolean features including five original features $A0$, $A1$, $B0$, $B1$, and R, 14 irrelevant features, and 28 additional redundant features. Among the 14 irrelevant features, only two features are uniformly random and each of the remaining 12 is completely correlated with one of the two features. Among the 28 additional redundant features, for each of $A0$, $A1$, $B0$ and $B1$, there are seven features that are correlated with it at various levels. The ratios of nonmatches are 0, 1/16, 2/16, ..., 6/16, respectively. CorrAL-46 is the same as CorrAL-47 except that it excludes R. The generation of the CorrAL-46 and CorrAL-47 is similar to the CorrAL, we extend the number of data instances discriminated by the optimal features $A0$, $A1$, $B0$, $B1$ into 1024 and then generate redundant features and irrelevant features.

Table I shows features selected by each algorithm. We use $A0$, $A1$, $B0$, $B1$ combined with subscripts 0, 1, ..., 6 to represent the newly introduced redundant features, with the value of the subscripts indicating the ratio of nonmatches. $I1$ and $I2$ are two irrelevant features using uniformly random. We use $I1$ or $I2$ combined with subscripts 1, 2, ..., 6 to represent irrelevant features which is completely correlated with $I1$ or $I2$.

In our experiment, there are two relevance threshold γ in FCBF, $FCBF_{(0)}$ and $FCBF_{(\log)}$. We set the γ as the default value 0 and the SU value of the $\lfloor N/\log N \rfloor$th ranked feature for each dataset, respectively.

We can see that most of the feature selection algorithms fail to remove the highly redundant feature R in CorrAL and CorrAL-47 except ReliefF and WFSDMS. However, the ReliefF tends to select several redundant features correlated with $A0$, $A1$, $B0$, and $B1$ in both CorrAL-46 and CorrAL-47. The feature R can be viewed as the redundant feature that can

[5]http://mlpy.sourceforge.net/docs/3.1/weighting.html

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DESCRIPTIONS OF UCI BENCHMARK DATASETS

	Data	Numerical	Catrgorial	Instances	Classes	Bayes	CART	SVM
1	Austr	8	6	690	2	68.59± 4.70	82.52± 4.50	85.62± 3.99
2	Blood	4	0	748	2	74.89± 4.40	73.78± 3.97	76.21± 0.41
3	*Car*	0	6	1728	4	80.54± 1.94	95.38± 1.50	88.43± 2.12
4	CrdApr	6	9	690	2	68.61± 5.22	83.26± 4.11	84.97± 3.80
5	*Hayes*	0	5	132	3	57.37± 12.16	77.76± 10.52	53.21± 11.02
6	Heart	7	6	270	2	79.81± 8.67	76.04± 7.05	81.67± 6.93
7	Iris	4	0	150	3	95.87± 4.99	94.87± 5.00	96.27± 4.38
8	Mamm	1	4	961	2	77.72± 3.85	79.58± 4.24	83.67± 3.41
9	Pima	8	0	768	2	73.49± 3.50	71.55± 5.56	76.47± 3.68
10	Seed	7	0	210	3	90.76± 6.05	92.05± 6.32	93.10± 5.39
11	Sonar	60	0	208	2	76.32± 7.74	70.55± 10.22	72.51± 8.23
12	*Voting*	0	16	435	2	88.80± 5.04	95.27± 2.89	95.47± 3.19
13	Wdbc	31	0	569	2	94.44± 2.94	91.97± 3.17	95.06± 2.53
14	Wine	13	0	178	3	97.64± 3.40	89.69± 6.167	98.37± 2.90
15	Wpbc	33	0	198	2	63.67± 10.68	70.22± 8.87	76.29± 2.00

TABLE III
CLASSIFICATION ACCURACY (IN PERCENT) OF UCI DATASETS WITH SUPPORT VECTOR MACHINE (SVM) CLASSIFIERS

Data	Raw	Dependency	Consistency	ReliefF	IRelief	InfoGain	CFS	FCBF	FCBF_log	WFSDMS
Austr	85.62± 3.99	85.75± 3.97	85.71± 4.17	85.72± 3.96	85.71±4.05	85.74± 4.22	85.71± 4.09	85.39± 4.00	85.25± 3.95	**85.93± 4.08**
Blood	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**	**76.21± 0.41**
Car	88.43± 2.12	77.78± 2.31	**89.28± 1.85**	88.56± 1.87	—	89.27± 1.88	88.47± 1.87	88.49± 1.77	80.43± 2.49	**89.28± 2.12**
CrdApr	84.97± 3.80	85.62± 4.34	85.78± 4.02	85.70± 3.62	85.70±3.87	85.51± 3.61	85.10± 3.69	85.59± 4.23	85.51± 3.64	**86.22± 3.75**
Hayes	53.21± 11.02	61.63± 13.80	61.52± 12.69	**62.02± 13.53**	—	58.47± 12.28	54.18± 11.88	57.87± 11.66	55.49± 13.27	61.63± 12.67
Heart	81.67± 6.93	80.00± 7.08	80.00± 8.14	82.30± 6.44	**84.11± 6.79**	82.15± 7.13	81.59± 6.33	80.56± 6.95	81.37± 7.76	81.74± 7.60
Iris	96.27± 4.38	96.00± 4.92	96.00± 5.19	96.27± 4.58	—	96.20± 4.67	87.40± 8.09	96.27± 4.67	**96.33± 4.68**	96.27± 4.26
Mamm	83.67± 3.41	83.72± 4.05	**83.74± 3.63**	83.72± 3.90	83.33± 3.33	83.62± 4.28	83.49± 3.78	83.27± 3.61	83.26± 3.69	**83.74± 3.81**
Pima	76.47± 3.68	77.47± 4.18	77.24± 3.71	**77.55± 4.13**	75.08± 3.69	77.08± 3.75	76.38± 3.68	76.50± 3.83	77.07± 3.77	77.27± 4.65
Seed	93.10± 5.39	92.86± 5.72	92.81± 4.76	93.19± 5.34	—	93.00± 6.33	91.05± 4.85	90.33± 6.29	90.38± 7.13	**93.24± 5.95**
Sonar	72.51± 8.23	72.93± 10.49	75.76± 9.85	**79.35± 7.72**	77.49 ± 6.02	75.28± 8.90	73.85± 8.24	73.06± 9.73	66.93± 9.10	73.99± 8.23
Voting	95.47± 3.19	95.63± 3.19	**95.64± 3.08**	95.63± 3.09	95.63 ± 3.00	95.63± 2.98	95.45± 2.69	95.61± 3.13	95.63± 3.30	**95.64± 2.93**
Wdbc	95.06± 2.53	95.36± 2.60	95.92± 2.25	**96.54± 2.07**	90.10 ± 3.64	95.32± 2.62	93.13± 3.08	94.55± 2.79	95.11± 3.00	95.78± 2.60
Wine	98.37± 2.90	95.30± 5.29	93.48± 6.21	98.35± 3.29	—	98.59± 2.57	98.43± 2.77	98.20± 3.18	95.57± 4.94	**98.77± 2.82**
Wpbc	76.29± 2.00	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**	**76.29± 2.00**
Ave.	**83.82**	**83.50**	**84.36**	**85.16**	—	**84.56**	**83.12**	**83.88± 4.55**	**82.72**	**84.80**

only be identified by a subset ($A0$, $A1$, $B0$, and $B1$), which is the type II redundant feature mentioned in previous section. This subset related redundant features often occurs in high-dimensional data and are hard to be removed by most heuristic search algorithms. The experiments show that our WFSDMS can find the optimal feature subset when there are redundant features that can only be identified based on feature subsets (e.g., the redundant feature R is defined by the subset of $A0$, $A1$, $B0$, and $B1$).

The results on IRelief show although it can avoid select irrelevant features in both CorrAL-46 and CorrAL-47, this algorithm fails to remove those relevant but redundant features such as R, $\mathbf{A0_0}$ etc.

We also notice that the consistency based methods often tend to select smaller feature subset in CorrAL-46, which does not contain the redundant feature R. The reason is the consistency metric tends to use an aggressive policy to estimate the classification risk of the feature subsets [5]. This policy can be viewed as a bias for small subset feature selection, which may enhance the classification performance of consistency based feature selection method, especially in some feature-number-sensitive classifiers. However, this bias may not be useful to find the true optimal feature subset.

We have also carried out the experiment on dataset with dependency. However, the dependency based selection algorithm failed to generate the nested candidate feature subsets due to the weak discrimination of each Boolean feature. That is the dependency of every feature is zero, when concerning the dependency of each feature in those three synthetic datasets.

B. Experiments on UCI Benchmark Data

In the following, we also employed 15 UCI benchmark datasets[6] to evaluate the performance of our WFSDMS and the other seven feature selection algorithms.[7] These datasets contain various numbers of features, instances, and classes, as shown in Table II. There are three datasets which contain purely nonnumerical features, and the remainders contain either purely continuous features or hybrid (discrete and continuous) features. To enable the dependency calculation of numerical features, we adopt a new method similar to Hu *et al.*'s [5] neighborhood approach according to Definition 8.

Definition 9 (Numerical Dependency): F is the feature set, Y is the label, for a set $X \subseteq F$, the dependency of X is defined as

$$D(X) = \sum_{\forall \hat{x}_i, \hat{y}_j} P(X = \hat{x}_i, Y = \hat{y}_j)_{\{P(Y=\hat{y}_j|X=\theta(\hat{x}_i))=1\}}$$

where $\theta(\hat{x}_i) = \left\{\hat{x}_j | \forall \hat{x}_j \text{ satisfied } \| \hat{x}_i - \hat{x}_j \| \leq \theta\right\}$, and $\theta > 0$, $||a - b||$ is the Euclid distance of a and b. $\hat{x}_j$ is the possible value of X in dataset.

The experimental results[8] are presented in Tables III–V. The results in Table III are evaluated with SVM, the results in Table IV are evaluated with CART, and the results in Table V are evaluated with Bayes method.

For the performance of SVM-based classification, as shown in Table III, WFSDMS comes with the highest accuracy in nine datasets. The average accuracy on the total 15 datasets are the second high among all the feature selection methods. After further analysis, the WFSDMSs accuracy on Sonar is much lower than the ReliefF', which may lower down the average accuracy of WFSDMS and lead to below the average accuracy of ReliefF. We use a diagram (Fig. 3), to visualize the

[6] http://archive.ics.uci.edu/ml/

[7] As the IRelief can only support two-classes datasets, we execute it in the ten two-classes datasets of Table II.

[8] As the suggestion by Hu *et al.* [5], we set $\theta = 0.14$. Each column of the dataset is normalized into [0, 1].

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
CLASSIFICATION ACCURACY (IN PERCENT) OF UCI DATASETS WITH CART CLASSIFIERS

Data	Raw	Dependency	Consistency	ReliefF	IRelief	InfoGain	CFS	FCBF	FCBF_log	WFSDMS
Austr	82.52± 4.50	84.67± 3.93	**86.52± 3.93**	86.09± 3.73	85.52± 3.64	83.45± 4.83	82.96± 4.03	85.57± 3.72	85.57± 3.86	85.70± 4.13
Blood	73.78± 3.97	75.94± 3.73	**76.53± 3.73**	76.31± 3.27	76.33± 1.80	75.79± 1.00	74.78± 4.10	72.74± 3.87	74.75± 2.20	76.39± 3.80
Car	95.38± 1.50	77.74± 2.38	95.35± 2.38	95.51± 1.48	—	95.42± 1.27	95.44± 1.32	95.39± 1.44	80.96± 2.43	**95.58± 1.44**
CrdApr	83.26± 4.11	84.51± 4.19	**86.62± 4.19**	85.99± 4.80	85.74± 4.14	83.54± 4.71	83.43± 4.47	85.46± 3.83	85.29± 3.93	85.35± 3.61
Hayes	77.76± 10.52	79.13± 9.13	79.08± 9.13	79.81± 7.92	—	78.35± 10.99	79.72± 9.97	77.71± 9.55	79.99± 7.48	**80.31± 7.16**
Heart	76.04± 7.05	80.19± 7.47	85.37± 7.47	80.96± 6.49	84.85± 5.98	77.07± 7.29	75.37± 7.49	84.00± 6.57	84.78± 6.57	**85.56± 6.67**
Iris	94.87± 5.00	95.53± 4.65	95.40± 4.65	94.47± 5.53	—	95.47± 5.98	93.00± 6.38	95.13± 5.59	95.20± 5.20	**95.67± 4.77**
Mamm	79.58± 4.24	79.57± 4.06	**83.86± 4.06**	81.31± 3.17	83.77± 4.08	83.77± 3.79	79.20± 3.81	83.16± 4.09	83.02± 4.17	83.84± 4.37
Pima	71.55± 5.56	71.46± 5.10	71.65± 5.10	71.43± 4.58	71.35± 5.07	**71.75± 4.68**	71.34± 4.17	70.12± 5.39	70.46± 5.42	70.99± 5.53
Seed	92.05± 6.32	89.33± 5.94	**93.52± 5.94**	91.81± 6.17	—	92.52± 6.14	90.76± 6.16	85.38± 7.57	86.33± 6.79	91.81± 6.02
Sonar	70.55± 10.22	72.39± 8.64	72.38± 8.64	**78.67± 10.05**	78.39± 8.06	78.02± 10.24	71.18± 11.03	71.36± 10.63	66.03± 9.71	75.45± 9.93
Voting	95.27± 2.89	95.93± 3.11	96.43± 3.11	**96.44± 3.13**	95.75± 3.50	95.80± 2.99	94.87± 3.15	94.67± 3.14	95.36± 3.02	95.38± 3.11
Wdbc	91.97± 3.17	93.23± 3.65	**94.64± 3.65**	94.24± 2.74	93.53± 3.01	93.04± 3.35	92.27± 3.45	93.50± 3.11	92.74± 3.22	94.48± 2.89
Wine	89.69± 6.17	92.46± 5.39	91.04± 5.39	**93.05± 6.40**	—	92.15± 5.82	90.82± 6.75	91.08± 6.30	91.41± 6.33	92.98± 5.19
Wpbc	70.22± 8.87	71.51± 9.69	71.14± 9.69	72.89± 8.40	71.75± 8.66	68.97± 9.29	69.03± 9.22	71.12± 8.97	68.80± 9.65	**73.09± 10.27**
Ave.	**82.97**	**82.91**	**85.30**	**85.27**	—	**84.34**	**82.95**	**83.76**	**82.71**	**85.51**

TABLE V
CLASSIFICATION ACCURACY (IN PERCENT) OF UCI DATASETS WITH BAYES CLASSIFIERS

Data	Raw	Dependency	Consistency	ReliefF	IRelief	InfoGain	CFS	FCBF	FCBF_log	WFSDMS
Austr	68.59± 4.70	69.27± 4.88	87.07± 3.84	85.84± 3.76	86.36± 4.38	69.36± 5.71	69.20± 5.13	86.61± 4.45	86.65± 3.68	**87.33± 4.11**
Blood	74.89± 4.41	75.66± 3.98	75.66± 4.49	74.65± 4.86	**76.83± 1.64**	75.63± 4.66	75.55± 4.62	68.63± 5.68	76.21± 0.41	75.66± 4.11
Car	80.54± 1.94	77.78± 2.32	82.40± 2.17	80.37± 1.86	—	**82.44± 2.51**	80.62± 2.01	80.50± 1.78	80.00± 2.25	82.31± 2.69
CrdApr	68.61± 5.22	68.81± 5.48	87.09± 2.99	85.51± 3.81	86.33± 3.87	69.11± 4.71	68.05± 5.54	86.81± 3.92	86.54± 4.03	**87.26± 3.82**
Hayes	57.37± 12.16	71.18± 12.45	70.04± 13.53	69.87± 13.57	—	57.99± 12.38	56.29± 11.19	58.76± 12.13	70.06± 13.70	**72.57± 12.96**
Heart	79.81± 8.67	78.67± 7.95	78.96± 7.29	79.93± 6.81	**82.89± 7.59**	79.93± 6.91	80.15± 7.09	82.63± 6.74	82.85± 7.27	79.85± 8.06
Iris	95.87± 4.99	**96.00± 4.92**	**96.00± 5.28**	95.87± 4.62	—	**96.00± 5.28**	89.27± 7.39	95.73± 4.79	95.87± 4.90	**96.00± 4.02**
Mamm	77.72± 3.85	77.81± 4.05	77.74± 4.06	77.79± 3.67	77.76± 4.05	77.73± 3.72	77.73± 4.01	76.80± 3.99	76.70± 4.43	**77.84± 3.78**
Pima	73.49± 3.50	**77.64± 4.45**	76.17± 4.01	76.79± 4.45	74.27± 4.12	76.76± 4.34	74.04± 3.85	73.46± 4.34	72.64± 3.72	**77.64± 4.04**
Seed	90.76± 6.05	**94.48± 4.48**	93.62± 5.34	90.62± 6.18	—	90.71± 6.29	91.10± 6.69	88.52± 5.94	88.43± 6.32	92.86± 4.95
Sonar	76.32± 7.74	75.19± 9.40	75.09± 9.26	78.74± 8.50	**80.71± 8.21**	79.33± 8.52	76.37± 9.52	74.98± 10.35	73.41± 9.37	77.11± 9.10
Voting	88.80± 5.04	94.47± 3.55	**95.64± 3.17**	95.63± 2.90	95.63± 2.87	**95.64± 2.83**	88.86± 4.90	91.32± 4.20	90.66± 3.80	91.50± 4.20
Wdbc	94.44± 2.94	96.63± 2.20	97.01± 2.23	96.10± 2.42	94.97± 2.91	94.38± 3.08	91.17± 3.23	91.42± 3.73	92.06± 3.22	**98.08± 2.79**
Wine	97.64± 3.40	97.52± 3.42	94.85± 4.74	97.48± 3.61	—	97.92± 3.26	97.70± 3.32	97.54± 3.48	96.22± 4.35	**98.09± 3.00**
Wpbc	63.67± 10.68	**78.17± 7.03**	77.20± 2.76	74.44± 8.26	67.69± 11.17	76.19± 2.29	65.51± 11.14	66.14± 10.90	63.04± 10.56	78.15± 8.06
Ave.	**79.24**	**81.95**	**84.30**	**83.97**	—	**81.27**	**78.77**	**81.32**	**82.09**	**84.82**

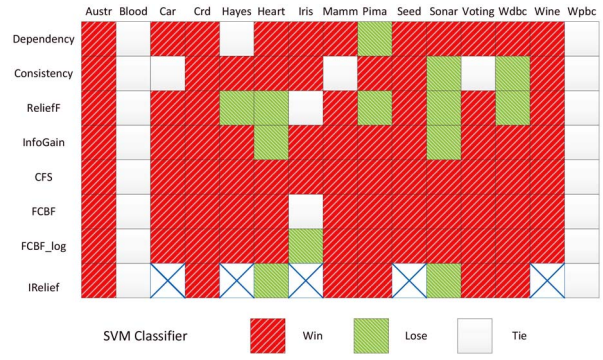


Fig. 3. Pair-wise accuracy comparison between WFSDMS and other methods with SVM classification on different datasets.

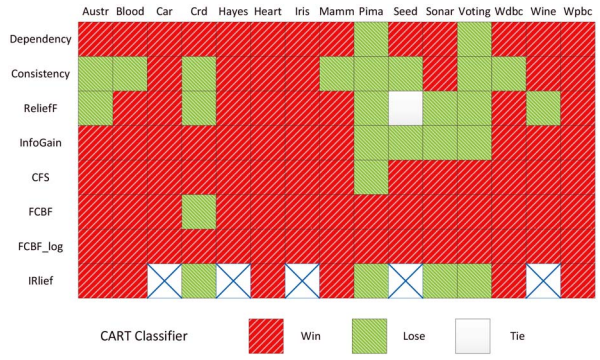


Fig. 4. Pair-wise accuracy comparison between WFSDMS and other methods with CART classification on different datasets.

pair-wise performances of our WFSDMS comparing with other methods with support vector machine (SVM) classification on different datasets. The red box in figure represents that WFSDMS'accuracy is higher than its comparable method's accuracy in current dataset. The green box shows WFSDMS' accuracy is lower than its comparable method's accuracy, and white box shows two methods are on par. The result in Fig. 3 shows the WFSDMS achieves better performance in most of dataset compared with other feature selection algorithms on SVM classification.

For the performance of CART-based classification, as shown in Table IV, WFSDMS comes with the highest accuracy in five datasets, which is a little below the consistency based feature selection methods (with six datasets). However, the average accuracy of WFSDMS is the highest in all the eight methods. Considering the pair-wise diagram shown in Fig. 4, WFSDMS can achieve better performance in most of the datasets. We then take a further analysis on the selected feature sets on WFSDMS and consistency methods, and the consistency methods tend to select smaller feature subsets comparing with WFSDMS. This may favor the short bias of the CART classification. However, we have proved in experiment 1, the short bias may be useful when considering improving the accuracy performance, it may not be able to find the true optimal feature subset.

With regard to the performance of Bayes-based classification, as shown in Table V, WFSDMS comes with the highest accuracy in eight datasets, which is more than all the other feature selection methods. The average accuracy of WFSDMS is the highest in all the eight methods. Considering the pair-wise diagram shown in Fig. 5, WFSDMS can achieve better performance in most of the datasets.

The UCI data experimental results on different classifiers also reminder us some interesting viewpoints. In the experiment of wrapping with SVM, there are two datasets, i.e., blood and Wpbc, which all the selection methods achieve the same classification accuracy as the original feature set, while these

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

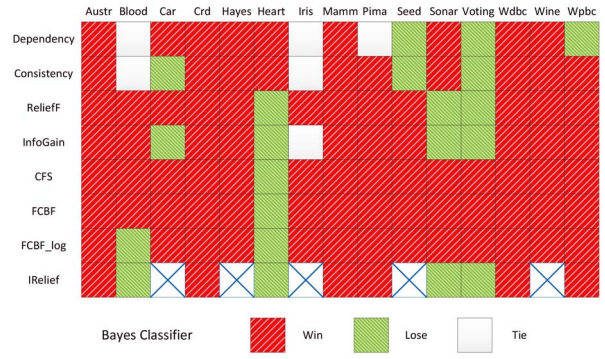


Fig. 5. Pair-wise accuracy comparison between WFSDMS and other methods with Bayes classification on different datasets.

two datasets training with original feature set cannot achieve the same classification accuracies when using the CART and Bayes. This may imply that SVM is not so sensitive to the input features as it will map all the features into a higher dimension with kernel functions. On the contrary, the Bayes is quite sensitive to the input features in classifications, this may explain why our method can achieve much better performance in the classify of Bayes than the other two classifiers.

Fig. 6 shows the relation between the number of selected features and the classification accuracies with different feature selection methods in wine and wpbc datasets. The results show our WFSDMS tends to select those features that are necessary (strong relevant) to the labels instead of selecting those features that are maximal positive related with the classification accuracy. This is why our WFSDMS performs much worse than other methods in the first selected feature, and then its classification performance will increase quickly with the increase of the number of selected features. This selection scheme can help our WFSDMS avoid the selection of highly relevant but redundant features, e.g., R in CorrAL, and find the optimal feature set.

C. Experiments on Face Recognition

In this section, we show results for applying our feature selection approach to the application of face recognition. In our experiments, the CMU PIE Database[9] is used. We manually remove background by cropping and registering all the faces[10] with a resolution of 64×64. Face examples in the database are shown in Fig. 9.

To handle the variations of illumination conditions and expressions in PIE database, the local binary patterns (LBP) [41] feature is used in our experiments. We divide each face (image) into 4×4 sub-regions (the size of each region is 16×16 pixels) and then calculate the $LBP^{u_2}_{8,1}$ features of every regions. Here, Superscript u_2 stands for using only uniform patterns [42], and $LPB_{P,R}$ stands for pixel neighborhoods which means P sampling points on a circle of radius of R.

The face recognition process combines all the LBP features of the 4×4 regions into a vector. This vector is used as the input features to train a classifier. In our experiments, each face region coding with the uniform patterns [41] will form a feature vector of 59 dimensions, thus the total dimension is 944 (59×16).

We then apply four feature selection approaches to find the most relevant features,[11] i.e., ReliefF, consistency, dependency, WFSDMS, into the face recognition process, and try to select the proper subset of face (image) regions. We randomly divide all the images of each person into three equal subsets and then use two subsets to apply feature selection approaches and train a SVM classifier, the third one subset is used as test set. We permutate between the training and testing datasets and then obtain the average classification accuracy and average number of selected features for each method. In our experiments, we repeat the random division 20 times. The results are shown in Figs. 7 and 8 where the horizontal axis, θ, is set from 3.1 to 5.1, this is based on the experimental observation of the dimensional ratio of the face recognition problem comparing with the previous experiments of UCI datasets, which contains lower dimensions and use $\theta = 0.14$ as their experimental optimal setting.

According to Fig. 7, the results show that the accuracies of our approach (WFSDMS) outperform most of other feature selection approaches and they can produce higher accuracy than the raw feature set. It is also suggested by Fig. 8 that our approach can reduce the face features (regions) more efficiently compared with other approaches with increasing θ. When θ increases to 5.1, the number of face regions (features) selected by our approach is reduced to minimum and at the same time our approach's accuracy reaches the maximum.

We also visualize the selected face regions by different feature selection approaches for further analysis. The results are shown in Fig. 9. The black blocks represent those unselected face regions. Each row in Fig. 9 represents the visualization result of a certain method drawing under different persons. The row 1, row 2, row 3, and row 4 represent the results of WFSDMS, dependency, consistency, and ReliefF, respectively. The results in Fig. 9 show that although our method selects least number of face regions (almost half of the 16 face regions) among all the four methods, it is still able to provide sufficient discrimination performance. Furthermore, the face and its division (4×4) in our experiments are symmetrical, which means that there may be some symmetrical face regions are redundant. And the results in Fig. 9 show that our method prefers to select asymmetrical face regions, which may remove those symmetrical redundancy in face recognition.

D. Analysis and Discussion

The simulated experimental results in Table I show that almost all of the feature selection algorithms can remove those irrelevant features, except two filter based methods, CFS-SF and $FCBF_{(0)}$, which implies that pure linear correlation based methods may encounter difficulties when processing

[9] http://vasc.ri.cmu.edu/idb/html/face/

[10] The dataset after processing can be downloaded from our website, http://www.nlict.zju.edu.cn/yliu/featureselection.html

[11] In this experiment, we compare the performance of wrapper based methods selecting sub-regions (features). As the calculation of the entropy for multiple dimensional vectors is intricate, this experiment does not include the method of InfoGain.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

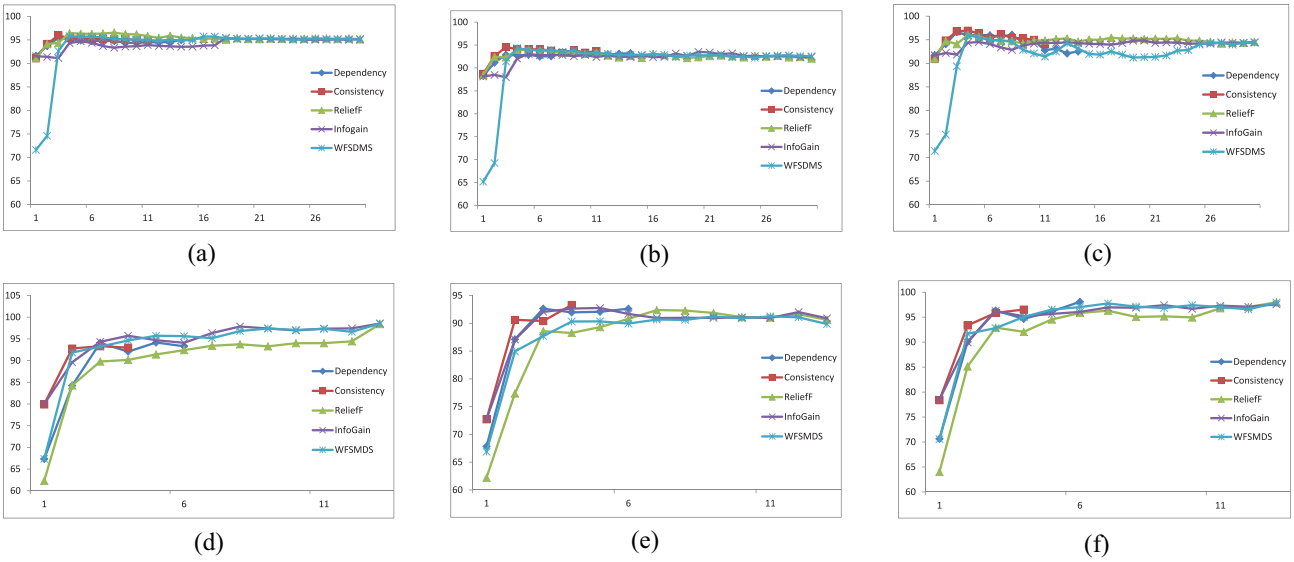


Fig. 6. Variation of average classification accuracies with the number of selected features. (a) Wpbc with SVM classification. (b) Wpbc with CART classification. (c) Wpbc with Bayes classification. (d) Wine with SVM classification. (e) Wine with CART classification. (f) Wine with Bayes classification.

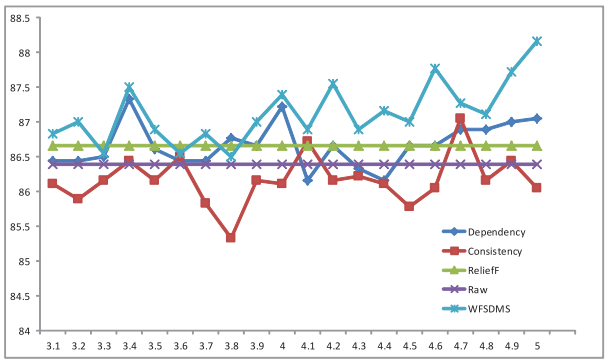


Fig. 7. Face recognition accuracy on different feature selection approaches with varied θ, $\theta \in [3.1, 5.1]$.

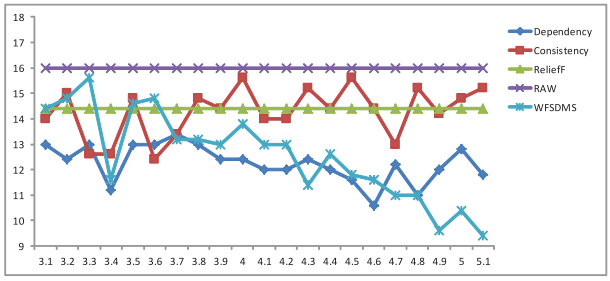


Fig. 8. Features selected by different approaches with varied θ, $\theta \in [3.1, 5.1]$.

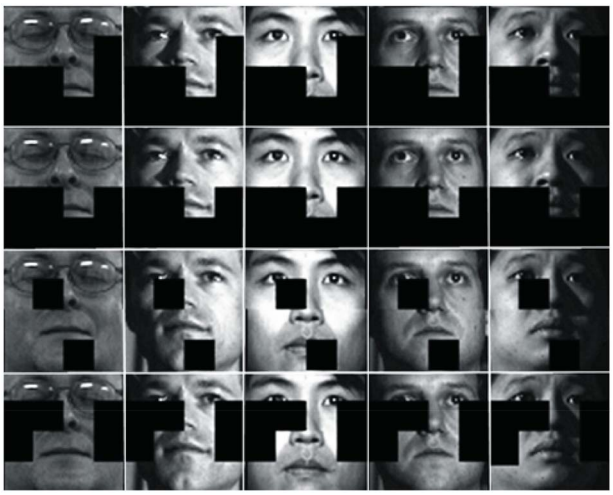


Fig. 9. Visualization of the selected features of different approaches ($\theta = 3.1$). The first row are the features selected by WFSDMS, the second row are the features selected by dependency, the third row are the features selected by consistency, and the last row are the features selected by ReliefF.

data with irrelevant features. Based on the observation of Table I, it also shows that the type II redundant feature will severely degrade the performance of feature selection algorithms. That's why those algorithms of CFS-CF, $FCBF_{(0)}$, $FCBF_{(\log)}$, Information Gain can almost find the optimal subset in Coral-46, while they fail to find the optimal subset in CorrAL and CorrAL-47. The results of those state-of-art algorithms in simulated experiments also reveal that current methods may already solve the problem of irrelevant feature removing, however, the remainder difficulty may be how to removing those relevant but redundant features especially in those data with type II redundant features.

Although it has been generally recognized that the classifier may not be the optimal evaluation criterion for the feature selection results due to its intrinsic bias, we still have no better choices especially when the ground true optimal feature subsets are not available. At least using varied classifiers working on large number of datasets to evaluate the performances of those feature selection methods can reveal the superior feature selection method in probability.

VI. Conclusion

This paper presents a feature selection algorithm using dependency margin of subsets. This dependency margin based

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

feature selection algorithm, which is estimated under the theories of Bayesian conditional independency and Markov blanket, employs independency to identify the strong relevant features and weakly relevant but nonredundant features and sorts the whole feature set with the sequence of strong relevant feature, weakly relevant but nonredundant features and remainder features. Then, a dependency margin based evaluation function is adopted to select features with forward greedy search algorithm from the sorted feature set. As the result, this approach can avoid the interference of relevant but redundant features and overcome the weakness (choosing weaker feature subset) of the forward greedy search based methods. Experimental results on simulated datasets, UCI benchmark datasets and practical face recognition case show the advantages of our method.

Several questions remained to be investigated in our feature works.

1) When processing the datasets with numerical values, a parameter θ, see Definition 8, is introduced to control the calculation of the dependency. Obviously, this parameter can be regarded as the granular control parameter and will affect the results of feature selection. Currently, we choose the value of θ empirically based on the dimension of problems. Further research will be on a principled and efficient way for parameter selection.
2) When concerning the relevance margin based evaluation function, there is a parameter α attached with the relevance function of unselected feature set. It is possible to discover the theories or models to choose proper value of α.

ACKNOWLEDGMENT

The authors would like to thank W. C. Liu for his preliminary work.

REFERENCES

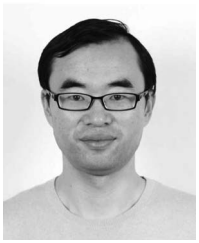
[1] H. Liu *et al.*, "Evolving feature selection," *IEEE Intell. Syst.*, vol. 20, no. 6, pp. 64–76, Nov./Dec. 2005.
[2] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, "Feature selection inspired classifier ensemble reduction," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1259–1268, Aug. 2014.
[3] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Jan. 2004.
[4] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn. (ICML)*, Bari, Italy, Jul. 1996, pp. 284–292.
[5] Q. Hu, W. Pedrycz, D. Yu, and J. Lang, "Selecting discrete and continuous features based on neighborhood decision error minimization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 137–150, Feb. 2010.
[6] J. M. Peña and R. Nilsson, "On the complexity of discrete feature selection for optimal classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1517–1522, Aug. 2010.
[7] S. Li and D. Wei, "Extremely high-dimensional feature selection via feature generating samplings," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 737–747, Jun. 2014.
[8] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013.
[9] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[10] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.
[11] X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using Laplacian regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2013–2025, Oct. 2011.
[12] J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1842–1849, 2008.
[13] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *J. Mach. Learn. Res.*, vol. 6, pp. 1855–1887, Jan. 2005.
[14] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.
[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.
[16] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proc. Speech Nat. Lang. Workshop*, San Francisco, CA, USA, 1992, pp. 212–217.
[17] H. Peng, F. Long, and C. H. Q. 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.
[18] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, nos. 1–2, pp. 155–176, 2003.
[19] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inform. Sci.*, vol. 178, no. 18, pp. 3577–3594, 2008.
[20] M. Bressan and J. Vitrià, "On the selection and classification of independent features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1312–1317, Oct. 2003.
[21] P. Maji and P. Garai, "Fuzzy-rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1166–1177, Aug. 2013.
[22] A. L. Oliveira and A. L. Sangiovanni-Vincentelli, "Constructive induction using a non-greedy strategy for feature selection," in *Proc. 9th Int. Workshop Mach. Learn.*, 1992, pp. 355–360.
[23] X. Geng, T. Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proc. 30th Annu. Int. ACM SIGIR Conf.*, 2007, pp. 407–414.
[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[25] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn. (ICML)*, Stanford, CA, USA, Jun./Jul. 2000, pp. 359–366.
[26] L. Yu and H. Liu, "Redundancy based feature selection for microarray data," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2004, pp. 737–742.
[27] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Jan. 2003.
[28] K. Javed, H. A. Babri, and M. Saeed, "Feature selection based on class-dependent densities for high-dimensional binary data," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 465–477, Mar. 2012.
[29] A. Appice, M. Ceci, S. Rawles, and P. A. Flach, "Redundant feature elimination for multi-class problems," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 5.
[30] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 853–858, May 2010.
[31] J. Biesiada and W. Duch, "Feature selection for high-dimensional data—A Pearson redundancy based filter," in *Computer Recognition Systems 2*. Berlin, Germany: Springer, 2008, pp. 242–249.
[32] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proc. IEEE Comput. Soc. Conf. Bioinformat. (CSB)*, Washington, DC, USA, Aug. 2003, pp. 523–528.
[33] P. Antal, A. Millinghoffer, G. Hullám, C. Szalai, and A. Falus, "A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction," in *Proc. Conf. Workshop J. Mach. Learn. Res.*, vol. 4. 2008, pp. 74–89.
[34] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Mach. Learn.*, New Brunswick, NJ, USA, Jul. 1994, pp. 121–129.
[35] L. Yu and H. Liu, "Redundancy based feature selection for microarray data," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD)*, New York, NY, USA, 2004, pp. 737–742.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.

[36] R. Duangsoithong and T. Windeatt, "Relevant and redundant feature analysis with ensemble classification," in *Proc. 7th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Kolkata, India, Feb. 2009, pp. 247–250.
[37] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann, 1988.
[38] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, 2003.
[39] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[40] Y. Sun, "Iterative relief for feature weighting: Algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, Jun. 2007.
[41] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[42] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

Yong Liu (M'11) received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science, both from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively.

He is currently an Associate Professor with the Institute of Cyber-Systems and Control, Department of Control Science and Engineering, Zhejiang University. His current research interests include machine learning, robotics vision, information processing, and granular computing. He has published over 30 research papers in machine learning, computer vision, information fusion, and robotics.



Feng Tang (M'09) received the bachelor's and master's degrees from the State Key Laboratory of Computer Aided Design and Computer Graphics, Zhejiang University, Zhejiang, China, in 2001 and 2004, respectively, and the Ph.D. degree from the University of California, Santa Cruz, Santa Cruz, CA, USA, in 2008.

He has been a Senior Researcher with Hewlett-Packard Laboratories, Palo Alto, CA, USA, since 2009. He has published over 40 papers in computer vision, machine learning, and multimedia.

Dr. Tang was the recipient of the Best Paper Award for Multimedia Modeling Conference in 2011 and the Best Paper Runner Up in the International Conference on Internet Multimedia Computing and Service in 2012.

Zhiyong Zeng received the B.S. degree in computer science and technology and the master's degree in computer science, both from Hangzhou Dianzi University, Hangzhou, China, in 2011 and 2014, respectively.

He is currently with the Meitu Corporation, Newton, MA, USA. His current research interests include deep learning and image processing. He has published several research papers in machine learning.

Authorized licensed use limited to: Zhejiang University. Downloaded on February 24,2021 at 05:28:41 UTC from IEEE Xplore. Restrictions apply.