Appl Intell (2012) 37:100–120
DOI 10.1007/s10489-011-0315-y

A novel feature selection method based on normalized mutual information

La The Vinh · Sungyoung Lee · Young-Tack Park · Brian J. d'Auriol

Published online: 23 August 2011

Abstract In this paper, a novel feature selection method based on the normalization of the well-known mutual information measurement is presented. Our method is derived from an existing approach, the max-relevance and min-redundancy (mRMR) approach. We, however, propose to normalize the mutual information used in the method so that the domination of the relevance or of the redundancy can be eliminated. We borrow some commonly used recognition models including Support Vector Machine (SVM), *k*-Nearest-Neighbor (*k*NN), and Linear Discriminant Analysis (LDA) to compare our algorithm with the original (mRMR) and a recently improved version of the mRMR, the Normalized Mutual Information Feature Selection (NMIFS) algorithm. To avoid data-specific statements, we conduct our classification experiments using various datasets from the UCI machine learning repository. The results confirm that our feature selection method is more robust than the others with regard to classification accuracy.

Keywords Feature selection · Mutual information · Minimal redundancy · Maximal relevance

1 Introduction

Feature selection is a technique for selecting a subset of relevant features, which contain information to help distinguish one class from the others, from a large number of features extracted from the input data. Feature selection is different from feature extraction [11], wherein a new set of features is formed by projecting the original feature space into a reduced-dimension space. In the present paper, we focus only on feature selection methods.

In pattern recognition, the identification of the most discriminative features is an important step [7], since it is common to have a large number of features, including relevant as well as irrelevant features, at the beginning of the pattern recognition process [11, 15]. Feeding a large set of features into a recognition model not only increases the computation burden but also causes the problem commonly known as the curse of dimensionality. Therefore, removing irrelevant features helps speed up the learning process and alleviates the effect of the curse of dimensionality. Due to the capabilities, feature selection has been largely applied in many applications, including text classification [6, 12], bio-informatics [8, 24, 32], intrusion detection [18, 27], and image retrieval [5, 9]. Furthermore, feature selection facilitates the data visualization and understanding [14, 17, 31].

So far, there is a great number of methods in the feature selection research area. Those methods can be categorized into three main directions namely *wrapper*, *embedded* and *filter*. *Wrapper* approaches [25, 29] make use of the classification accuracy to evaluate the usefulness of features at each step. However, repeatedly training such classifiers often requires high computational cost, making the *wrapper* based methods impractical with large datasets. Besides, the performance of *wrapper* approach may strictly depend on the classifier being used in the evaluation.

Embedded methods [4, 33] also use particular classifiers to find feature subsets. They, however, select features in the training phase of the classifier. Thus, *embedded* methods can utilize extra information of the cost function to guide the

L.T. Vinh · S. Lee (✉) · B.J. d'Auriol
Dept. of Computer Engineering, Kyung Hee University, Seoul, Korea
e-mail: sylee@oslab.khu.ac.kr

Y.-T. Park
School of IT, Soongsil University, Seoul, Korea
e-mail: park@ssu.ac.kr

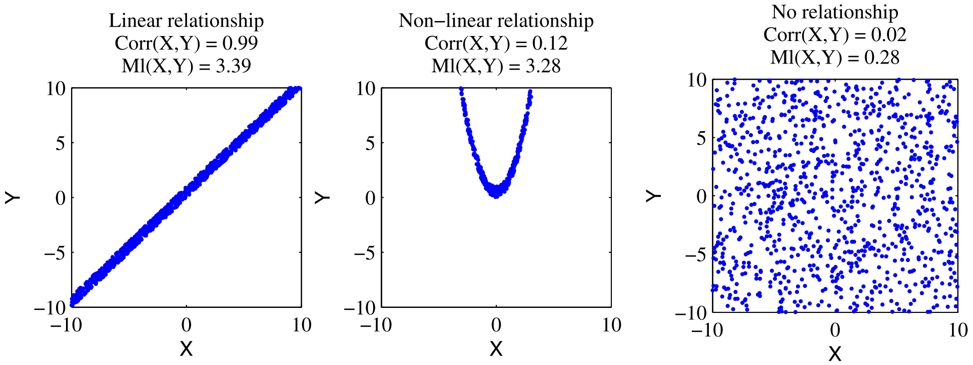

Fig. 1 Different types of the relationship between variables X and Y. The *left figure* shows a linear relationship captured by high values of both the correlation (Corr) and the mutual information (MI). The *middle figure* shows a non-linear relationship which is still well described by the high MI value, but Corr fails to reflect this relationship. The *right figure* shows two unrelated variables, hence both Corr and MI produce very low values

search direction. *Embedded* approaches are reported to be much faster than those of *wrapper*; the performance, however, also depends on the classifier [24].

Filter algorithms [2, 10, 23] utilize simple measurements such as correlation, mutual information to estimate the goodness of features. As a result, *filter* methods are classifier-independent and effective regarding computational cost. In the following paragraphs, we will provide more details of the feature evaluation criterion in filter algorithms.

In a correlation-based feature selector, a subset of features (S) is selected so that the below potential measurement is maximized

$$P_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{1}$$

where S is a subset of k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$), and $\overline{r_{ff}}$ is the average feature-feature inter-correlation. The correlation in (1) is computed by

$$\overline{r_{xy}} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y}, \tag{2}$$

where μ_x, μ_y, σ_x, and σ_y are the mean and standard deviation values of x and y, respectively. However, it is well known that the correlation is not able to describe non-linear relationships among variables as depicted in Fig. 1. Furthermore, the computation of (2) requires that all the features must be numerical variables, it is another weakness of the correlation-based feature selection method.

The information-based method utilizes a simple measurement, hence it has the advantage of low computation cost. In addition, as we point out below, the mutual information is capable of capturing non-linear relationships and is suitable for both numerical and categorical data. Therefore, in our work, we utilize the information measurement to estimate the potential of the features. On the topic of searching algorithms, since an exhaustive search over a large feature space is impractical, greedy forward selection and backward elimination are often used [2, 19, 23, 26]. Here, we exploit greedy forward selection, wherein each feature is appended to the feature set based on its quality.

The rest of our paper is organized as the following. In Sect. 2, we present some existing work in the area of mutual information based feature selection. Also in this section, we analyze the limitations of the previous work. After that we propose our method to overcome the addressed limitations in Sect. 3. Our experiments and discussions are presented in Sect. 4. Finally, our conclusions and future works are given in Sect. 5.

2 Related work

In this section we first present the fundamental background of mutual information based feature selection methods. After that we review some recently proposed methods in that area, and also point out their improvements as well as their limitations.

In mutual information based feature selection methods, mutual information is used to quantitatively analyze the relationship between any two features or between a feature and a class variable. The mutual information of two random variables X and Y is defined as

$$I(X;Y) = \int_{\Omega_Y}\int_{\Omega_X} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy, \tag{3}$$

where Ω_X and Ω_Y are the sample spaces of X and Y, $p(x)$, $p(y)$, and $p(x,y)$ are the probability density functions of

Fig. 2 Mutual information of categorical variables. The *left table* contains ten objects which have two categorical attributes A_1 and A_2. The *right table* shows the joint and marginal probabilities

Object	A_1	A_2
1	M	A
2	M	B
3	F	B
4	F	A
5	M	C
6	F	C
7	M	C
8	F	C
9	F	A
10	M	B

	A	B	C	$P(A_1)$
M	0.1	0.2	0.2	0.5
F	0.2	0.1	0.2	0.5
$P(A_2)$	0.3	0.3	0.4	

$$I(A_1,A_2)=P(M,A)\log_2\left(\frac{P(M,A)}{P(M)P(A)}\right)+P(M,B)\log_2\left(\frac{P(M,B)}{P(M)P(B)}\right)+$$
$$P(M,C)\log_2\left(\frac{P(M,C)}{P(M)P(C)}\right)+P(F,A)\log_2\left(\frac{P(F,A)}{P(F)P(A)}\right)+$$
$$P(F,B)\log_2\left(\frac{P(F,B)}{P(F)P(B)}\right)+P(F,C)\log_2\left(\frac{P(F,C)}{P(F)P(C)}\right)=0.05$$

X, Y, and (X,Y), respectively. In the case of discrete variables, the integration notation is replaced by the summation notation as

$$I(X;Y)=\sum_{y\in\Omega_Y}\sum_{x\in\Omega_X}p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{4}$$

Equation (4) computes the mutual information based on probability distributions of discrete variables, hence we can apply that to both numerical as well as categorical data. An example of computing the mutual information of categorical data is given in Fig. 2. Additionally, Fig. 1 demonstrates that non-linear relationships can be well described by the mutual information. The mutual information can also be represented by the entropy as

$$I(X;Y)=H(X)-H(X|Y), \tag{5}$$

where

$$H(X)=-\sum_{x\in\Omega_X}p(x)\log_2(p(x)) \tag{6}$$

and

$$H(X|Y)=-\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}p(x,y)\log_2(p(x|y)) \tag{7}$$

are entropy functions which measure the uncertainties of random variables. Based on (5), (6), and (7), we can define the mutual information as the amount of uncertainty in X which is removed by knowing Y. Figure 3 illustrates the relationship between the mutual information and the entropy.

In pattern recognition, our target is to determine the class label from feature values. Therefore, we expect a feature set that can remove as much of the uncertainty of the class variable as possible. This can be achieved by finding a feature set $S_i=\{X_1,X_2,\dots,X_i\}$ to maximize the following joint mutual information

$$I(S_i;C)=\sum_{c\in\Omega_C}\sum_{s_i}p(s_i)\log_2\left(\frac{p(s_i,c)}{p(s_i)p(c)}\right). \tag{8}$$

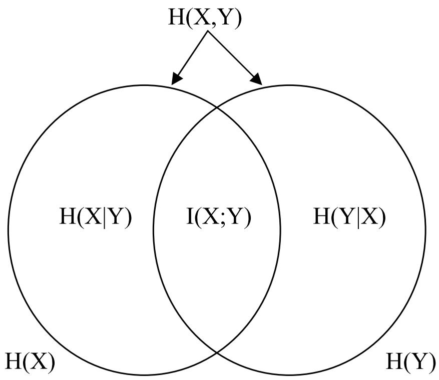

Fig. 3 The relationship between the mutual information and the entropy

Hereafter, we use C and X to denote the class and feature variables, respectively, and Ω_C is the set of all possible class labels.

Methods using (8) are referred as Max-Dependency (MD) approaches. Regardless of the searching algorithm, MD faces difficulties in estimating the multivariate density functions, which requires not only a high computational cost but also a large number of samples. For example, suppose that we have K features, each of which has an integer value from 1 to N. Then, to estimate the mutual information we have to know the joint density at each of K^N combinations. Therefore, an exponential complexity is required for the computation.

Because of this, Battiti proposed an heuristic approximation of MD. In his work, only bivariate mutual information functions were computed, including feature-feature mutual information $I(X_i;X_j)$ and class-feature mutual information $I(C;X_i)$ [2]. The selection criterion aimed at maximizing the class-feature mutual information (CFMI) and minimizing the feature-feature mutual information (FFMI). Since the CFMI represents the discrimination ability of a feature (relevance), while the FFMI contains information about the redundancy or the similarity among features, the method in [2] serves as a starting point for the later max-relevance and min-redundancy approaches [10, 19, 23]. Battiti's feature selection algorithm (MIFS) selects a feature (X_i) at each step so that the following feature potential measurement is maximized

$$f(X_i) = I(C; X_i) - \beta \sum_{X_s \in S_{i-1}} I(X_s; X_i), \tag{9}$$

where function f measures the goodness of a feature, S_{i-1} is the set of selected features in the previous $i-1$ steps, X_i is any non-selected feature, and β is a manually tuned parameter used to make the left and the right terms in the subtraction comparable.

In [19], the author analyzed the disadvantages of Battiti's criterion and then proposed an improved one, the MIFS-U, represented by

$$f(X_i) = I(C; X_i) - \beta \sum_{X_s \in S_{i-1}} \frac{I(C; X_s)}{H(X_s)} I(X_s; X_i). \tag{10}$$

Despite the improvement made by the later work, both of the above methods require a parameter (β) to be estimated manually. If β is too large, the right term dominates, so both algorithms tend to select features based on minimum redundancy. In contrast, if β is too small, the algorithms favor maximum-relevance features. Unfortunately, there is no way to optimize the value of β.

The authors of [23] presented a parameter-free feature selection algorithm called max-relevance and min-redundancy (mRMR). In [23], the authors pointed out that (8) can also be rewritten in the form

$$\begin{aligned} I(S_i; C) &= H(S_i, C) - H(S_i) \\ &= H(S_{i-1}, X_i, C) - H(S_{i-1}, X_i), \end{aligned} \tag{11}$$

where H is the joint entropy function. From (11), the authors of [23] showed that if a set of previously selected features (S_{i-1}) is given, Max-Dependency can be solved by finding X_i to maximize $H(S_{i-1}, X_i, C)$ and minimize $H(S_{i-1}, X_i)$. They then proved that the former is equal to maximizing $I(C; X_i)$, and the latter is corresponding to finding the minimum of $\sum_{X_s \in S_{i-1}} I(X_s; X_i)$. Hence, they come up with the below goodness measurement

$$f(X_i) = I(C; X_i) - \frac{1}{|S_{i-1}|} \sum_{X_s \in S_{i-1}} I(X_s; X_i). \tag{12}$$

The basic idea in [23] is similar to the one introduced by Battiti. However, Peng and his colleges provided a solid theoretical background of the method and eliminated the manually tuned parameter by averaging the feature-feature mutual information in the right term of the subtraction in (9). Although, mRMR does not always produce better results than do MIFS and MIFS-U [10], it eliminates the difficulty of parameter selection while producing results comparable to those of MIFS and MIFS-U.

Recently, the authors of [10] pointed out the drawback of mRMR, which was still the unbalance between the two terms of the subtraction. From (5) we can see that

$$\begin{aligned} I(C; X_i) &= H(C) - H(C|X_i) \le H(C) \\ &= -\sum_{c \in \Omega_C} p(c) \log_2(p(c)), \end{aligned} \tag{13}$$

where Ω_C is the sample space of the class variable C. Based on Jensen's inequality, it is clear that

$$I(C; X_i) \le \log_2\left(\sum_{c \in \Omega_C} p(c) \frac{1}{p(c)}\right) = \log_2(|\Omega_C|). \tag{14}$$

Therefore, in a two-class recognition problem ($|\Omega_C| = 2$), $I(C; X_i)$ is bounded in the range $[0, 1]$. By using similar proof we can conclude that

$$\frac{1}{|S_{i-1}|} \sum_{X_s \in S_{i-1}} I(X_s; X_i) \le \log_2(|\Omega_X|), \tag{15}$$

where Ω_X is the sample space of the features. Since $|\Omega_X|$ can have any arbitrary large value, the right term of the subtraction in (12) greatly varies and can dominate the left term (bounded in $[0, 1]$). In such a case, the algorithm is biased toward the less redundant features.

Based on the above observation, Pablo et al. introduced so-called normalized mutual information [10]. The authors showed that the mutual information between two random variables should be divided by the minimum value of the entropies in order to produce a normalized value in the range $[0, 1]$. Then they presented a selection strategy (NMIFS) using the following feature quality estimation

$$f(X_i) = I(C; X_i) - \frac{1}{|S_{i-1}|} \sum_{X_s \in S_{i-1}} \frac{I(X_s; X_i)}{\min(H(X_s), H(X_i))}. \tag{16}$$

It can be seen from (16) that NMIFS achieves a good balance between relevance and redundancy in two-class recognition systems, wherein both terms of the subtraction are within the range $[0, 1]$. Problems may occur when the number of classes increases [28]. In that case, the left-side mutual information breaks the upper bound and may dominate the right term. Hence, NMIFS may suffer from the same limitation as that in MIFS and MIFS-U when β is too small. Furthermore, because NMIFS assigns different normalizing weights to the features, it may select unexpected features. For example, if X_i and X_j are two features with the same relevance; X_i, however, is less random than X_j or $\frac{1}{H(X_i)} > \frac{1}{H(X_j)}$. In such a case, X_j may have a smaller weight in the right term of (16), making the overall potential $f(X_j)$ bigger than that of X_i; as a result, the method biases toward the noisier feature.

To summarize the common problem of the existing works [2, 10, 19, 23] in mutual information based feature selection, we reformulate the problem as the following: given a dataset with N features $X_1, X_2, \dots, X_N$, and a set of $i-1$ selected indexes ($S_{i-1} = \{s_1, s_2, \dots, s_{i-1}\}$), the next feature ($X_{s_i}$) is selected so that the redundancy ($RD(X_{s_i}) = \sum_{s \in S_{i-1}} I(X_s; X_{s_i})$) is minimized and the relevance ($RL(X_{s_i}) = I(C; X_{s_i})$) is maximized. However, because the two problems may not have a common solution, we would like to find a scale factor (β) so that a feature X_{s_i} maximizing $RL(X_{s_i}) - \beta \times RD(X_{s_i})$ will be a feasible solution for the minimization as well as the maximization. The existing solutions are summarized below

- MIFS and MIFS-U: β is manually selected by experiments,
- mRMR: $\beta = \frac{1}{|S_{i-1}|}$,
- NMIFS: $\beta(X_s; X_{s_i}) = \frac{1}{|S_{i-1}|} \times \frac{1}{\min(H(X_s), H(X_{s_i}))}$.

Although a significant improvement has been made [10], there are still some limitations of the existing works as we pointed out. Hence, in the next sections, we propose a new method to overcome those limitations.

Algorithm 1: Quantization algorithm

```
Input  : $M$—Total number of features
         $X(1..M)$—Training data
         $\xi$—The quantization error
Output: $N$—Number of quantization levels
         $Y(1..M)$—Quantized data
Quantization
    $N = 2$
    while 1 do
        $MaxError = -1e + 16$
        for $m = 1$ to $M$ do
            $Upper = \max(X(m))$
            $Lower = \min(X(m))$
            $Step = (Upper - Lower)/N$
            $Partition = [Lower : Step : Upper]$
            $CodeBook = [Lower - Step, Lower : Step : Upper]$
            $[Y(m), QError] =$
            $Quantiz(X(m), Partition, CodeBook)$
            if $Qerror > MaxError$ then
                $MaxError = QError$
        if $MaxError < \xi$ then
            Break;
        $N = N + 1$
end
```

3 The proposed method

As we discuss above, even though Pablo et al. proposed NMIFS to overcome the limitations of the previous methods including MIFS, MIFS-U and mRMR, there are still some situations in which NMIFS may cause unexpected feature selections. Therefore, in this section, our focus is to resolve the limitations of NMIFS addressed in Sect. 2.

Before going into the detail of our method, we first consider the upper bound of the mutual information of random variables. Since any continuous variable can be quantized into discrete form, we assume that two discrete random variables X and Y are given along with their marginal and joint distributions. Hence, the joint mutual information of X and Y is computed using (4). From (5) and (6), we can see that

$$I(X; Y) \le \min(H(X), H(Y)). \tag{17}$$

Applying Jensen's inequality to the definition of the entropy, we have

$$H(X) \le \log_2\left(\sum_{x \in \Omega_X} p(x) \frac{1}{p(x)}\right), \tag{18}$$

$$H(X) \le \log_2 (|\Omega_X|). \tag{19}$$

From (17) and (19), it is obvious that

$$I(X; Y) \le \min\bigl(\log_2(|\Omega_X|), \log_2(|\Omega_Y|)\bigr). \tag{20}$$

In our method, every feature is quantized using the same number of levels (N), which is decided so that the expected quantization error is achieved. The quantization algorithm is depicted in Algorithm 1 below. As can be seen, we gradually increase the number of quantization levels until the quantization error is smaller than a predefined small constant ξ, the expected quantization error. In our experiments, we selected $\xi = 0.01$ because smaller values did not make any improvement regarding the accuracy but created extra computation burden. From the algorithm, we can see that $|\Omega_X| = N$ for every feature X. Therefore

$$I(X; Y) \le \log_2(N). \tag{21}$$

Obviously, $\log_2(N)$ is an upper bound of the mutual information $I(X, Y)$ and does not depend on X or Y (hence, we call $\log_2(N)$ a feature-independent upper bound).

To eliminate the problem of unequal normalizing weights, we propose to use the feature-independent upper bound in (21) to normalize the mutual information instead of using (17) as in [10]. Therefore, our normalized feature-feature mutual information is calculated by

$$NI(X;Y) = \frac{I(X;Y)}{\log_2(N)}. \tag{22}$$

Clearly, the normalized feature-feature mutual information is always within the range [0, 1]. Therefore, to achieve a balance between the relevance and the redundancy, we divide the class-feature mutual information by $\log_2|\Omega_C|$. The normalized class-feature mutual information is now defined as

$$NI(C;X) = \frac{I(C;X)}{\log_2(|\Omega_C|)}. \tag{23}$$

Using the normalized mutual information functions defined in (22) and (23), we measure the potential of a feature as

$$f^1(X_i) = NI(C;X_i) - \frac{1}{|S_{i-1}|}\sum_{X_s \in S_{i-1}} NI(X_s;X_i). \tag{24}$$

In order to clarify our improvement, we compare f^1 with the other two measurements in (12) and (16) in terms of the classification accuracy, we denote them as f^2 and f^3, respectively. In addition, to validate the effect of the imbalance between the relevance and the redundancy that we point out above, we combine normalized class-feature mutual information with the same feature-feature mutual information as in [10]. In this way, the goodness of a feature is measured by

$$f^4(X_i) = NI(C;X_i) - \frac{1}{|S_{i-1}|}\sum_{X_s \in S_{i-1}} \frac{I(X_s;X_i)}{\min(H(X_s),H(X_i))}. \tag{25}$$

Furthermore, experiments to compare our method with other common methods such as MIFS [2], MIFS-U [19], GainRatio [16] and SBMLR (an embedded method using Bayesian L1 regularization) [4] are presented in the appendix section to avoid a mess of statistics. The following pseudo-code in Algorithm 2 illustrates the selection process using greedy forward searching strategy.

Algorithm 2: Mutual Information-based Feature Selection Using Greedy Forward Searching

```
Input  : $M$—Total number of features
         $N$—Total number of data samples
         $K$—Number of features to be selected
         $X_{ij}$—Feature values, where $i = 1, 2, \ldots, M$
and $j = 1, 2, \ldots, N$
         $C_j$—Class labels of the data samples, where
$j = 1, 2, \ldots, N$
         $a$—Index of the selected measurement
Output: $S_k$—The selected feature index, where
         $k = 1, 2, \ldots, K$
Forward
    $S = \varnothing$
    //Normalize the features
    for $m = 1$ to $M$ do
        $\mu_m$ = Mean value of $X_m$
        $\sigma$ = Standard deviation of $X_m$
        $X_m = X_m - \mu_m$
        $X_m = X_m/\sigma_m$
    //Convert features into discrete form using linear
    quantization
    $\overline{X} = Quantiz(X)$
    //Start selecting features
    for $k = 1$ to $K$ do
        for $i = 1$ to $M$ do
            Compute $f^a(\overline{X_i})$
        $s = \mathrm{argmax}_{i \notin S}(f^a(\overline{X_i}))$
        $S = S \cup s$
end
```

4 Experiments and discussions

For our experiments, we use 12 datasets from the UCI machine learning repository [1]. Table 1 provides brief information about these datasets. To ensure objective and accurate comparison results and to avoid data-specific statements, we select datasets of different class number, sample number, and feature type, as depicted in Table 1.

Regarding classification methods, we propose to use k-Nearest-Neighbor (kNN, $k = 3$), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). We utilize MatlabArsenal toolbox [30] with WEKA [16] integrated to implement our recognition experiments. The accuracy is measured using the ten-fold cross validation rule. Tables 2, 3, and 4 summarize the classification rates of the three classifiers in the 12 datasets. Each sub-table contains the number of features in the first column and the recognition accuracies in columns 2 to 5, which correspond to the four feature potential measurements f^1, f^2, f^3, and f^4. Besides the average accuracy we also measure the significance of the difference between our method and the others by using paired t-tests [13]. Those t-values are put on the right side of each accuracy. Although the tables are convenient for highlighting insignificant differences, they are limited in representing the overall trend. Therefore, we provide a more general view of the results in Figs. 5,

Table 1 Brief information about the datasets used in the experiments

	Dataset	# Class	# Samples	# Features	Type of features
1	Arrhythmia	16	452	279	Continuous, Discrete
2	Hill Valley	2	1212	100	Continuous
3	Image Segmentation	7	2310	18	Continuous
4	Ionosphere	2	351	33	Continuous, Discrete
5	Isolet	26	7797	617	Continuous
6	Libras Movement	15	360	90	Continuous
7	Madelon	2	2600	500	Continuous
8	Multiple Features	10	2000	649	Continuous, Discrete
9	Landsat Satellite	6	6435	36	Discrete
10	(Connectionist Bench) Sonar	2	208	60	Continuous
11	Spambase	2	4601	57	Continuous, Discrete
12	Breast Cancer (Diagnostic)	2	569	31	Continuous

6, and 7. In the following paragraphs, we analyze the results to show that we can successfully overcome the addressed limitations of the other methods. The accuracy produced by our feature selection algorithm is often higher or at least comparable to those yielded by the other methods.

Arrhythmia dataset: It can be seen that f^1 produces the highest accuracy, which is on average about 26% higher than that of f^2, and the difference in the accuracies increases as the number of features increases. f^4 shows a slightly better result than f^3 (about 3–5% higher, especially when combined with an LDA classifier); it, however, is still worse than f^1, which has the highest results in 15 out of 18 tests with the Arrhythmia dataset.

Hill Valley dataset: This dataset sees almost the same accuracies in all four selection methods. With kNN and LDA classifiers, f^2 and f^1, respectively, produce higher results than do the other methods although the disparity is often not greater than 2%. As can be seen, t-values are rarely higher than 2.26 (or p-value < 0.05). It means that the accuracy differences are not statistically significant.

Image Segmentation dataset: Although the four results approach to each other as the feature number goes up, f^2 and f^3 are often the lowest accurate methods with high t-values (high significant differences). On average, f^4 is slightly better than f^1 (about 1.3% higher accuracies).

Ionosphere dataset: While the four feature selection methods do not create any significant differences when combined with kNN and SVM classifiers (almost all t-values are much smaller than 2.26). f^3 and f^4 have about 3% lower average accuracies than those yielded by f^1 and f^2 in case of using LDA recognition model.

Isolet dataset: With this dataset, f^3 often produces the worst results, this significant weakness is also supported by the very high t-values. f^1 is a little better than f^4 when the number of features is greater than 5. f^1 and f^2 are almost similar with only about 0.6% average distance in the accuracy.

Libras Movement dataset: It is obvious that f^3's accuracies are often significantly lower than those of the others (lower accuracies, high t-values). The differences among f^1, f^2, and f^4 are insignificant since almost all the t-values are much smaller than 2.26.

Madelon dataset: No significant disparity is presented with LDA recognition model; however, when combined with kNN and SVM, f^1 proves to be the superior measurement, with about 5% higher accuracies than those of the others.

Multiple features dataset: f^1 and f^2 are a slightly better than the other two methods if the number of features is less than 10. However, with 10 to 15 features, f^3 and f^4 are better than f^1 and f^2. The four methods approach to similar results when the number of features keeps increasing.

Landsat Satellite and Breast Cancer datasets: Similar results are observed in these datasets regardless of the classifier or the feature selection method. There is no dominant measurement among the four, and the difference of classification rates (between any two selection criterions) is approximately 1–2%.

Sonar dataset: There is no superior among the four methods when the LDA classifier is used. However, while f^1, f^3, and f^4 maintain similar accuracies with kNN and SVM, f^2 loses its competitiveness and obviously becomes the weakest method (about 10% lower recognition rates in almost all the kNN tests).

Spambase dataset: When kNN and LDA classifiers are used, the average accuracies are similar; however, f^3 and f^4 are significantly better than f^1 because they have small standard deviations leading to high t-values as can be seen in Table 2 and 4. Although, f^2's accuracies are clearly lower

Table 2 *k*NN classification accuracies of the 12 datasets. Bold items highlight significant differences in comparison with f^1 (t-value > 2.26 or p-value < 0.05)

k-Nearest-Neighbor (*k*NN, $k = 3$)									
Arrhythmia					Hill Valley				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	51.53	47.99/0.57	55.17/1.66	54.22/1.24	5	53.47	53.40/0.04	52.31/0.54	52.31/0.54
10	**60.00**	58.21/0.85	**55.52/2.35**	57.98/1.41	10	**55.20**	**52.48/2.69**	54.05/0.62	54.05/0.62
15	**60.63**	51.53/1.49	**54.60/2.91**	60.60/0.02	15	53.63	54.29/0.62	54.13/0.33	54.13/0.33
20	**63.99**	**50.29/3.60**	**56.17/3.90**	60.38/2.24	20	53.72	51.82/1.58	54.46/0.39	54.46/0.39
25	**64.19**	**54.24/5.30**	**56.84/4.15**	63.30/0.52	25	53.88	53.39/0.54	53.96/0.05	53.96/0.05
30	**65.24**	**30.58/4.79**	**57.06/4.67**	61.11/1.74	30	53.14	53.80/1.06	53.39/0.16	53.39/0.16
Image Segmentation					Ionosphere				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
3	**92.81**	**83.77/5.38**	**88.35/2.70**	**96.71/2.97**	5	89.18	90.30/0.93	88.32/0.99	88.32/0.99
6	**96.19**	**95.11/2.26**	**93.68/2.41**	96.02/0.58	10	87.13	88.27/1.06	87.18/0.03	87.18/0.03
9	**95.84**	95.41/0.87	95.89/0.40	95.76/1.50	15	87.98	88.03/0.04	85.69/1.81	85.69/1.81
12	**95.84**	**94.55/3.30**	95.58/1.20	95.58/1.20	20	85.99	85.47/0.39	84.84/1.32	84.84/1.32
15	**94.98**	**94.33/2.57**	**96.10/2.65**	**96.10/2.65**	25	83.40	85.16/1.17	84.27/1.17	84.27/1.17
18	95.50	95.50/0.00	95.50/0.00	95.50/0.00	30	84.32	84.89/0.69	84.32/0.00	84.32/0.00
Isolet					Libras Movement				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**48.67**	49.02/1.80	**41.12/9.58**	**51.84/5.96**	5	**63.43**	63.91/0.24	**47.17/7.16**	61.66/0.72
10	**64.22**	**61.40/7.14**	**51.71/17.07**	**61.86/4.39**	10	**69.77**	72.13/1.14	**56.58/4.97**	71.00/0.53
15	**70.55**	70.75/0.54	**59.31/20.52**	**64.92/6.46**	15	69.84	72.41/2.07	64.28/1.95	70.40/0.28
20	**72.98**	72.18/1.42	**64.41/17.74**	71.73/1.77	20	71.85	72.96/0.64	66.52/1.55	72.72/0.47
25	**75.05**	**73.44/4.20**	**65.35/14.76**	**73.22/2.79**	25	73.72	73.54/0.11	**66.51/2.82**	73.25/0.20
30	**76.07**	**74.89/5.21**	**65.77/19.22**	**73.99/4.09**	30	74.10	74.89/0.58	72.16/1.14	75.36/0.77
Madelon					Multiple Features				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**59.92**	**54.96/3.65**	**56.08/2.70**	**56.19/2.56**	5	**89.35**	90.65/1.72	**86.25/3.18**	**87.45/2.71**
10	**58.77**	**54.08/5.27**	**52.38/6.29**	**52.38/6.29**	10	**93.60**	94.60/1.96	94.35/1.39	**96.60/6.80**
15	**58.00**	**52.50/3.34**	**51.19/4.86**	**51.19/4.86**	15	**96.95**	**95.35/2.42**	**98.00/5.55**	97.45/1.50
20	**56.00**	**52.38/5.34**	**53.42/2.93**	**53.42/2.93**	20	**97.95**	**97.05/3.67**	98.05/0.30	98.20/1.34
25	**55.81**	**51.77/3.51**	54.00/2.07	54.00/2.07	25	97.85	97.35/1.79	98.35/2.12	98.20/1.91
30	**55.73**	**51.08/5.94**	**52.42/3.99**	**52.42/3.99**	30	**98.25**	**97.35/2.86**	97.90/1.48	97.95/1.11
Landsat Satellite					Sonar				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	85.76	85.97/0.54	86.03/0.49	85.95/0.37	5	74.04	70.65/1.42	75.51/0.78	75.51/0.78
10	88.70	89.40/1.52	88.28/2.18	88.83/0.35	10	**79.84**	**70.65/2.99**	85.45/1.81	85.45/1.81
15	**89.77**	**90.85/3.77**	89.70/0.29	89.62/1.07	15	**83.18**	**72.68/2.61**	85.13/0.75	85.13/0.75
20	90.19	90.94/2.10	89.93/1.30	90.50/1.57	20	**85.07**	**70.73/6.89**	86.13/0.46	86.13/0.46
25	90.74	90.67/0.25	90.27/2.13	90.86/1.10	25	**87.49**	**76.01/5.75**	87.47/0.02	87.47/0.02
30	90.89	90.85/0.27	90.78/1.17	91.03/1.49	30	**86.09**	**75.10/3.29**	87.04/0.51	87.04/0.51
Spambase					Breast Cancer				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	86.63	80.92/1.66	79.18/1.85	79.18/1.85	5	**94.02**	**91.22/2.58**	95.07/1.96	95.07/1.96
10	**89.91**	**88.29/2.93**	89.13/0.69	89.09/0.74	10	**93.84**	**90.68/2.43**	93.84/0.00	93.84/0.00
15	90.46	89.57/1.94	90.59/0.34	90.57/0.28	15	93.50	94.36/1.15	93.14/0.80	93.14/0.80
20	**89.59**	90.24/1.34	**91.65/3.54**	**91.63/3.57**	20	96.32	95.25/0.85	96.32/0.00	96.32/0.00
25	**90.13**	90.46/0.64	91.04/1.01	**91.57/2.36**	25	97.36	96.48/1.17	97.19/1.00	97.19/1.00
30	**90.26**	90.59/0.80	**91.05/3.63**	**91.05/3.63**	30	96.83	97.01/0.27	96.83/0.00	96.83/0.00

Table 3 SVM classification accuracies of the 12 datasets. Bold items highlight significant differences in comparison with f^1 (t-value > 2.26 or p-value < 0.05)

Support Vector Machine (SVM)									
Arrhythmia					Hill Valley				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**60.52**	46.94/1.57	**56.73/2.72**	57.87/1.35	5	50.49	51.07/0.96	51.24/1.79	50.58/0.19
10	**65.55**	**42.38/2.40**	**58.26/7.35**	**59.17/4.36**	10	51.15	51.32/0.25	51.15/0.00	50.74/1.10
15	**64.65**	**25.93/4.29**	**60.23/2.45**	63.37/0.61	15	51.15	50.49/1.50	51.32/0.52	50.66/1.20
20	**67.27**	**9.78/37.30**	**60.67/5.74**	64.08/1.43	20	**50.74**	50.74/0.00	**51.81/2.90**	50.99/0.61
25	**67.28**	**10.63/45.53**	**60.47/4.45**	64.73/1.07	25	51.15	50.99/0.51	51.15/0.01	51.40/0.58
30	**68.18**	**10.41/44.36**	**60.07/3.88**	66.11/0.71	30	51.32	51.48/0.42	50.82/1.40	50.91/0.96
Image Segmentation					Ionosphere				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
3	**77.97**	75.19/1.54	75.19/1.60	**88.35/5.97**	5	88.27	89.69/1.12	90.02/1.55	90.02/1.55
6	**90.00**	**87.32/9.34**	**84.42/4.41**	**91.77/3.46**	10	**92.28**	92.02/0.39	**90.84/3.00**	**91.14/2.45**
9	**92.64**	**87.58/12.41**	92.38/1.51	92.77/1.41	15	93.16	92.86/0.45	94.60/2.21	94.60/2.21
12	**93.29**	**91.47/4.92**	93.29/0.00	93.33/1.00	20	94.89	94.00/1.04	93.77/1.82	93.77/1.82
15	**93.38**	**92.42/4.14**	**94.24/2.33**	**94.24/2.37**	25	94.88	94.85/0.04	95.19/0.47	95.19/0.47
18	93.90	93.90/0.00	93.85/0.55	93.90/0.00	30	95.16	95.14/0.04	95.73/1.50	95.73/1.50
Isolet					Libras Movement				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**53.24**	52.69/0.68	**46.70/6.39**	**56.28/3.99**	5	**47.54**	51.96/1.63	**40.77/2.34**	**44.77/2.37**
10	**68.90**	**65.56/6.72**	**57.12/17.58**	**66.14/6.48**	10	**61.68**	64.05/0.72	**51.63/4.78**	64.86/1.44
15	**73.66**	**74.70/2.65**	**64.35/20.25**	**70.99/3.69**	15	**72.19**	71.70/0.30	**57.46/5.36**	71.02/0.67
20	**76.40**	76.14/0.57	**70.19/18.47**	77.02/1.19	20	**75.65**	77.04/0.92	**63.49/4.33**	75.39/0.16
25	**78.81**	78.50/0.74	**71.04/17.15**	78.62/0.40	25	**77.30**	80.05/1.80	**72.29/2.38**	77.34/0.05
30	**80.11**	80.12/0.03	**71.99/19.72**	79.83/0.58	30	77.57	81.93/2.14	74.84/1.67	79.33/1.47
Madelon					Multiple Features				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**61.65**	**53.42/8.52**	**53.73/6.80**	**54.12/6.87**	5	**90.10**	91.65/1.85	**86.85/3.33**	**86.60/3.68**
10	**64.00**	**53.81/9.37**	**56.62/6.27**	**56.69/6.90**	10	**93.65**	94.45/1.65	**94.95/5.46**	**96.80/7.47**
15	**62.35**	**53.85/5.69**	**55.31/5.50**	**55.12/5.77**	15	**97.85**	**95.85/4.67**	97.85/0.00	97.65/0.65
20	**61.92**	**53.50/5.74**	**57.73/5.59**	**57.46/6.14**	20	98.20	97.95/0.86	98.45/0.96	98.45/0.83
25	**61.31**	**54.12/8.29**	**57.23/6.87**	**57.31/5.68**	25	**98.65**	**98.10/2.40**	98.55/0.69	98.25/1.56
30	**61.54**	**53.35/8.27**	**55.73/5.55**	**55.96/5.65**	30	**98.75**	98.35/1.50	**98.35/2.45**	98.45/1.41
Landsat Satellite					Sonar				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	85.50	84.82/2.02	85.56/0.17	85.27/1.24	5	73.67	75.94/0.98	72.67/0.47	73.17/0.28
10	87.19	87.72/1.53	87.27/0.36	87.13/0.27	10	76.06	75.94/0.05	76.46/0.14	77.37/0.42
15	**88.94**	88.94/0.01	**88.38/2.93**	88.75/1.10	15	78.44	75.53/1.32	80.30/1.05	80.77/1.22
20	**89.17**	**89.90/3.84**	89.32/0.80	89.36/1.65	20	81.77	75.53/2.20	81.82/0.04	82.75/0.82
25	**89.79**	89.95/1.93	89.98/0.87	**90.01/2.95**	25	**83.18**	**75.51/2.74**	83.75/0.35	84.25/0.51
30	90.35	90.23/0.69	90.33/0.44	90.35/0.02	30	**82.25**	**74.10/3.59**	**85.63/3.30**	**85.63/3.30**
Spambase					Breast Cancer				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	88.22	87.31/2.11	87.87/0.75	87.78/0.93	5	**95.42**	**92.60/2.75**	94.54/1.62	94.54/1.62
10	**90.85**	85.57/1.79	**89.83/2.80**	**89.87/2.64**	10	**95.78**	**92.42/3.36**	95.60/1.00	95.78/0.00
15	90.94	85.78/1.67	90.72/0.69	90.74/0.63	15	95.25	94.36/0.96	94.72/1.96	94.89/1.01
20	91.46	86.81/1.49	91.59/0.63	91.57/0.46	20	97.00	96.29/0.85	97.18/0.42	97.18/0.54
25	92.09	86.22/1.96	91.55/1.65	91.52/1.71	25	97.53	97.36/0.36	97.53/0.00	97.36/1.00
30	92.02	86.33/1.95	91.72/1.14	91.72/1.08	30	97.53	97.88/0.79	97.53/0.00	97.53/0.00

Table 4 LDA classification accuracies of the 12 datasets. Bold items highlight significant differences in comparison with f^1 (t-value > 2.26 or p-value < 0.05)

Linear Discriminant Analysis (LDA)									
Arrhythmia					Hill Valley				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**24.17**	21.92/0.44	**14.41/2.80**	22.75/0.40	5	51.32	50.91/1.62	51.07/1.00	51.07/1.00
10	**33.96**	25.04/1.54	**27.28/3.06**	35.83/0.78	10	51.40	51.15/1.15	51.57/0.61	51.57/0.61
15	**44.74**	**27.04/3.14**	**34.63/4.44**	38.90/1.74	15	51.57	51.07/1.33	51.40/0.40	51.40/0.40
20	**50.89**	**7.00/11.68**	**39.99/4.07**	**41.34/2.53**	20	51.65	51.15/1.97	51.48/0.56	51.48/0.56
25	**55.22**	**10.27/11.47**	**39.41/5.14**	**45.32/2.64**	25	51.57	51.24/1.31	51.65/0.36	51.65/0.36
30	**55.30**	**15.54/8.90**	**41.40/5.13**	49.10/2.16	30	**51.49**	**50.66/2.74**	51.90/1.47	51.90/1.47
Image Segmentation					Ionosphere				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
3	**78.14**	**68.14/5.01**	**73.03/3.13**	**81.65/2.31**	5	**84.29**	84.29/0.01	**80.56/2.51**	**80.56/2.51**
6	**84.59**	**73.16/14.71**	**78.18/4.76**	**87.32/5.81**	10	**84.83**	85.72/0.74	**81.44/2.55**	**81.44/2.55**
9	**87.92**	**81.77/8.70**	87.62/1.00	87.97/0.22	15	**83.67**	85.11/1.89	**81.11/3.23**	**81.11/3.23**
12	**89.87**	**88.23/2.98**	89.78/1.00	89.87/0.00	20	85.68	86.55/1.01	81.70/2.03	81.70/2.03
15	**89.09**	**88.66/2.38**	88.57/0.87	88.57/0.87	25	86.85	87.71/1.96	84.82/1.17	84.82/1.17
18	88.79	88.79/0.00	88.79/0.00	88.79/0.00	30	**85.99**	**87.70/2.71**	84.23/0.91	84.23/0.91
Isolet					Libras Movement				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	**48.31**	47.86/1.00	**41.07/8.55**	49.99/2.10	5	**46.21**	46.81/0.28	**39.18/2.62**	44.62/0.82
10	**61.25**	**57.37/7.10**	**48.04/15.71**	**58.05/5.19**	10	**54.37**	54.94/0.32	**46.06/2.86**	52.62/1.45
15	**63.68**	**65.55/3.76**	**54.14/14.40**	**61.70/2.64**	15	**59.30**	60.98/0.88	**47.77/6.65**	58.72/0.70
20	**65.27**	**66.90/3.70**	**60.83/10.63**	**66.60/2.63**	20	**60.43**	**65.09/2.37**	**52.01/4.85**	62.77/1.74
25	**67.31**	66.27/1.72	**60.81/9.52**	67.82/0.68	25	**64.54**	67.58/1.80	**56.84/4.76**	63.61/0.64
30	**68.21**	68.05/0.27	**62.18/13.03**	69.31/1.84	30	**65.01**	67.55/2.13	**58.33/4.58**	64.81/0.23
Madelon					Multiple Features				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	60.62	60.04/0.56	60.19/0.42	60.19/0.42	5	**88.35**	87.50/1.08	**81.75/8.45**	**81.95/6.59**
10	60.42	59.62/0.65	59.88/0.43	59.88/0.43	10	90.30	90.90/1.05	90.20/0.26	**91.70/2.35**
15	60.69	59.81/0.72	59.58/1.34	59.58/1.34	15	**94.55**	**92.40/2.90**	**95.90/2.61**	94.60/0.18
20	60.54	60.08/0.53	60.08/0.61	60.08/0.61	20	95.30	94.95/0.70	95.80/1.63	95.00/0.97
25	60.92	59.38/2.08	59.54/1.96	59.54/1.96	25	95.75	95.85/0.19	96.10/1.41	95.95/0.45
30	60.15	59.08/1.05	59.46/1.19	59.46/1.19	30	96.15	95.75/0.95	96.35/0.69	96.75/1.86
Landsat Satellite					Sonar				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	81.03	80.22/2.11	81.38/1.36	80.98/0.19	5	69.72	68.27/0.54	71.22/0.59	71.22/0.59
10	82.33	81.94/1.42	82.14/1.09	81.99/1.13	10	70.65	67.77/0.93	72.58/1.29	72.58/1.29
15	82.64	82.39/0.94	82.50/0.59	82.41/1.60	15	75.05	69.32/1.59	75.94/0.51	75.94/0.51
20	82.44	82.38/0.28	82.60/0.83	82.60/1.63	20	77.39	72.54/2.18	78.01/0.28	78.01/0.28
25	82.13	82.24/0.63	82.33/1.17	82.27/0.93	25	76.96	73.47/1.13	78.39/0.82	78.39/0.82
30	82.38	82.38/0.01	82.33/1.00	82.55/1.94	30	78.05	73.51/1.39	77.96/0.05	77.96/0.05
Spambase					Breast Cancer				
# Fea	f^1	f^2/t	f^3/t	f^4/t	# Fea	f^1	f^2/t	f^3/t	f^4/t
5	83.48	83.63/0.30	84.57/2.24	84.57/2.24	5	**94.03**	**91.04/3.79**	94.20/0.54	94.20/0.54
10	**86.57**	85.89/1.05	**87.92/4.35**	**87.92/4.35**	10	**94.56**	**91.56/3.43**	94.56/0.00	94.56/0.00
15	87.00	87.48/0.69	87.42/1.19	87.42/1.19	15	94.38	94.20/0.42	94.02/1.50	94.02/1.50
20	**87.42**	87.57/0.28	**88.37/3.00**	**88.37/3.00**	20	95.96	95.25/1.21	95.96/0.00	95.96/0.00
25	**87.94**	88.07/0.25	**89.13/4.50**	**89.13/4.50**	25	**96.49**	**95.60/3.00**	96.67/1.00	96.67/1.00
30	**88.70**	88.59/0.27	**89.81/4.02**	**89.81/4.02**	30	96.67	96.14/1.41	96.67/0.00	96.67/0.00

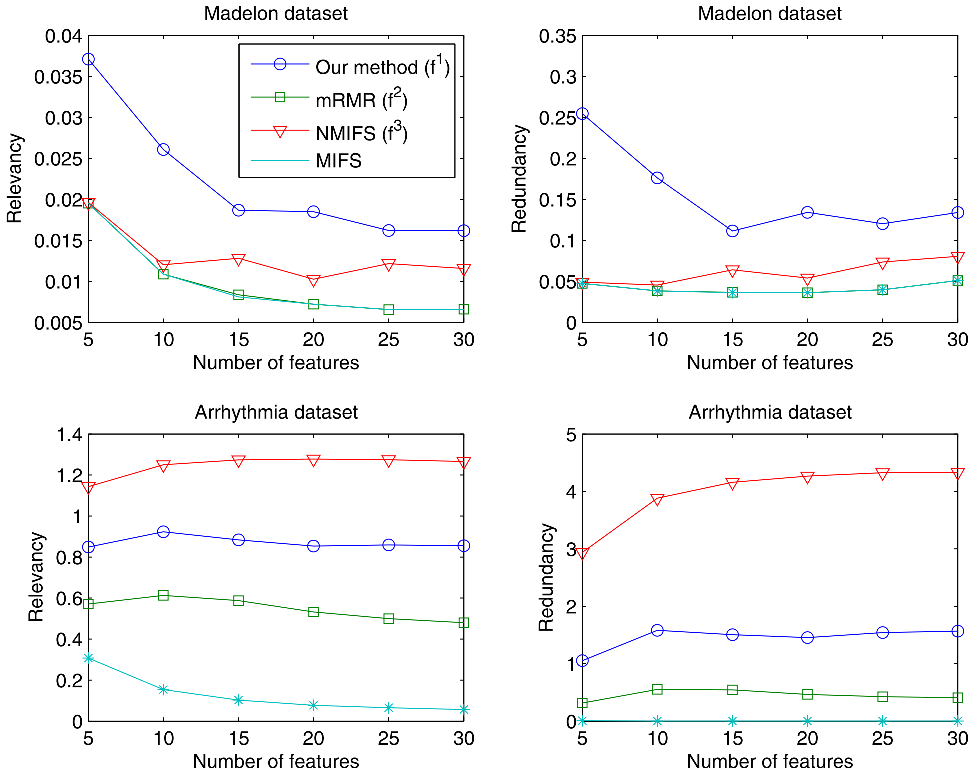

Fig. 4 Redundancy and relevancy of the selected features

than those of f^1 when they are used with SVM, the differences are not statistically significant because of the low t-values.

Overall, we can see that f^2 is often the worst criterion; f^1, in contrast, is often one of the two best measurements. Even if it does not have the highest result (for example with the Sonar dataset), the difference between f^1 and the best method is not significant. It is also worth noting that f^4 often produces better results than does f^3. Furthermore, from Tables 2, 3, and 4, we summarize the number of times that each feature selection method produces the highest results and show the statistics in Fig. 8. It is obvious that f^1 proves to be the most outstanding method. Among the other three selection criterions, f^4, in general, is a little better than f^2 and f^3. Hence, the statistics provide another reason for us to conclude that f^1 is the most superior method with f^4 occupying the second position, f^2 and f^3 competing for the lowest rank. Since f^4 differs from f^3 only in the class-feature normalization, it is clear that the normalization of class-feature mutual information has a positive effect on the quality of the selected feature set. The superiority of f^1 illustrates the efficiency of the constant normalizing weights in our methods because f^1 and f^4 are different only in these weights.

In addition, Fig. 4 shows an analysis of the redundancy (RD) and relevancy (RL) of the selected features. Those two quantities are computed as below (derived from the method in [34]).

$$RD(X_1, X_2, \dots, X_N) = \frac{1}{N(N-1)} \sum_{i \neq j} I(X_i; X_j) \qquad (26)$$

$$RL(X_1, X_2, \dots, X_N) = \frac{1}{N} \sum_{i} I(C; X_i) \qquad (27)$$

As can be seen, MIFS is biased toward the less redundant features. As a result, the classification accuracy is low because of less relevant features. It is also clear that our method gives higher priority to selecting the relevant features in case of low-redundancy dataset (Madelon). Whereas, it pays more attention to selecting the less redundant features if the dataset has high redundancy (Arrhythmia). In other words, our method is less prone to a specific kind of feature than the others.

So far, we have proven that our method not only inherits the advantages of the parameter-free methods like mRMR and NMIFS but also overcomes their limitations. By proposing to normalize the class-feature mutual information, we are able to avoid the imbalance between the relevance and the redundancy, which can be seen in mRMR and NMIFS, as we pointed out in Sect. 2. To resolve the problem of unequal normalizing weights, we present a feature-independent up-

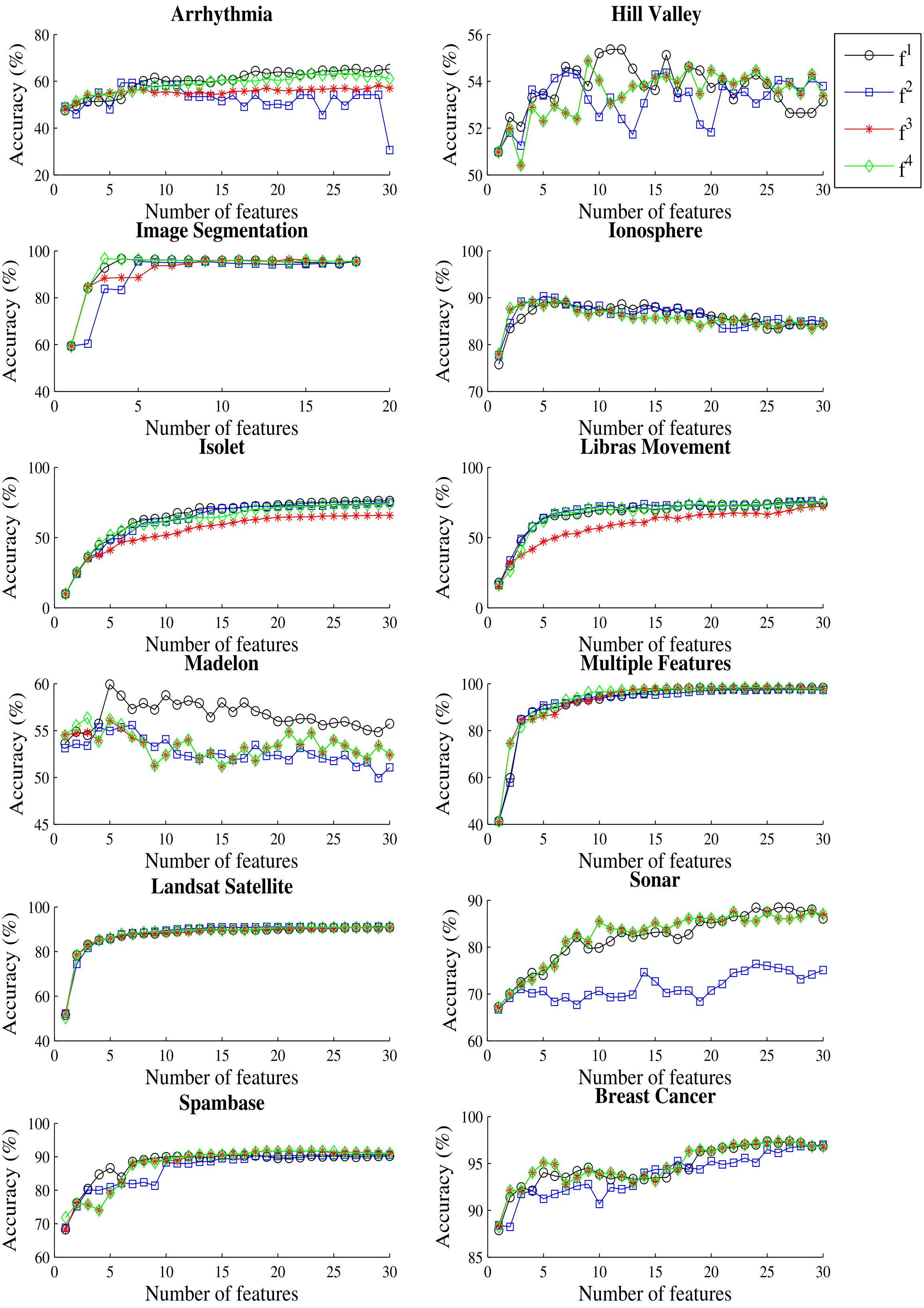

Fig. 5 *k*NN classification accuracies of the 12 datasets

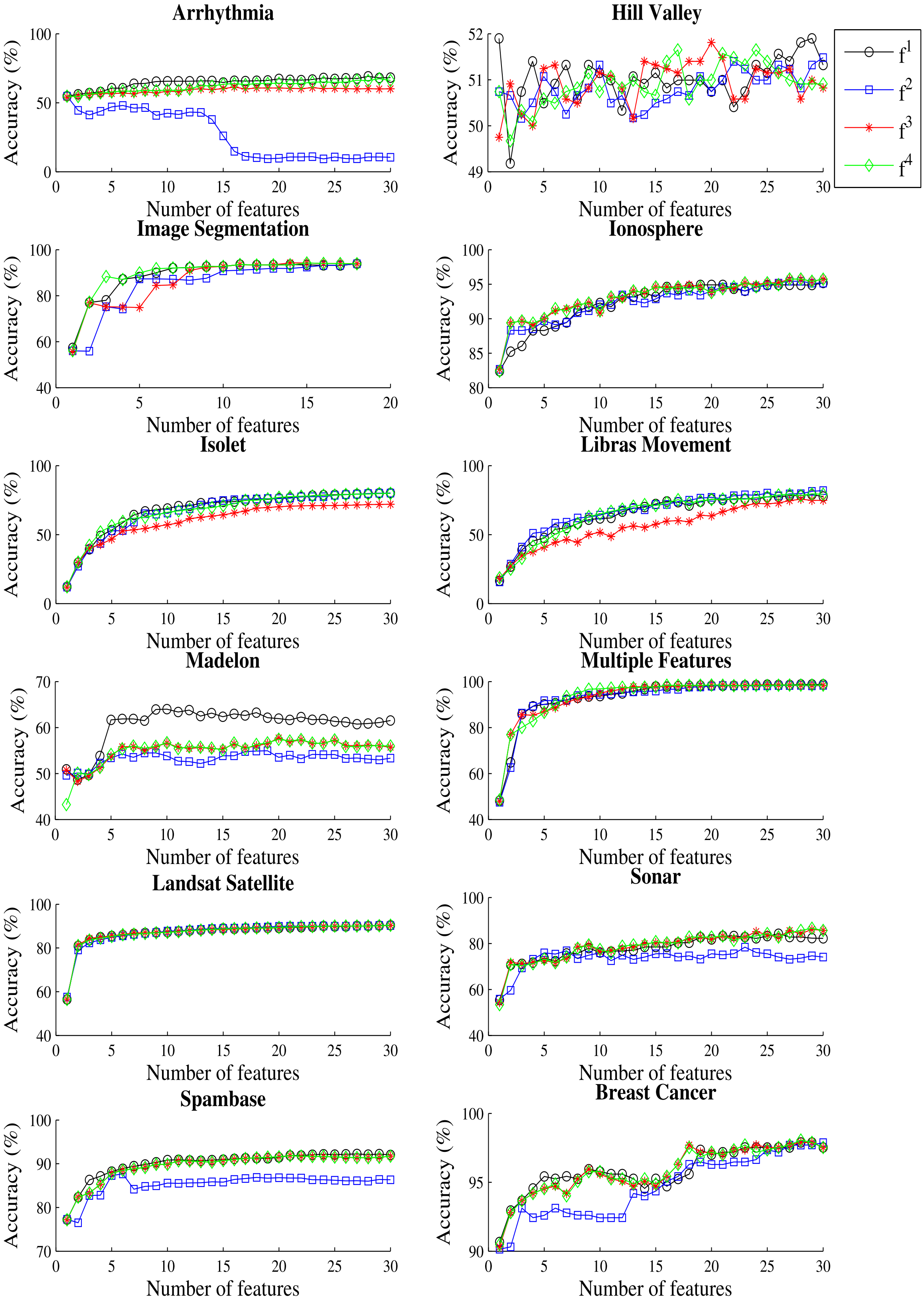

Fig. 6 SVM classification accuracies of the 12 datasets

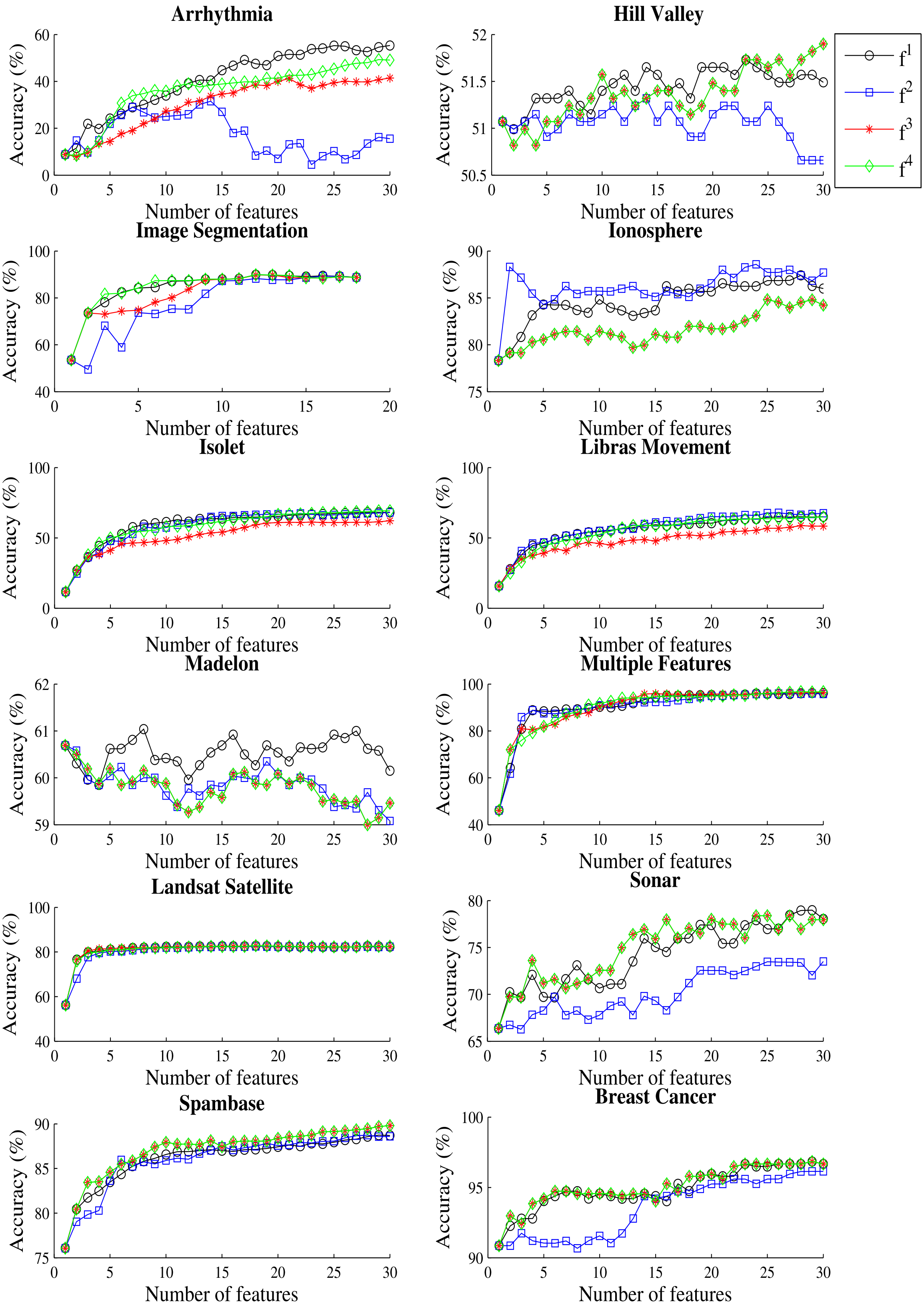

Fig. 7 LDA classification accuracies of the 12 datasets

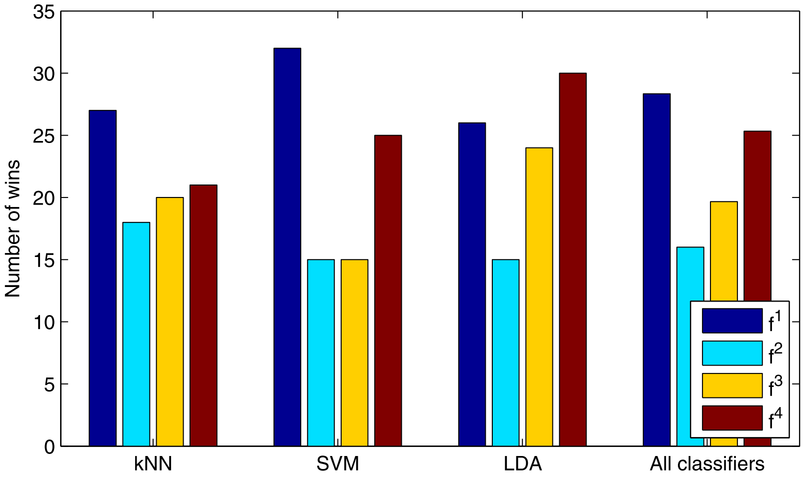

Fig. 8 Number of times each method achieves the highest accuracy. The *rightmost group* shows the average number of all the three classifiers. There are $12 \times 6 = 72$ tests for each classifier and a total of $72 \times 3 = 216$ tests for all three classifiers. The results of those tests are presented in Tables 2, 3, and 4

per bound of the mutual information, which then acts as the normalizing factor.

5 Conclusion

In conclusion, we have reviewed some recently developed algorithms for mutual information-based feature selection. We discussed the limitations of each method, and based on our observations, we proposed our own method derived from the NMIFS with two improvements, the normalization of the mutual information and the feature-independent normalizing weights. To clarify these improvements, we conducted comprehensive experiments using 12 datasets of different characteristics from the UCI machine learning repository. The experimental results confirmed our analysis and provided obvious evidences, allowing us to conclude that our method achieves a better feature set in terms of classification accuracy.

In the present paper, we limited our scope to the selection criterion only, since this is an important basis on which to develop different searching algorithms. For our future work, we are going to integrate our selection criterion into more advanced searching strategies such as branch and bound [21], genetic search [3, 20, 22]. Another problem that should be considered is the quantization process. Currently, a very simple linear quantization method is utilized; we expect that a better quantization algorithm may bring higher quality to the feature selection method.

Acknowledgements This work was supported by the Industrial Strategic Technology Development Program (10035348, Development of a Cognitive Planning and Learning Model for Mobile Platforms) funded by the Ministry of Knowledge Economy (MKE, Korea), and was partially supported by Microsoft Research Asia. We also would like to acknowledge Dr. Zheng Zhao and his colleges for their collection of feature selection algorithms.

Appendix A: Accuracies of kNN using our method (f^1), MIFS [2], MIFSU [19], and Gain Ratio (GR) [16]

Arrhythmia				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	50.43	52.01/0.72	47.46/0.41	50.71/0.12
10	**60.98**	**51.58/5.11**	**44.73/2.54**	**55.64/2.76**
15	**63.36**	**47.06/4.21**	**50.06/3.37**	**53.72/4.37**
20	**62.32**	**41.78/3.18**	**45.21/2.50**	**54.90/5.00**
25	**64.05**	**30.55/3.77**	**49.28/3.26**	54.39/1.64
30	**65.47**	**54.28/8.49**	**41.00/3.49**	**59.12/2.51**

Hill Valley				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	51.98	51.73/0.30	52.31/0.21	52.39/0.26
10	**52.72**	53.05/0.43	52.63/0.10	**50.33/2.35**
15	52.47	52.56/0.08	53.79/1.42	52.63/0.15
20	53.21	52.97/0.23	53.30/0.11	52.72/0.43
25	53.29	52.88/0.54	52.80/0.98	52.30/1.12
30	52.88	52.30/0.59	53.30/0.65	52.31/0.32

Image Segmentation				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
3	**92.64**	**62.34/21.99**	**88.61/2.74**	**88.14/2.86**
6	**96.10**	**84.76/13.01**	94.33/1.60	**88.79/11.67**
9	**95.97**	**92.55/3.86**	**92.55/3.86**	93.81/2.07
12	**95.76**	**93.51/4.62**	**93.55/3.39**	95.58/0.94
15	**94.33**	95.02/1.40	**93.64/2.84**	94.72/1.40
18	**95.06**	95.06/0.00	95.06/0.00	95.06/0.00

Ionosphere				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	87.43	87.44/0.01	87.15/0.22	84.26/1.78
10	85.42	85.43/0.01	85.98/0.98	87.13/1.76
15	82.17	84.56/0.69	84.86/0.84	85.69/1.11
20	79.62	83.12/1.09	83.70/1.37	83.44/1.15
25	80.86	82.27/1.04	81.99/0.71	82.85/1.76
30	81.14	81.97/0.69	81.97/0.79	81.40/0.28

Isolet				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**49.16**	**36.89/13.61**	49.15/0.03	**12.08/73.07**
10	**64.67**	**43.81/14.35**	**52.20/15.81**	**19.99/44.03**
15	**70.72**	**45.53/38.02**	**57.52/21.87**	**28.56/37.62**
20	**72.72**	**47.98/60.18**	**57.88/26.20**	**32.99/43.25**
25	**74.87**	**49.71/38.95**	**57.33/23.66**	**34.73/25.93**
30	**75.70**	**51.44/51.21**	**57.80/27.50**	**40.08/16.89**

Libras Movement				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**62.97**	58.27/2.01	**58.10/2.84**	**46.06/4.84**
10	**69.43**	69.73/0.12	71.75/1.20	**50.79/4.98**
15	**69.75**	**73.52/2.31**	73.46/1.94	**53.13/6.89**
20	**70.91**	73.01/1.08	74.04/2.00	**56.20/6.83**
25	**72.84**	74.70/1.09	76.65/2.15	**66.15/2.31**
30	**73.43**	76.09/1.73	**77.21/2.34**	**68.68/3.04**

Madelon				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**61.00**	**53.35/10.69**	**72.12/4.47**	59.19/1.21
10	**58.54**	**52.00/4.94**	**62.19/2.78**	**73.65/9.52**
15	**58.42**	**51.15/8.28**	57.81/0.55	**80.42/12.86**
20	**56.69**	**51.96/3.16**	56.81/0.14	**76.12/20.03**
25	**56.88**	**51.50/4.38**	55.12/2.25	**71.46/13.26**
30	**56.46**	**51.00/4.85**	**53.42/2.79**	**67.58/7.83**

Multiple Features				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**89.25**	87.75/1.13	87.10/2.14	**77.85/4.15**
10	**93.95**	**89.65/5.02**	92.05/1.91	**89.10/5.99**
15	**97.25**	**92.40/7.83**	**94.40/5.90**	**93.40/7.61**
20	**97.85**	**94.45/5.35**	**95.20/3.57**	**93.65/8.78**
25	**98.15**	**95.25/6.50**	**95.85/4.64**	**95.30/5.57**
30	**98.35**	**95.65/4.10**	**95.70/3.40**	**95.60/6.40**

Landsat Satellite				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	86.31	86.11/0.48	86.22/0.22	82.45/1.70
10	**88.95**	89.06/0.33	**89.76/3.12**	**87.93/3.58**
15	90.33	90.12/0.57	90.40/0.14	89.39/2.06
20	90.38	90.47/0.24	90.74/1.06	90.19/0.71
25	90.99	90.63/0.94	90.69/1.10	90.82/0.87
30	90.77	90.88/0.31	90.97/0.65	90.72/0.40

Sonar				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**73.88**	74.37/0.23	72.97/0.41	**70.04**/3.22
10	**80.78**	**72.15/2.30**	**67.68/3.43**	74.49/1.84
15	**80.23**	**71.14/2.54**	79.23/0.29	76.90/0.95
20	**85.11**	**74.97/5.14**	**75.33/3.21**	81.18/1.69
25	**85.57**	**73.02/4.03**	**75.87/3.94**	85.09/0.16
30	**88.47**	**77.90/4.27**	**76.92/3.96**	84.57/1.44

Spambase				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**86.74**	**73.59/3.10**	83.81/0.94	86.20/0.54
10	**89.20**	**78.13/3.82**	88.46/0.62	88.44/0.63
15	**89.66**	**78.37/3.90**	88.65/1.00	89.33/0.37
20	**89.57**	**81.53/3.68**	89.07/0.38	88.57/0.68
25	**89.55**	86.15/2.20	88.68/0.68	89.18/0.29
30	**89.63**	87.39/1.97	89.59/0.06	90.16/1.31

Breast Cancer				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**93.83**	**90.85/2.69**	92.09/1.32	94.38/0.45
10	94.38	91.36/1.60	92.61/1.41	94.56/0.58
15	**93.31**	92.96/0.43	**95.07/2.72**	94.02/1.50
20	95.77	95.08/1.49	95.07/1.06	95.42/0.48
25	**97.00**	96.48/0.66	95.61/2.05	**95.76/3.29**
30	97.18	97.36/0.31	97.18/0.00	97.01/1.00

Appendix B: Accuracies of SVM using our method (f^1), MIFS [2], MIFSU [19], and Gain Ratio (GR) [16]

Arrhythmia

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**58.20**	**39.66/2.45**	**33.19/2.92**	58.64/0.34
10	**63.53**	**39.22/2.95**	**8.86/44.01**	60.20/2.23
15	**63.73**	**39.23/3.11**	**10.18/56.01**	60.20/2.13
20	**65.05**	**34.11/3.92**	**10.41/40.70**	60.87/2.20
25	**66.58**	**14.42/8.18**	**7.78/24.51**	50.10/1.94
30	**67.92**	**18.15/8.36**	**8.65/36.23**	**43.59/2.70**

Hill Valley

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	50.00	50.00/0.01	51.65/1.21	50.17/0.13
10	50.33	50.25/0.23	51.57/1.65	50.49/0.32
15	50.66	50.41/0.51	50.66/0.01	49.92/1.07
20	51.07	50.99/0.18	51.15/0.20	50.09/1.33
25	51.07	50.58/0.94	50.99/0.28	49.59/1.63
30	50.58	51.32/1.44	50.58/0.01	51.15/1.35

Image Segmentation

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
3	**78.79**	**53.16/11.03**	79.09/0.17	75.54/1.67
6	**89.70**	**75.84/14.04**	**86.93/5.58**	**74.81/22.40**
9	**92.60**	**86.67/6.79**	**87.53/8.92**	**83.72/6.31**
12	**93.16**	93.29/0.44	**92.21/4.30**	93.29/0.57
15	**93.38**	93.98/2.09	93.16/1.10	**93.64/2.70**
18	**94.11**	94.20/1.01	94.11/0.02	**94.29/2.45**

Ionosphere

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	83.92	84.48/0.98	85.61/1.74	84.45/0.37
10	87.59	87.31/0.32	88.43/1.15	87.86/0.30
15	89.29	88.72/0.56	89.01/0.42	88.14/0.95
20	**90.71**	89.30/2.22	**88.45/2.44**	89.86/1.96
25	90.72	89.88/1.96	89.87/1.96	90.13/1.03
30	89.58	90.16/1.50	90.15/0.99	90.43/1.14

Isolet

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**53.26**	**42.12/9.66**	52.66/1.42	**13.15/42.09**
10	**69.14**	**49.56/14.06**	**58.15/16.21**	**23.74/29.65**
15	**74.00**	**53.29/22.51**	**64.13/21.82**	**32.86/36.84**
20	**76.77**	**57.46/24.99**	**66.17/18.52**	**37.84/44.69**
25	**78.85**	**59.07/27.46**	**66.97/23.07**	**38.98/44.86**
30	**80.15**	**61.63/40.87**	**68.05/25.39**	**43.85/20.25**

Libras Movement

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**53.31**	52.85/0.13	54.27/0.30	**33.59/4.39**
10	**63.64**	69.36/2.10	66.00/1.14	**40.16/9.26**
15	**67.99**	**75.55/3.18**	72.77/1.58	**46.58/8.31**
20	**71.42**	**76.45/2.77**	**76.62/2.47**	**54.23/6.21**
25	**79.22**	80.60/0.77	81.38/1.17	**63.90/5.79**
30	**80.25**	82.77/2.15	83.42/2.18	**67.03/6.04**

Madelon

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**61.54**	**52.77/12.21**	**66.69/5.56**	**57.54/4.38**
10	**64.19**	**53.12/21.65**	65.19/1.21	**71.15/8.75**
15	**63.46**	**52.12/13.40**	62.58/1.74	**77.15/10.66**
20	**62.42**	**53.31/10.82**	60.85/1.96	**77.77/16.59**
25	**61.96**	**53.54/10.11**	**60.31/2.65**	**74.58/14.84**
30	**60.23**	**53.19/5.81**	**59.08/3.26**	**72.65/14.94**

Multiple Features

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**90.00**	**87.85/3.17**	**85.95/3.28**	**77.15/4.72**
10	**93.95**	92.65/1.57	94.85/1.43	**89.75/5.23**
15	**97.80**	**94.55/5.57**	**96.50/2.62**	**94.05/5.76**
20	**98.35**	**96.35/4.82**	**97.30/2.40**	**95.60/5.55**
25	**98.65**	**96.85/3.55**	**97.40/2.95**	**96.30/5.48**
30	**98.70**	**97.35/3.95**	**97.50/2.98**	**96.95/3.80**

Landsat Satellite

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**85.05**	84.83/0.57	85.27/0.73	**83.95/2.84**
10	**87.21**	87.91/1.96	**87.66/2.58**	86.85/1.55
15	88.80	88.79/0.01	88.97/0.70	88.41/1.88
20	89.34	89.63/1.23	89.73/1.64	89.31/0.14
25	89.81	90.27/1.90	90.15/1.68	89.99/1.60
30	90.36	90.43/0.30	90.46/0.38	90.36/0.01

Sonar

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	69.56	68.13/0.51	69.52/0.01	62.48/1.22
10	75.87	74.13/0.40	70.68/1.61	69.25/2.24
15	79.69	74.63/1.19	72.70/2.13	75.40/2.05
20	74.94	74.05/0.32	73.60/0.56	77.37/1.63
25	77.89	76.46/0.76	74.03/1.57	78.89/0.62
30	**79.80**	**73.06/3.08**	74.01/1.95	79.85/0.02

Spambase

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**87.98**	**74.48/6.34**	87.65/0.91	**86.37/2.32**
10	**90.65**	**71.75/7.02**	**88.42/4.32**	**89.52/3.69**
15	**91.33**	**77.09/5.31**	**86.24/2.28**	90.91/1.33
20	**91.70**	**78.05/5.41**	86.44/1.83	89.35/1.29
25	**92.44**	**84.16/3.12**	86.29/2.17	89.70/1.71
30	**92.59**	**85.18/2.69**	86.81/2.22	90.31/1.39

Breast Cancer

# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**94.56**	**92.11/2.32**	93.15/2.24	94.38/0.37
10	94.73	92.47/2.05	93.16/1.40	94.20/1.01
15	94.56	95.08/0.51	94.74/0.17	94.91/1.50
20	97.01	96.13/1.24	96.67/0.42	97.37/0.62
25	97.37	97.02/1.00	97.19/0.44	97.20/0.55
30	97.55	97.37/1.00	97.37/1.00	97.37/1.00

Appendix C: Accuracies of LDA using our method (f^1), MIFS [2], MIFSU [19], and Gain Ratio (GR) [16]

Arrhythmia					Hill Valley				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**21.24**	21.92/0.09	**13.22/2.35**	16.16/1.36	5	50.90	50.99/0.29	51.40/1.96	51.07/0.68
10	32.34	22.58/1.27	43.27/1.74	25.41/1.68	10	50.99	50.99/0.00	51.40/1.86	51.15/0.69
15	**42.97**	**21.92/2.55**	39.92/0.50	33.16/1.92	15	51.40	51.23/0.52	51.56/0.80	51.23/0.81
20	**52.33**	**27.03/3.22**	**31.19/2.60**	**33.90/3.85**	20	51.15	50.99/0.61	51.15/0.01	51.07/0.23
25	**54.19**	43.63/1.57	**28.61/3.42**	**35.82/2.72**	25	51.40	51.15/0.63	51.23/0.80	50.74/2.07
30	**58.07**	46.68/1.90	**30.79/4.03**	**26.23/4.42**	30	51.57	51.23/0.72	51.48/0.37	52.80/1.74
Image Segmentation					Ionosphere				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
3	**76.71**	**42.25/22.53**	**70.48/3.55**	73.68/2.00	5	84.30	83.45/0.99	84.03/0.27	82.31/1.76
6	**84.33**	**60.91/22.98**	**72.16/6.78**	**75.02/22.47**	10	84.01	84.85/1.39	84.85/1.15	83.43/0.63
9	**87.53**	**72.03/12.73**	**71.47/22.13**	**82.51/2.91**	15	84.01	84.58/0.61	85.15/1.49	83.70/0.31
12	**89.78**	89.57/0.51	**87.23/6.74**	89.65/1.00	20	84.30	84.56/0.27	84.57/0.28	83.99/0.46
15	89.26	89.18/0.21	89.57/1.65	89.48/0.99	25	83.97	85.44/1.21	84.88/0.72	83.68/0.29
18	89.05	89.05/0.00	89.05/0.00	89.05/0.00	30	83.40	85.13/1.53	84.84/1.19	83.68/0.36
Isolet					Libras Movement				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**48.15**	**35.42/9.90**	47.89/0.77	**10.57/42.96**	5	**48.17**	46.14/0.98	46.47/1.85	**37.23/3.53**
10	**61.52**	**41.12/18.83**	**48.72/29.10**	**19.20/31.87**	10	**51.57**	**58.99/2.39**	**60.36/4.65**	**45.65/2.28**
15	**64.01**	**43.91/20.84**	**53.85/22.05**	**27.57/33.54**	15	**57.68**	62.23/1.35	63.48/2.24	**48.15/3.85**
20	**65.44**	**47.24/18.25**	**56.38/14.76**	**32.26/36.08**	20	**60.83**	63.08/0.64	64.50/1.28	**49.13/3.33**
25	**67.27**	**49.19/19.32**	**57.39/10.84**	**33.22/32.61**	25	**63.83**	63.28/0.19	65.02/0.50	**50.90/4.69**
30	**68.35**	**51.62/23.51**	**58.63/14.80**	**36.32/19.85**	30	**66.31**	63.46/1.08	66.23/0.04	**53.26/4.64**
Madelon					Multiple Features				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	60.62	61.04/2.08	60.38/0.60	60.08/0.47	5	**88.65**	**82.90/3.83**	**78.95/10.58**	**74.55/6.69**
10	60.62	60.65/0.08	59.96/1.95	61.08/1.02	10	**89.60**	88.25/1.09	88.85/0.63	**84.90/3.84**
15	60.27	60.81/1.06	60.04/0.52	61.38/1.62	15	**94.60**	**90.35/3.49**	**90.75/5.27**	**88.20/10.06**
20	60.00	60.12/0.31	59.92/0.43	60.38/0.47	20	**95.30**	**92.35/2.93**	**92.50/3.32**	**90.25/9.27**
25	60.15	60.15/0.00	59.77/1.23	60.69/0.84	25	**95.80**	**93.60/2.63**	**93.65/3.17**	**91.75/8.06**
30	60.08	59.62/0.82	59.50/1.53	60.19/0.13	30	**96.50**	**94.70/2.59**	**94.00/3.90**	**92.75/8.36**
Landsat Satellite					Sonar				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**80.84**	79.92/2.01	81.21/1.19	**78.35/4.45**	5	72.38	72.48/0.03	70.02/1.32	69.05/0.98
10	**82.36**	**81.09/3.99**	**81.32/2.77**	**80.89/3.05**	10	**74.93**	71.43/1.07	**69.10/2.46**	70.50/1.21
15	**82.77**	82.22/1.96	**82.07/2.40**	82.35/1.60	15	**78.21**	**74.38/2.46**	74.88/2.09	**71.98/2.76**
20	82.38	82.28/0.23	82.24/0.46	82.39/0.05	20	**78.24**	**70.55/6.04**	74.83/1.29	74.90/1.33
25	82.25	82.27/0.04	82.19/0.17	81.99/2.01	25	75.40	70.90/1.12	74.33/0.47	74.43/0.63
30	82.45	82.10/1.21	82.03/1.67	82.58/1.93	30	77.81	74.36/1.23	74.83/1.09	77.33/0.32
Spambase					Breast Cancer				
# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t	# Fea	f^1	$MIFS/t$	$MIFSU/t$	GR/t
5	**83.16**	**77.07/2.84**	**85.72/5.10**	83.18/0.03	5	**93.85**	**91.21/3.50**	**92.45/2.45**	93.67/0.43
10	**86.74**	**78.25/4.92**	86.42/0.90	**85.70/2.57**	10	**94.38**	92.97/1.35	**92.80/2.38**	94.38/0.00
15	**87.18**	**80.92/3.15**	86.11/0.82	86.55/1.76	15	94.91	95.08/0.44	94.72/0.33	94.91/0.00
20	**87.03**	**82.33/2.46**	87.83/0.65	87.07/0.15	20	96.14	95.26/1.86	95.08/1.77	95.96/0.30
25	**88.18**	**84.42/3.49**	87.57/0.48	86.70/1.89	25	96.66	96.14/0.65	95.61/1.60	96.66/0.00
30	**88.79**	**85.61/3.35**	87.96/0.83	87.74/1.69	30	96.84	96.84/0.02	96.49/0.79	96.66/0.55

Appendix D: Accuracies of our method (f^1), and SBMLR [4]

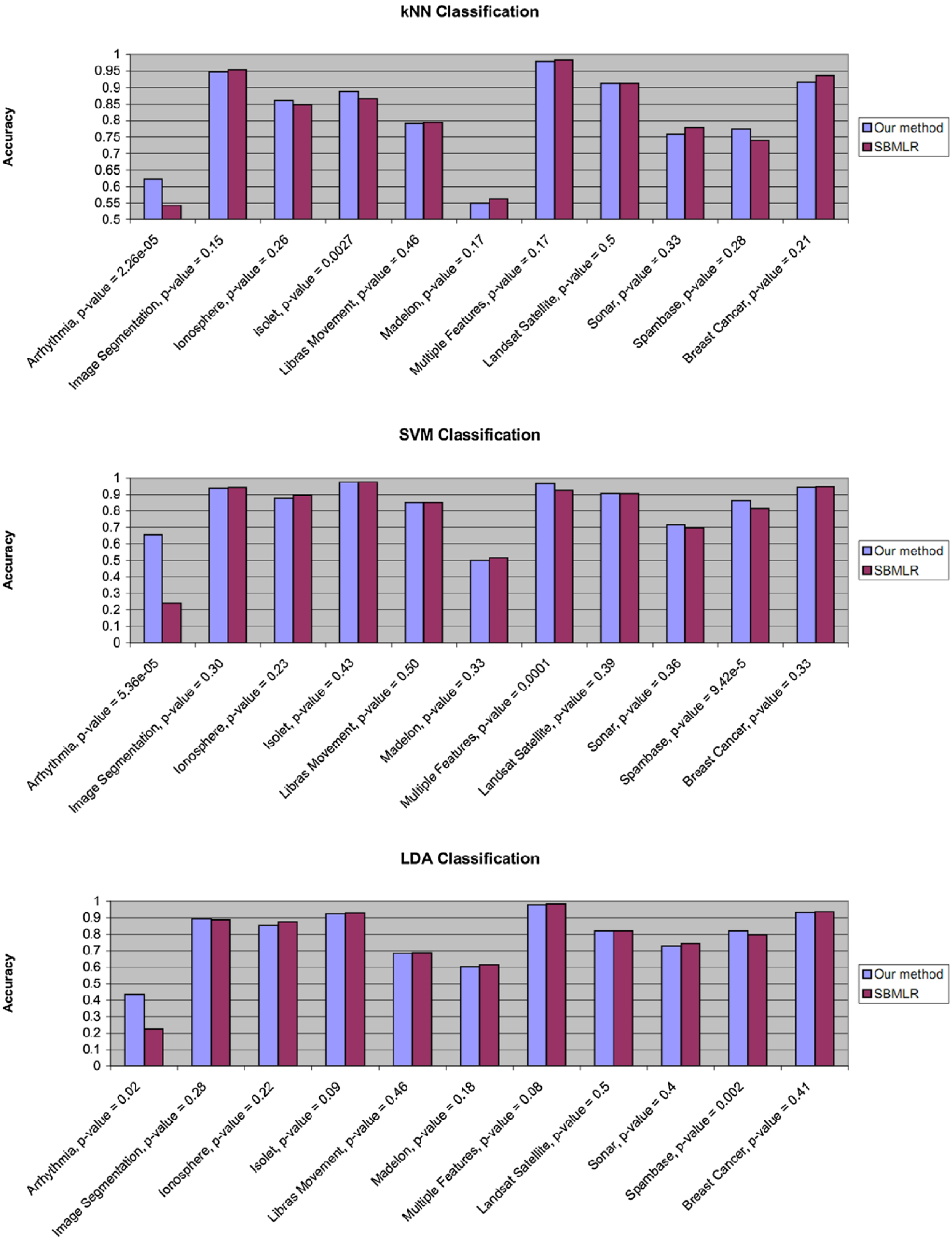

Fig. 9 Classification accuracies of our method and SBMLR. In this experiment, we first used SBMLR to select a subset of features, then our method was executed to select the same number of features. As can be seen, our accuracies are higher than those of SBMLR in all the cases in which significant differences are observed ($p\text{-}value < 0.05$)

References

1. Asuncion A, Newman DJ (2007) Uci machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://www.ics.uci.edu/~mlearn/MLRepository.html
2. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
3. Bhanu B, Lin Y (2003) Genetic algorithm based feature selection for target detection in sar images. Image Vis Comput 1(7):591–608
4. Cawley GC, Talbot NLC, Girolami M (2007) Sparse multinomial logistic regression via Bayesian l1 regularisation. Adv Neural Inf Process Syst 19:209–216
5. Chang T-W, Huang Y-P, Sandnes FE (2009) Efficient entropy-based features selection for image retrieval. In: Proceedings of the 2009 IEEE international conference on systems, man and cybernetics, pp 2941–2946
6. Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney MW (2007) Feature selection methods for text classification. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pp 230–239
7. Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156
8. Dimililer N, Varoglu E, Altinçay H (2009) Classifier subset selection for biomedical named entity recognition. Appl Intell 31:267–282
9. Dy JG, Brodley CE, Kak A, Broderick LS, Aisen AM (2003) Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Trans Pattern Anal Mach Intell 25(3):373–378
10. Estévez PA, Tesmer M, Perez CA, Zurada JM (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20(2):189–201
11. Fodor IK (2002) A survey of dimension reduction techniques. Technical report, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory
12. Forman G, Alto P (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
13. Goulden CH (1956) Methods of statistical analysis, 2nd edn. Wiley, New York
14. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
15. Hall MA (1999) Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: An update. SIGKDD Explor 11(1):10–18
17. Kamimura R (2011) Structural enhanced information and its application to improved visualization of self-organizing maps. Appl Intell 34:102–115
18. Khor K-C, Ting C-Y, Amnuaisuk S-P (2009) A feature selection approach for network intrusion detection. In: Proceedings of the 2009 international conference on information management and engineering, pp 133–137
19. Kwak N, Choi C-H (2002) Input feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159
20. Li Y, Zeng X (2010) Sequential multi-criteria feature selection algorithm based on agent genetic algorithm. Appl Intell 33:117–131
21. Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. IEEE Trans Comput 26(9):917–922
22. Oh I-S, Lee J-S, Moon B-R (2004) Hybrid genetic algorithms for feature selection. IEEE Trans Pattern Anal Mach Intell 26(11):1424–1437
23. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
24. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):1367–4803
25. Shen K-Q, Ong C-J, Li X-P (2008) Novel multi-class feature selection methods using sensitivity analysis of posterior probabilities. In: Proceedings of the IEEE international conference on systems, man and cybernetics, pp 1116–1121
26. Shie J-D, Chen S-M (2008) Feature subset selection based on fuzzy entropy measures for handling classification problems. Appl Intell 28:69–82
27. Tsang C-H, Kwong S, Wang H (2007) Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognit 40(9):2373–2391
28. Vinh LT, Thang ND, Lee Y-K (2010) An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. In: Proceedings of the 10th IEEE/IPSJ international symposium on applications and the Internet, pp 395–398
29. Xia H, Hu BQ (2006) Feature selection using fuzzy support vector machines. Fuzzy Optim Decis Mak 5(2):187–192
30. Yan R (2006) MatlabArsenal toolbox for classification algorithms. Informedia School of Computer Science, Carnegie Mellon University
31. Yang HH, Moody J (1999) Data visualization and feature selection: New algorithms for nongaussian data. In: Advances in neural information processing systems. MIT Press, Cambridge, pp 687–693
32. Yu L, Liu H (2004) Redundancy based feature selection for microarray data. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, pp 737–742
33. Yuan G-X, Chang K-W, Hsieh C-J, Lin C-J (2010) A comparison of optimization methods and software for large-scale l1-regularized linear classification. J Mach Learn Res 11:3183–234
34. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H (2010) Advancing feature selection research—asu feature selection repository. Technical report, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University

La The Vinh received his B.S. and M.S. from Hanoi University of Technology, Vietnam, in 2004 and 2007, respectively. Since September 2008, he has been working on his PhD degree at the Department of Computer Engineering at Kyung Hee University, Korea. His research interests include digital signal processing, pattern recognition and artificial intelligence.

Sungyoung Lee received his B.S. from Korea University, Seoul, Korea. He got his M.S. and PhD degrees in Computer Science from Illinois Institute of Technology (IIT), Chicago, Illinois, USA in 1987 and 1991 respectively. He has been a professor in the Department of Computer Engineering, Kyung Hee University, Korea since 1993. He is a founding director of the Ubiquitous Computing Laboratory, and has been affiliated with a director of Neo Medical ubiquitous-Life Care Information Technology Research Center, Kyung Hee University since 2006. He is a member of ACM and IEEE.

Young-Tack Park received his B.S. from Seoul National University, Korea in 1978. He got M.S. and PhD degrees in Computer Science from KAIST Korea and University of Illinois at Urbana-Champaign, USA in 1980 and 1992, respectively. Since 1993, he has been a professor at School of Computing, Soongsil University, Korea.

Brian J. d'Auriol received the BSc(CS) and Ph.D. degrees from the University of New Brunswick in 1988 and 1995, respectively. Currently, he is an Assistant Professor in the Department of Computer Engineering at Kyung Hee University, Global Campus, Republic of Korea. Previously, he had been a researcher at the Ohio Supercomputer Center, USA and Assistant Professor at The University of Texas at El Paso, USA and at The University of Manitoba, Canada; and a Visiting Assistant Professor at The University of Akron, USA, and Wright State University, USA. He has organized and chaired the International Conference on Communications in Computing (CIC) 2000-2008 and the 11th Annual International Symposium on High Performance Computing Systems (HPCS'97) in 1997. He has published over 75 papers in international journals and conferences. His research includes information and data visualization, optical bus computing models and ubiquitous sensor networks. He is a member of the ACM and IEEE (Computer Society).