A New Feature Selection Scheme Using a Data Distribution Factor for Unsupervised Nominal Data

Tommy W. S. Chow, *Senior Member, IEEE*, Piyang Wang, and Eden W. M. Ma

***Abstract*—A new efficient unsupervised feature selection method is proposed to handle nominal data without data transformation. The proposed feature selection method introduces a new data distribution factor to select appropriate clusters. The proposed method combines the compactness and separation together with a newly introduced concept of singleton item. This new feature selection method considers all features globally. It is computationally inexpensive and able to deliver very promising results. Eight datasets from the University of California Irvine (UCI) machine learning repository and a high-dimensional cDNA dataset are used in this paper. The obtained results show that the proposed method is very efficient and able to deliver very reliable results.**

***Index Terms*—Clustering, feature ranking, unsupervised feature selection.**

I. Introduction

FEATURE reduction is a generic term for a process that aims at reducing features with certain criteria. It is an important preprocessing tool for high-dimensional datasets as high dimensionality is unaffordable for many existing algorithms. There are two common approaches to achieve this goal: 1) feature extraction and 2) feature selection [1]. Feature extraction transforms a given feature set into a feature subset with fewer features. As the feature set is transformed, the physical meaning of each original feature is blurred. Thus, referring to the original feature set is usually required for feature-subset interpretation. Principle component analysis, which is designed to capture the variance of a dataset in terms of principle components, is a popular unsupervised feature-extraction method. In contrast, feature selection reduces features from a given feature set without transformation. It retains the physical meaning of the selected features and provides clues for data collection or further analysis. There are numerous supervised feature selection methods [2], [3], but unsupervised selection method is rare.

Usually, feature selection is performed in a way of feature ranking [12], [13], in which selected features are individually, instead of globally, considered. Thus, the selected features may not always be the optimal features in representing a given dataset. The feature selection can be formalized as a combinational optimization problem finding a feature set to maximize the quality of the hypothesis learned from these features through ranking all features. In some of the well-known numerical feature selection approaches, features are ranked in mutual information (MI) [12], [13], [22] or Bayesian discriminant [21]. Top ranked features are selected iteratively, with MI-based criteria and a forward searching process. In these cases, the number of features to be selected may be difficult to be determined.

Feature selection conducted in the previously described ways [12], [13] usually leads to effective data dimensionality reduction, but these methods are only designed for numerical-data problems. Each unsupervised feature selection inherits the characteristics of its employed clustering algorithm, inclusive of the evaluation criteria [4]. Unsupervised feature selection schemes, such as [5]–[9], are able to handle with the numerical data easily, but they are inept to handle nominal data such as color and brand name. This is because their evaluation criteria involve distance calculation which is not applicable to nominal data that carry no order information. Converting nominal data into binary data and inserting order into nominal data are common approaches when the nominal data are handled by using a numerical feature selection scheme. However, such a data conversion usually results in generating a lot of irrelevant features. In addition, the conversion may make the interpretation of a feature selection result difficult or even impossible. To transform a dataset from nominal to numerical, inserting order is required [10]. However, the performance of this approach is quite dataset-dependent. The method proposed in [11], called SUD (Sequential Backward Selection algorithm to determine the relative importance of variables for Unsupervised Data) in this paper, is an unsupervised feature selection method. The SUD can handle the nominal data without any transformation or conversion. It uses entropy similarity measurement to determine the importance of features with respect to the underlying clusters. Features are ranked according to the entropy similarity measurement in the following ways. First, each feature in the feature subset is removed in turn to perform entropy calculation on the remaining feature subset without that particular feature. The feature, which has been taken out from the feature subset, that scores the minimum entropy is regarded as the least important in the subset and removed in the next iteration. Fig. 1 shows the mechanism of SUD in a block diagram. Due to the iteratively entropy estimation, the SUD is highly computationally expensive. If there are m features in the dataset, $(2+m)\times(m-1)/2$ entropy estimations are required. In addition, it is worth noting that the measurement of SUD may suffer from inevitable problems posed by high dimensionality, such as estimates of entropy, probability density function, and similarity in high-dimensional space [12], [13].

Manuscript received March 22, 2007; revised September 25, 2007. This paper was recommended by Associate Editor P. S. Sastry.

The authors are with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: eetchow@cityu.edu.hk).

Digital Object Identifier 10.1109/TSMCB.2007.914707

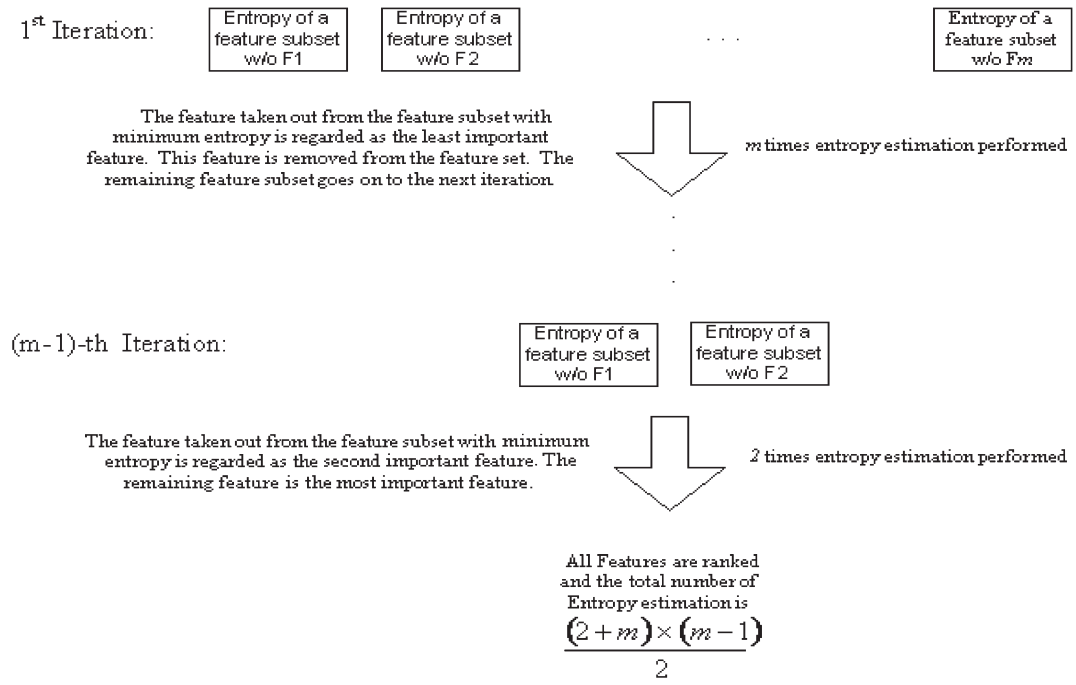


Fig. 1. Block diagram of the SUD.

In this paper, an efficient unsupervised feature selection scheme for nominal data (UFSN) is proposed. The proposed feature selection scheme, called UFSN, is able to directly process nominal dataset. As a result, the shortcomings of converting data from nominal into binary are eliminated. To the best of our knowledge, today, only the SUD and the proposed scheme are able to perform the unsupervised feature selection on the nominal data without binary conversion or data transformation. In this paper, we propose a very computationally efficient approach for handling the nominal-data feature selection by means of ranking features. In the proposed method, we use a relevance index (REL), which will be detailed in a later section, to determine the importance of every feature. The evaluated REL value of each feature is compared and ranked. This enables appropriate features to be ranked and selected. It is worth noting that the evaluation of REL is not affected by the sequence of feature selection. In the obtained results, the proposed scheme is found to be up to 100 times more efficient than the SUD because the UFSN does not require the iterative calculation of entropy. As there is no class label provided, a clustering algorithm must be firstly used to generate a set of clustering results, which are called cluster descriptions in this paper. These cluster descriptions are generated by using a clustering algorithm with different parameter settings. A data distribution factor (DDF) is newly introduced to select an appropriate cluster description from the provided cluster descriptions for further measurement. DDF is the combination of both compactness and separation. Similar objects are grouped into the same cluster for obtaining a high compactness. For the separation part, another new idea of singleton item is introduced for handling the nominal data. We show that the more singleton items there are in a separation among clusters, the higher the entropy will be. This indicates an important relationship between the separation and the number of singleton items. Clearly, they have a very similar nature in terms of clustering information. In addition, it is worth noting that the determination of singleton item is very computationally efficient.

Usually, without the knowledge on how much information is contained in the selected features, most feature selection methods have a difficulty in stopping the selection process analytically. Instead, the selection process can only be terminated at a predetermined point that results in the selected features either missing out certain useful information or including certain redundancy features. This problem is overcome in this paper by introducing the cluster number as stopping criteria. Our obtained results show that the proposed method can be up to 100 times faster compared with the SUD. A REL using the concept of singleton item is then developed to evaluate features individually. Based on the selected cluster description, this paper shows that the newly proposed scheme can deliver very promising results. This paper is organized as follows. Section II presents the proposed feature selection method. Section III displays the simulation platform for all the datasets. Section IV shows our intensive simulations based on eight UCI datasets and a cDNA dataset with 12 600 features. At last, a conclusion is drawn in Section V.

II. Proposed Feature Selection Mechanism

In this paper, a dataset with m features and N nominal data instances, which are also referred as transactions in [14]–[16], is considered. The nominal dataset consists of finite values in each feature, and there is no order information among these values. For example, there is no order information in color, i.e., red is not closer to blue than green. The proposed feature selection scheme is composed of three parts. First, a clustering algorithm that can handle the nominal data is used to generate a set of cluster descriptions, which describe the data

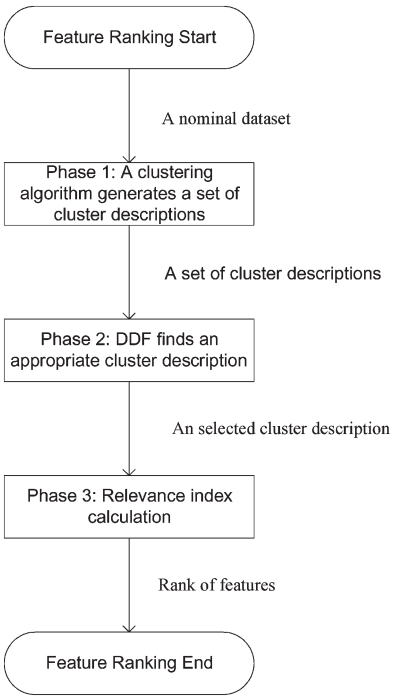


Fig. 2. Procedure of the proposed method.

characteristics in a clustering sense. Then, the DDF is used to select an appropriate cluster description for further relevance ranking. Finally, a newly developed feature REL is applied for feature ranking. Fig. 2 shows the block diagram of the proposed scheme. A clustering algorithm is employed in Phase 1 to generate a set of cluster descriptions, which are the output of Phase 1. From these descriptions, only one cluster description selected by the DDF in Phase 2 will be passed to Phase 3 for feature relevance rank. Features are ranked in Phase 3, and the number of relevant features is determined by a threshold called IrrThreshold. A feature with a relevance value higher than the IrrThreshold is recognized as a relevant feature; otherwise, it is an irrelevant feature.

A. Phase 1—Generation of Cluster Descriptions

Clustering is an unsupervised process aiming at grouping similar objects into the same cluster and separating dissimilar objects into different clusters. Any clustering algorithm that is able to handle the nominal data can be used in this phase to generate a set of cluster descriptions with different parameter settings. In this paper, two different transactional clustering algorithms, namely, Clustering with sLOPE (CLOPE) [14] and small–large ratio (SLR) [15], are used in this phase for comparison since they do not use order information to cluster data.

CLOPE needs larger height-to-width ratio of the histogram to indicate a better intracluster similarity. In its clustering algorithm, criterion functions should be defined and optimized so that the intracluster similarity and the intercluster dissimilarity can be maximized. The shape of every cluster and the number of transactions in a given dataset must be considered in order to have the criterion function defined.

There are two phases in the SLR, which are the allocation phase and the refinement phase. Each transaction is read one

Nominal Data

Instance \ Feature	F1	F2	F3	F4	F5	F6
1	a	a	?	?	f	d
2	a	c	h	i	g	g
3	b	d	j	i	?	?
4	b	d	h	i	j	?

Connect each value of a feature with its corresponding feature's name

Transactional -Nominal Data

Transaction	Information of transaction
1	F1_a, F2_a, F5_f, F6_d
2	F1_a, F2_c, F3_h, F4_i, F5_g, F6_g
3	F1_b, F2_d, F3_j, F4_i
4	F1_b, F2_d, F3_h, F4_i, F5_j

Fig. 3. Transactional–nominal data format.

by one in the allocation phase and then assigned to a cluster in order to minimize the total cost of clustering. The transaction may be assigned to the existing cluster or a new cluster. In the refinement phase, the support values of items computing and clusters searching and the intermediate support values of items calculating are performed in sequence. All the data used in this paper are in a form of transactional–nominal data, which is modified to fit the algorithms without changing the meaning of data.

The conversion of nominal data to fit into transactional clustering is discussed next. For transactional data, each value is unique; number of items is not restricted, and permutation of items is irrelevant. If there are any missing values, some form of "?" must be supplemented to avoid disorder. Obviously, the nominal data format, which is unable to distinguish features, is not suitable for the transactional clustering. As a result, the nominal data are transformed into transactional–nominal format by means of connecting each value to the name of the feature to form an item (Fig. 3). Missing values are ignored instead of supplemented with dummy values, and the permutation of the features can be altered. While inheriting the merits of the transactional data, the fundamental meaning of each feature is reserved. The number of value is variable, which is capped at the number of features. This transactional–nominal format is used to present the nominal data throughout this paper.

In this phase, a set of cluster descriptions with different parameter settings is generated. A transactional clustering algorithm is used to generate a set of cluster descriptions. CLOPE [14] and SLR [15], which are the commonly used efficient algorithms for processing the transactional data, are used to make a comparative study. The CLOPE and the SLR, which do not require transforming data into binary type, have one and three user input parameters, respectively. The CLOPE uses the height-to-width ratio of the cluster histogram to determine clusters. The parameter r is used to control the tightness of the clusters. SLR is an enhanced version of LargeItem [16], which employs the large-item idea from association rule. The SLR introduces a middle item, which is an item belonging to neither a large item nor a small item. Three user input parameters (minimum support, MinSup, damping factor λ, and SLR threshold α) are required. If the support of an item in a cluster is larger than the minimum support, it is considered a large item. On the contrary, if the support of an item in a cluster is smaller than the maximum ceiling, which is defined as the product of

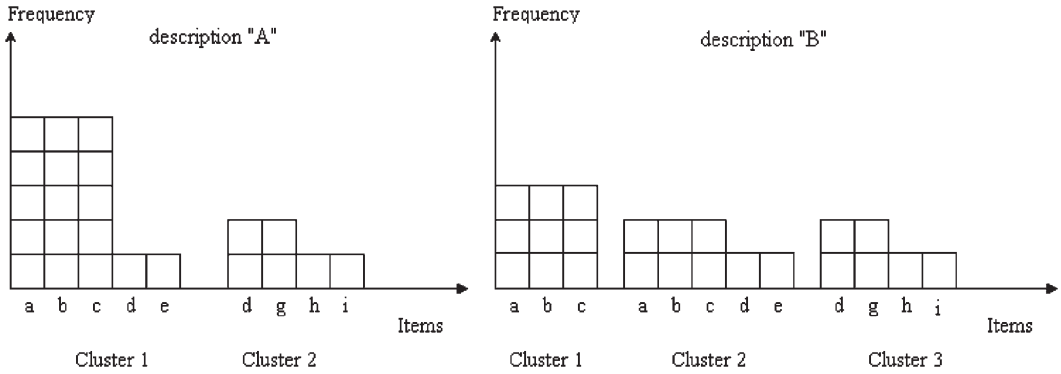


Fig. 4. Item distribution in cluster descriptions "A" and "B."

minimum support and damping factor, it is considered a small item. If the small-to-large-item ratio is greater than the SLR threshold α, the transaction is regarded as an excess transaction and handled separately. In this paper, the parameters are set according to [15].

B. Phase 2—Cluster Description Selections

In this phase, an appropriate cluster description is selected by means of DDF. An appropriate cluster description is a cluster description with the highest DDF and with a number of clusters greater than one. The DDF is defined as follows:

$$\begin{aligned}\mathrm{DDF} &= \frac{\sum_{j=1}^{k}|C_j|\left[\frac{\sum_{F_i\_\nu\in D(C_j)}\mathrm{Frequency}(F_i\_\nu,C_j)}{|D(C_j)|\times|C_j|}\right]}{\sum_{j=1}^{k}|C_j|} \\ &\quad+\frac{\sum_{F_i\_\nu\in\mathrm{Singleton}}\mathrm{Frequency}(F_i\_\nu)}{\sum_{j=1}^{k}\sum_{F_i\_\nu\in D(C_j)}\mathrm{Frequency}(F_i\_\nu,C_j)} \\ &= \frac{\sum_{j=1}^{k}\left[\frac{\sum_{F_i\_\nu\in D(C_j)}f(F_i\_\nu,C_j)}{|D(C_j)|}\right]}{\sum_{j=1}^{k}|C_j|} \\ &\quad+\frac{\sum_{F_i\_\nu\in\mathrm{Singleton}}f(F_i\_\nu)}{\sum_{j=1}^{k}\sum_{F_i\_\nu\in D(C_j)}f(F_i\_\nu,C_j)}\end{aligned} \tag{1}$$

where $f(F_i\_\nu, C_j)$ is the frequency of value F_i_ν in cluster C_j, $|D(C_j)|$ is the number of the distinct values in cluster j, and $|C_j|$ is the number of instances of cluster C_j.

The first part of DDF evaluates the compactness of the cluster description. As one of the purposes of clustering is to group similar objects into the same cluster, a high compactness within a cluster means the objects in the cluster exhibit high similarity. The average frequency of values within the cluster is determined by $\sum_{F_i\_\nu\in D(C_j)} f(F_i\_\nu, C_j)/|D(C_j)| \times |C_j|$, where $|D(C_j)| \times |C_j|$ is the maximum total possible frequency of values. The average frequency of values is weighted by the ratio of the number of instances.

The second part of DDF evaluates the separation of the cluster description. Moreover, another purpose of clustering is to group dissimilar objects into different clusters. In the following example, the concept of entropy is used to illustrate the relationship between the number of singleton items and its cluster separation.

Example 1: A dataset consists of seven transactions: $\mathrm{t1} = \{\mathrm{a},\mathrm{b},\mathrm{c}\}$, $\mathrm{t2} = \{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d}\}$, $\mathrm{t3} = \{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{e}\}$, $\mathrm{t4} = \{\mathrm{a},\mathrm{b},\mathrm{c}\}$, $\mathrm{t5} = \{\mathrm{d},\mathrm{g},\mathrm{h}\}$, $\mathrm{t6} = \{\mathrm{d},\mathrm{g},\mathrm{i}\}$, and $\mathrm{t7} = \{\mathrm{a},\mathrm{b},\mathrm{c}\}$. Cluster description "A" consists the following: $\mathrm{C1} = \{\mathrm{t1},\mathrm{t2},\mathrm{t3},\mathrm{t4},\mathrm{t7}\}$ and $\mathrm{C2} = \{\mathrm{t5},\mathrm{t6}\}$. Cluster description "B" consists the following: $\mathrm{C1} = \{\mathrm{t1},\mathrm{t4},\mathrm{t7}\}$, $\mathrm{C2} = \{\mathrm{t2},\mathrm{t3}\}$, and $\mathrm{C3} = \{\mathrm{t5},\mathrm{t6}\}$. Fig. 4(a) and (b) shows the frequency of the items in cluster descriptions "A" and "B," respectively.

Obviously, for description "A," item "d" appears in both clusters 1 and 2. This means that "a," "b," "c," "e," "g," "h," and "i," which appear in only one cluster, are all singleton items. Similarly, "e," "g," "h," and "i" are singleton items in description "B." Therefore, there are seven singleton items in description "A" and four in description "B." The entropies for cluster descriptions "A" and "B" are 0.9597 and 0.53, respectively. By using the concept of entropy, this typical example shows that a clear separation among the clusters has more singleton items. The frequency of the singleton values is counted in the second part of the DDF. A cluster description with the highest DDF and the number of clusters greater than one is chosen for relevance rank in Phase 3.

C. Phase 3—Relevance Rank

Based on the selected cluster description, relevance value of each feature is evaluated. A feature is a relevant feature when its relevance value is higher than or equal to the threshold, IrrThreshold. The REL, $\mathrm{REL}(F_i)$, of feature i is defined as follows:

$$\mathrm{REL}(F_i) = \frac{|\mathrm{Singleton}|}{|F_i|} \times \frac{N - \mathrm{Miss}(F_i)}{N} \tag{2}$$

where $|\mathrm{Singleton}|$ is the number of singleton values, $|F_i|$ is the number of values in ith feature f_i, N is the number of instances,

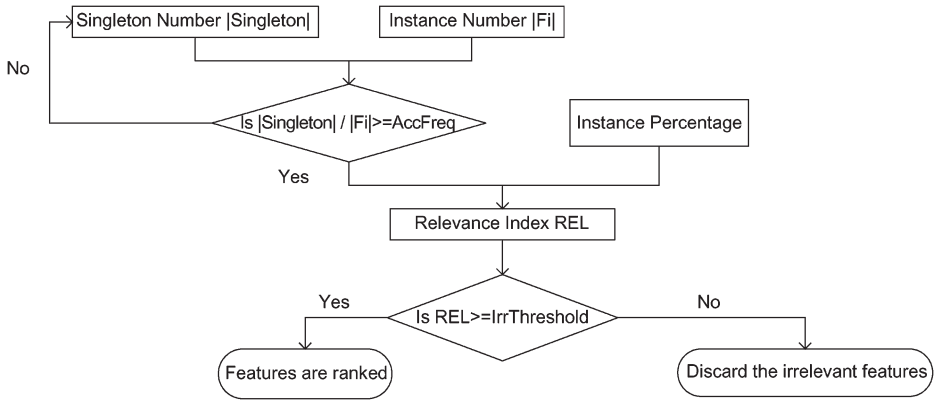


Fig. 5. Feature ranking process.

TABLE I
THREE CLUSTERS AND THEIR FEATURE VALUE DISTRIBUTIONS

Data Instances	F1	F2	F3	F4	F5	F6	F7	Cluster Number
1	1_0	2_1	3_0	4_1	5_0	6_1	7_0	1
2	1_0	2_1	3_1	4_0	5_0	6_1	7_0	1
3	1_0	2_1	3_0	4_0	5_0	6_1	7_0	1
4	1_1	2_0	3_1	4_0	5_1	6_2	7_0	2
5	1_1	2_0	3_0	4_0	5_1	6_2	7_2	2
6	1_1	2_0	3_1	4_0	5_1	6_2	7_2	2
7	1_1	2_1	3_0	4_0	5_0	6_0	7_3	3
8	1_1	2_1	3_1	4_0	5_0	6_0	7_3	3
9	1_1	2_1	3_0	4_1	?	6_0	7_3	3
Relevance Index $REL(f_i)$	$\frac{1}{2}\times 1$ = 0.5	$\frac{1}{2}\times 1$ = 0.5	$\frac{0}{2}\times 1$ = 0	$\frac{0}{2}\times 1$ = 0	$\frac{1}{2}\times\frac{8}{9}$ = 0.444	$\frac{3}{3}\times 1$ = 1	$\frac{2}{3}\times 1$ = 0.67	

and $\mathrm{Miss}(F_i)$ is the number of instances with a missing value in F_i.

A relevant feature is a feature grouping instances according to the cluster description closely. If all values of a particular feature are singleton values, this feature groups the instances exactly according to the cluster description. Hence, higher $|\mathrm{Singleton}|/|F_i|$ means that the feature groups the instances more closely to the cluster description. Since there may be missing values in features, $[N - \mathrm{Miss}(F_i)]/N$ is used to weigh the instance value percentage of a feature.

As aforementioned, a singleton value is defined as a value appearing in a single cluster. There are, however, cases that such definition seems too stringent. A parameter acceptable frequency ($\mathrm{AccFreq}$) is proposed to loosen the definition and satisfy those cases if required. If a value appears in more than one cluster and it mostly occurs in one cluster (i.e., its frequency in one cluster is greater than or equal to $\mathrm{AccFreq}$), it is still regarded as a singleton value. For example, there are 98 instances in cluster A and two instances in cluster B among 100 instances containing the value of "1_T." Since the $\mathrm{AccFreq}$ is 98%, "1_T" is a singleton value if $\mathrm{AccFreq} = 95\%$ and not a singleton value when $\mathrm{AccFreq} = 99\%$. When $\mathrm{AccFreq} = 100\%$, the definition of singleton value is the same as the original one, i.e., the value occurs in a cluster only. Throughout this paper, $\mathrm{AccFreq} = 100\%$ is used in all investigations. It is worth noting that the $\mathrm{REL}(F_i)$, which is the label used for the subsequent unsupervised clustering, is related to the IrrThreshold and the $\mathrm{AccFreq}$. Apparently, the $\mathrm{REL}(F_i)$ is a synthetic class label, which is different from the real labels used for supervised clustering method.

By comparing the ratio $|\mathrm{Singleton}|/|F_i|$ with the parameter $\mathrm{AccFreq}$, the finalized number of singleton number can be determined. As aforementioned, $\mathrm{AccFreq}$ is set to 100% as a reference. After obtaining the $|\mathrm{Singleton}|/|F_i|$ and $[N - \mathrm{Miss}(F_i)]/N$, the REL is obtained. The REL of each feature is calculated and compared. Thus, we are able to discard features with REL less than IrrThreshold. Apparently, a feature with the highest REL value is regarded as the most relevant feature. Fig. 5 shows the process of feature ranking.

Example 2: This example illustrates how features are ranked. The concept of REL, IrrThershold, and $\mathrm{AccFreq}$ is also elaborated. The dataset used in this example has nine data instances and seven features in three clusters. Table I shows the data and the REL with respect to each feature. According to the index, the relevance of features should be in an order of $\mathrm{F6} > \mathrm{F7} > \mathrm{F1} = \mathrm{F2} > \mathrm{F5} > \mathrm{F3} = \mathrm{F4}$. It is clear that feature F6 is the most relevant feature, whereas features F3 and F4 are the irrelevant ones. The distribution of feature F6 is identical to the cluster description, whereas feature F3 is random. Feature F4, which appears once in two different clusters, has only one different value. Feature F7 provides more information than features F1, F2, and F5, as the former distinguishes two classes while the latter distinguishes one cluster from others. The REL of F5 is less than those of F1 and F2 because F5 has one missing value, whereas F1 and F2 do not contain any missing value. For instance, if IrrThreshold is set at zero, features F3 and F4 are

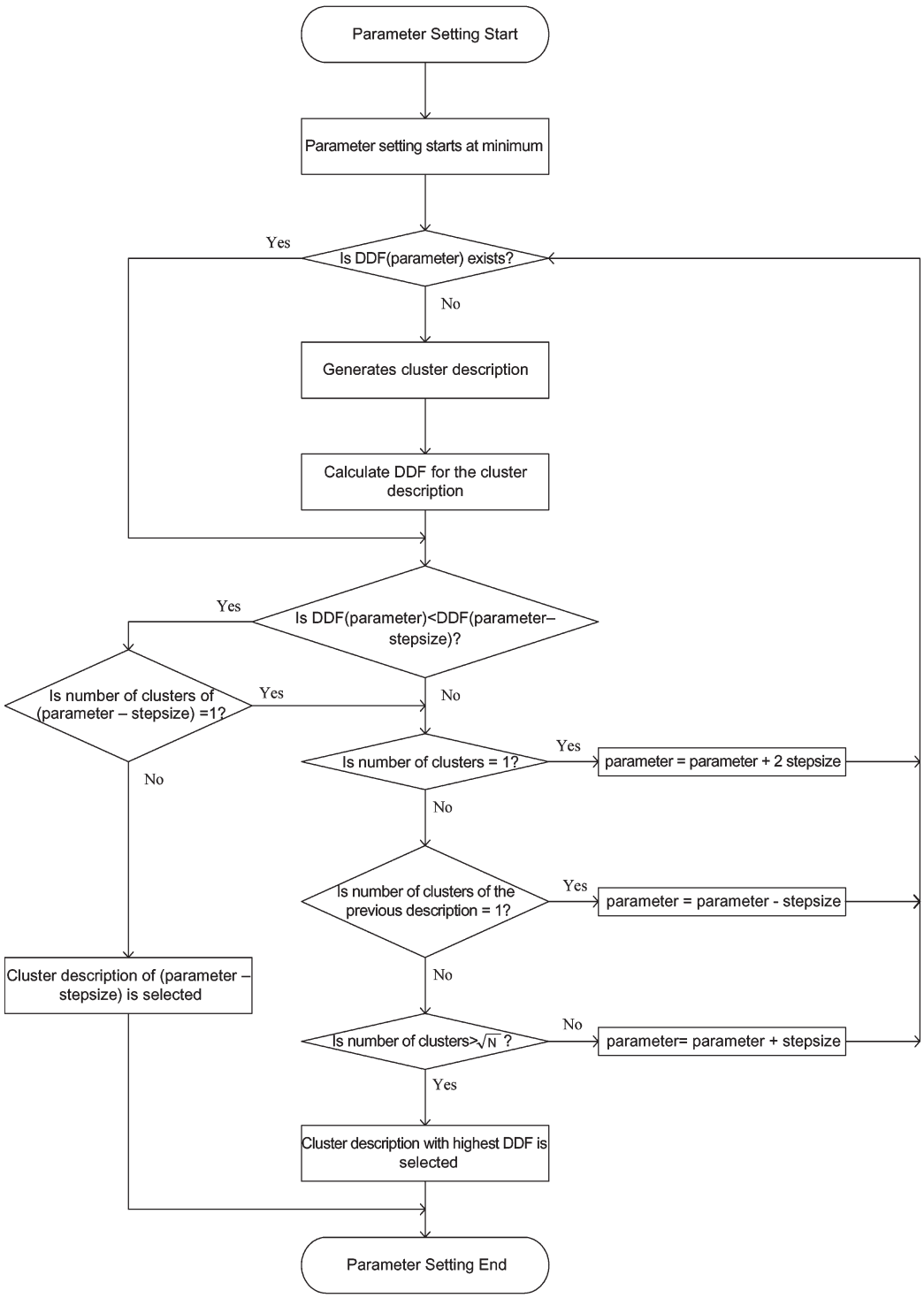


Fig. 6. Enhanced version of the proposed feature selection scheme.

recognized as irrelevant features. If IrrThreshold is set at 0.5, features F1–F5 are considered irrelevant. If $\mathrm{AccFreq}$ is ≤ 0.75, the REL of feature F7 is changed from 0.67 to 1.

D. Enhanced Version

To reduce the computational resources, an enhanced version of UFSN, called EUFSN, is designed. In some clustering algorithms, the number of clusters can be roughly estimated by their parameters. For example, the number of clusters tends to increase when r in CLOPE increases. Based on this property, the enhanced scheme changes the parameter automatically. Fig. 6 shows a block diagram of the EUFSN. In the enhanced scheme, one of the stopping criteria is that the number of clusters is greater than $\sqrt{N}$, where N is the number of data instances. This criterion is used to prevent choosing a cluster description with the number of clusters close to the number of instances. In some datasets, the second part of DDF is very low

TABLE II
PARAMETER CHANGE OF THE ENHANCED SCHEME

Initial:	Iteration	Parameter	DDF	Cluster Number
Minimum Parameter=0.1; Step size=0.1; Instance Number=100	1	0.1	1.47	1
	2	0.3	1.47	1
	3	0.5	1.47	1
	4	0.7	1.47	1
	5	0.9	1.04	2
	6	0.8	1.47	1
	7	1.0	1.85	2
	8	0.9	1.04	2
	9	1.0	1.85	2
	10	1.1	0.85	3
	Selected	1.0	1.85	2

for all cluster descriptions. In these cases, the DDF increases when the number of clusters increases. Hence, it is suggested that only the cluster descriptions with the number of clusters between 2 and $\sqrt{N}$ [17] are evaluated.

Instead of generating all clustering descriptions in Phase 1 and evaluating them in Phase 2, the enhanced scheme combines Phases 1 and 2 to save the computational resources. First of all, the enhanced scheme generates a cluster description with a parameter at the minimum and evaluates the description. Then, the parameter is raised by a user-predefined step-up size and generates a cluster description with the new parameter setting, if necessary. The parameter step-up repeats and goes on to generate and evaluate cluster description until the cluster description with the highest DDF value is determined. The cluster description with the highest DDF proceeds to Phase 3.

In case there is a relationship, either direct or inverse proportion, between parameter settings and number of clusters and that the step size is small, it is probable that the consecutive cluster descriptions are the same. If the number of clusters of a cluster description is equal to one, it is very likely that the number of the consecutive cluster description is also equal to one. The enhanced scheme, hence, doubles the step size of the parameter and investigates the number of clusters in the cluster description. The process repeats until there is more than one cluster in the cluster description. Likewise, a step-down size is set to generate the skipped cluster description for the DDF evaluation.

Table II is used to demonstrate how the parameter r varies in the enhanced scheme. First, the initial r is set at 0.1, whereas the step size is set at 0.1. Since the number of clusters in the first four iterations is one, it increases with a doubled step size of 0.2. At the fifth iteration, the number of clusters is two, whereas the number of clusters at the fourth iteration is one. Due to the cluster description with r of 0.8 being skipped, a step-down size is set to determine the cluster description with r equals 0.8. It is then found that there is only one cluster in the cluster description when r is 0.8. The parameter r, again, is increased by a doubled-step size, i.e., 0.2, for the next iteration. Similarly, the cluster description with r of 0.9 should be determined in the eighth iteration. However, the cluster description with r of 0.9 has already been determined in a prior iteration. As a result, the cluster description generation and evaluation when r is equal to 0.9 is skipped. Although the DDF of the cluster description with r of 0.8 is greater than that of 0.9, the cluster description with r of 0.8 has one cluster only and cannot be selected. The scheme continues by stepping up the parameter until the peak of DDF is determined or the number of clusters of a cluster description is greater than $\sqrt{N}$, where N is the number of data instances.

III. SIMULATION PLATFORM

In this paper, we use eight real datasets from the UCI machine learning repository [18] and a cDNA dataset with 12 600 features about prostate cancer [19] to demonstrate the ability of the proposed method. In order to simulate a real-world situation, all class labels of the UCI datasets and the cDNA dataset are removed. During the feature selection process, no class label is involved. Classification accuracy is provided to present a brief idea of the results from the proposed feature selection method. The classification accuracy is generated by using the J48 decision tree in Weka [20], which is a modified version of a C4.5 decision tree, on the selected feature subset and the class label. All classification accuracies are obtained by the J48 decision tree with a tenfold cross validation in Weka. In the tenfold cross validation, the original data are separated into ten subsamples. Of the ten subsamples, a single subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. The cross-validation process is then repeated ten times, with each of the ten subsamples used exactly once as the validation data.

In a real application where the class label is not available, it is difficult to determine if the selected features are correct or not. In the problem studied in this paper, we assume that the relevant features are the features describing the synthetic cluster labels. Thus, the classification accuracy of the selected feature subset and the synthetic cluster labels provides a measure on the performance.

IV. RESULTS

In this paper, the SUD [11] is compared with the proposed UFSN, where $\mathrm{AccFreq}$ is set at one for all comparisons. The method CLOPE used in Phase 1 is called UFSN-CLOPE, and the scheme based on CLOPE with enhanced scheme is called EUFSN-CLOPE. The UFSN-CLOPE generates cluster descriptions where a setting of r starts from 0.1 and a step size of 0.1. It stops when the number of clusters is greater than $\sqrt{N}$. Similar to the UFSN-CLOPE, the EUFSN-CLOPE starts at r of 0.1 and a step size of 0.1. The SLRs used in Phase 1 with α at 0.5, 1.5, and 2.5 are called UFSN-SLR-0.5, UFSN-SLR-1.5, and UFSN-SLR-2.5, respectively. For UFSN-SLR-0.5, UFSN-SLR-1.5, and UFSN-SLR-2.5, the cluster description is generated with a minimum support at 0.6 and λ varying from 0.4 to 1.

The comparative computational time among UFSN-CLOPE, EUFSN-CLOPE, UFSN-SLR-0.5, UFSN-SLR-1.5, UFSN-SLR-2.5, and SUD is shown in Table III. It shows that the SUD, which delivers the same level of classification accuracy, is a very computational demanding method compared with the proposed scheme. The UFSN-CLOPE is extremely similar to the EUFSN-CLOPE with much longer computational time. UFSN-SLR-0.5, UFSN-SLR-1.5, and UFSN-SLR-2.5 are faster than EUFSN-CLOPE except dataset "cDNA." Table IV illustrates the selected cluster description of each

TABLE III
COMPARISON OF THE COMPUTATIONAL TIME (IN SECONDS) OF DIFFERENT METHODS

Dataset	UFSN-CLOPE	EUFSN-CLOPE	UFSN-SLR-0.5	UFSN-SLR-1.5	UFSN-SLR-2.5	SUD
Agaricus-lepiota	43784	8685	1255	897	833	Terminated after 66th day
Balance-Scale	77	30	14	13	12	1314
Breast-Cancer	63	14	12	8	8	1216
Breast-Cancer-Wisconsin	175	57	46	51	51	7530
Hepatitis	30	13	8	6	6	1753
House-Votes-84	133	26	10	11	11	9938
Lung-Cancer	7	3	1	1	1	948
Soybean-small	16	3	1	1	1	603
cDNA	5642	2368	21170	25734	21406	Terminated after 72 hours

TABLE IV
SELECTED CLUSTER DESCRIPTION OF EACH DATASET IN THE PROPOSED SCHEME

	UFSN-CLOPE	EUFSN-CLOPE	UFSN-SLR-0.5	UFSN-SLR-1.5	UFSN-SLR-2.5
Dataset	r		λ		
Agaricus-lepiota	4.4	0.9	1	0.8	0.5
Balance-Scale	0.3	0.3	0.4	0.4	0.4
Breast-Cancer	1.1	0.4	0.6	0.8	0.8
Breast-Cancer-Wisconsin	0.7	0.7	1	1	1
Hepatitis	1	1	1	1	N/A
House-Votes-84	0.9	0.9	0.7	0.7	0.7
Lung-Cancer	0.9	0.9	0.9	0.9	0.9
Soybean-small	1.1	1.1	1	N/A	N/A
cDNA	0.6	0.6	0.4	0.6	0.7

TABLE V
INFORMATION OF THE UCI DATASETS

	Agaricus-lepiota	Balance-Scale	Breast-Cancer	Breast-Cancer-Wisconsin	Hepatitis	House-Votes-84	Lung-Cancer	Soybean-small
Number of Features	22	4	9	9	19	16	56	35
Number of instances	8124	625	286	699	155	434	32	47
Number of classes	2	3	2	2	2	2	3	4
Number of missing values	2480	0	9	16	167	376	5	0

dataset in UFSN-CLOPE, EUFSN-CLOPE, UFSN-SLR-0.5, UFSN-SLR-1.5, and UFSN-SLR-2.5.

A. UCI Datasets

Eight nominal UCI datasets are used to demonstrate the ability of the proposed method. The UCI repository provides benchmark datasets with class label for machine learning. Table V depicts the detail information of each dataset. The classification accuracies of different feature subsets conducted via EUFSN-CLOPE, UFSN-SLR-0.5, and SUD are shown in Fig. 7. Meanwhile, Fig. 7 shows a brief comparison between the proposed schemes (both EUFSN-CLOPE and UFSN-SLR-0.5) and the SUD.

Classification accuracies with respect to datasets "Agaricus-Lepiota" and "Soybean-Small" are discussed in detail in the following paragraphs. The dataset "Agaricus-Lepiota" consists of 22 features and 8124 data instances with 2480 missing data in feature 11 (stalk–root). The dataset is composed of two classes: "edible" and "poisonous." The features describe different fundamental characteristics such as odor and cap color. This is the only one UCI dataset with different cluster descriptions selected by the UFSN-CLOPE and the EUFSN-CLOPE. The cluster description with r of 4.4, smaller than

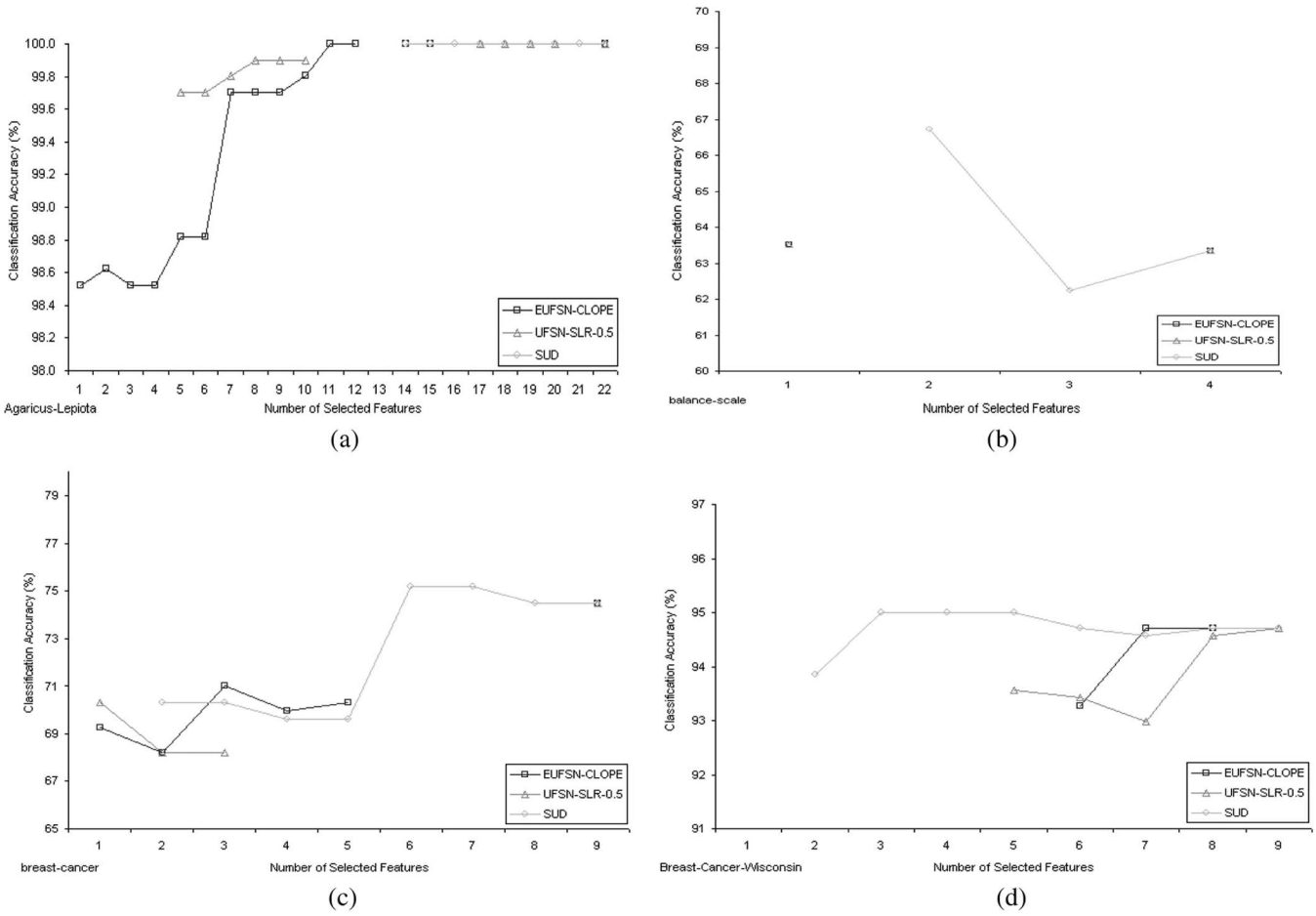


Fig. 7. Classification accuracies for different numbers of features selected by EUFSN-CLOPE, UFSN-SLR-0.5, and SUD of the following datasets: (a) "Agaricus-Lepiota," (b) "Balance-Scale," (c) "Breast-Cancer," (d) "Breast-Cancer-Wisconsin."

$\sqrt{N}$, is selected by the UFSN-CLOPE. This is a typical case that the DDF value of a cluster description increases when r increases. Classification accuracies of various numbers of features selected by the UFSN-CLOPE and the EUFSN-CLOPE for dataset "Agaricus-Lepiota" are shown in Fig. 8. As the feature ranking is based on the cluster description, a cluster description with too many clusters may cause confusion about the feature relevance measurement.

The EUFSN-CLOPE reduces the dataset from 22 features to 11 features (50.0% reduced) without lowering the classification accuracy. When the number of features is reduced to one (95.5% reduced), the EUFSN-CLOPE is still able to maintain an accuracy of 98.5%, i.e., the most important feature is picked. In general, the classification accuracy in the cluster description generated by the EUFSN-CLOPE, the UFSN-SLR-0.5, the UFSN-SLR-1.5, and the UFSN-SLR-2.5 is comparable. The EUFSN-CLOPE outperforms the others when there is only one feature left. The UFSN-CLOPE and the EUFSN-CLOPE use 12.2 and 2.5 h to rank the features, respectively, whereas the UFSN-SLR-0.5, the UFSN-SLR-1.5, and the UFSN-SLR-2.5 use less than 0.5 h to rank the features of the same dataset. In addition, it takes more than 60 days to rank the eight features via SUD, and the process is subsequently terminated on the 66th day. As shown in the presented results, the proposed schemes are more efficient than the SUD.

The dataset "Soybean-Small" consists of 35 features and 47 data instances with no missing data. The "Soybean-Small" dataset has four class labels represented by D1–D4. Its features describe the planting situation such as temperature and area damages for soybean disease diagnosis. The classification accuracy of the original dataset is 97.9%. Both the UFSN-CLOPE and the EUFSN-CLOPE select the cluster description with r at 1.1, and the number of features is reduced from 35 to 5 (85.7% reduction) with 100% classification correctness, whereas the classification accuracy of SUD is 36.2% with 14 features. In terms of computational efficiency, the SUD used 603 s to rank the features, whereas the UFSN-CLOPE and the EUFSN-CLOPE required only 16 and 3 s, respectively.

To sum up, the classification accuracy and the number of selected features by the SUD are about the same as that of the proposed schemes. Nevertheless, the computational time of SUD, which is about 100 times on average, is substantially longer than the proposed scheme. It is clear that the proposed schemes select relevant features in a more efficient way compared with other methods.

B. cDNA

As there is no benchmark dataset with huge number of features available for examining the capability of handling

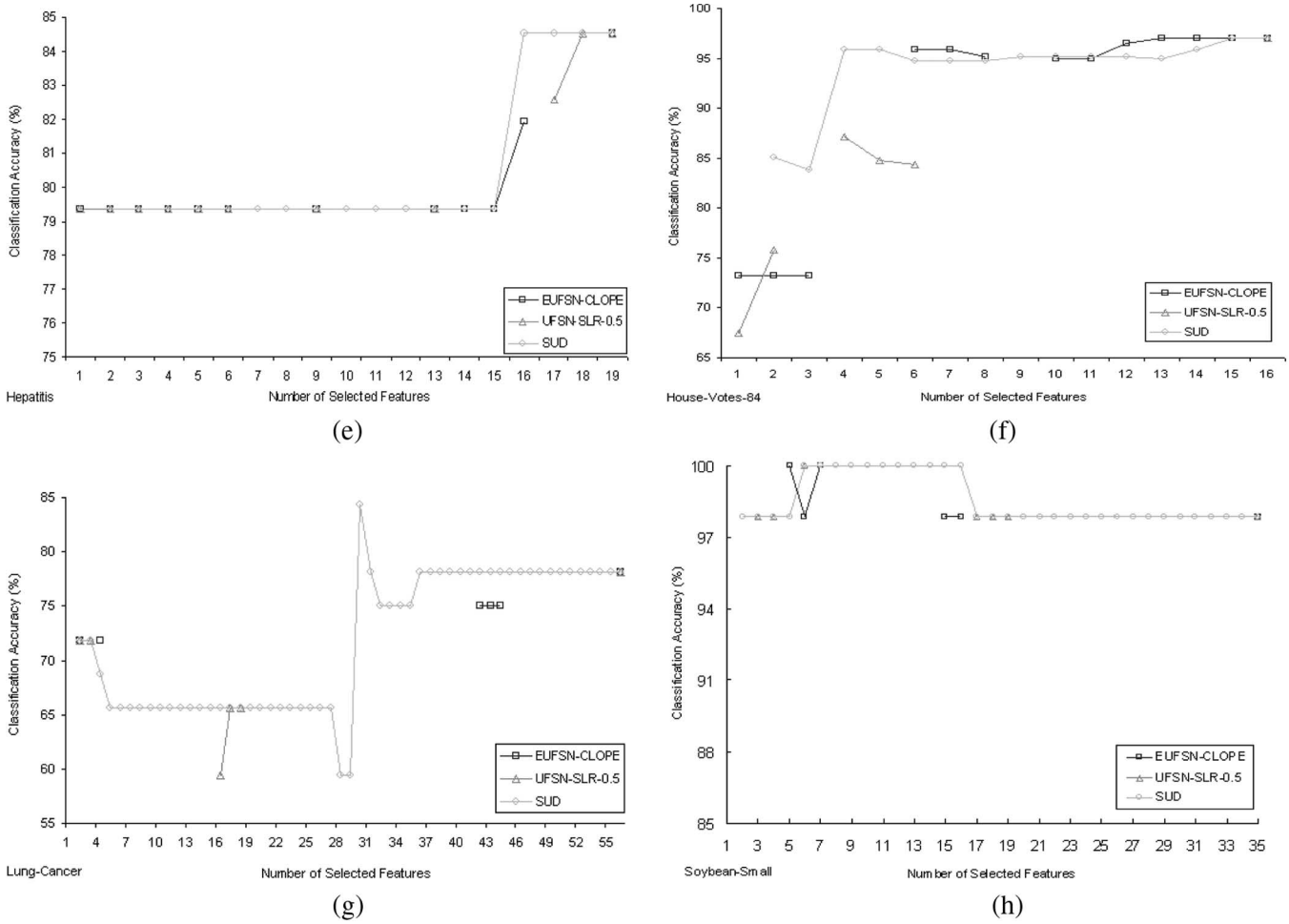


Fig. 7. (*Continued*.) Classification accuracies for different numbers of features selected by EUFSN-CLOPE, UFSN-SLR-0.5, and SUD of the following datasets: (e) "Hepatitis," (f) "House-Votes-84," (g) "Lung-Cancer," and (h) "Soybean-Small."

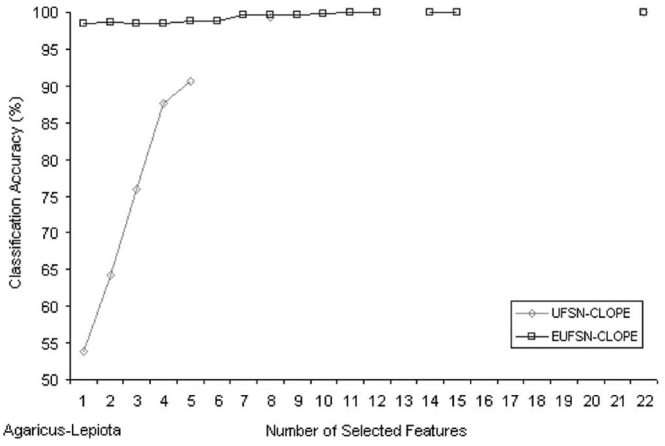


Fig. 8. Classification accuracies for different numbers of features selected by UFSN-CLOPE and EUFSN-CLOPE for dataset "Agaricus-Lepiota."

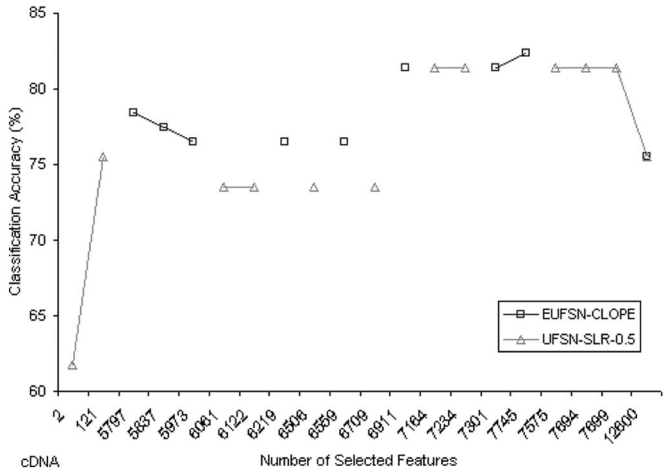


Fig. 9. Classification accuracies for different numbers of features selected by EUFSN-CLOPE, UFSN-SLR-0.5, and SUD of dataset "cDNA."

large number of feature datasets, a numerical cDNA dataset with 12 600 features is transformed into nominal format and subsequently used for illustration. The dataset "cDNA" is composed of 12 600 features and 102 data instances with no missing data. It contains two classes represented by A and B. Both the EUFSN-CLOPE and the UFSN-CLOPE select the cluster description with r of 0.6. The number of features is trimmed down from 12 600 features to 5797 (54% reduced) with an accuracy level of 78.4%, whereas the classification accuracy of the full dataset is 75.5%. The EUFSN-CLOPE used approximately 40 min to select the features, whereas the UFSN-CLPOE used 94 min. The SUD was terminated after 72 h without any features ranked. The proposed schemes are

proved again to be more efficient than the SUD. As there are too many features to be shown in a graph, Fig. 9 shows a part of the classification accuracy for various numbers of features selected by the EUFSN-CLOPE and the UFSN-SLR-0.5. The number of features in the minimum feature subset of the EUFSN-CLOPE is 5797, whereas it is one for the UFSN-SLR-0.5.

V. Conclusion

An efficient unsupervised feature selection scheme is developed to perform the nominal-data feature selection. The proposed scheme can be used with different clustering algorithms, for instance, CLOPE. The DDF is introduced as a stopping criterion for selecting cluster description for relevance ranking. A singleton item, which is proved to be similar in nature of finding the higher entropy, is developed for efficient clustering. Based on the selected cluster description, the relevance of the features is measured by using the proposed REL. The user is allowed to adjust the threshold, IrrThreshold, to control the number of features to be included. The relevance of features is measured individually and does not suffer from the problem caused by high dimensionality. The SUD is compared with the proposed scheme, and the obtained results show that the proposed scheme is a reliable and efficient feature selection methodology.

References

[1] R. Thawonmas and S. Abe, "Feature reduction based on analysis of fuzzy regions," in *Proc. IEEE Int. Conf. Neural Netw.*, 1995, vol. 4, pp. 2130–2133.

[2] G. V. Lashkia and L. Anthony, "Relevant, irredundant feature selection and noisy example elimination," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 2, pp. 888–897, Apr. 2004.

[3] L. Yu and H. Liu, "Efficiently handling feature redundancy in high-dimensional data," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 685–690.

[4] K. Z. Mao, "Identifying critical variables of principal components for unsupervised feature selection," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 2, pp. 339–344, Apr. 2005.

[5] S. Y. M. Shi and P. N. Suganthan, "Unsupervised similarity-based feature selection using heuristic Hopfield neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2003, vol. 3, pp. 20–24.

[6] J. G. Dy and C. E. Brodley, "Feature subset selection and order identification for unsupervised learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 247–254.

[7] P. Mitra, C. A. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[8] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering—A filter solution," in *Proc. 2nd IEEE Int. Conf. Data Mining*, 2002, pp. 115–122.

[9] J. Basak, R. K. De, and S. K. Pal, "Unsupervised feature selection using neuro-fuzzy approach," *Pattern Recognit. Lett.*, vol. 19, no. 11, pp. 997–1006, Sep. 1998.

[10] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognit.*, vol. 35, no. 4, pp. 835–846, 2002.

[11] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proc. 9th IEEE Int. Conf. Tools Artif. Intell.*, 1997, pp. 532–539.

[12] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 213–224, Jan. 2005.

[13] D. Huang and T. W. S. Chow, "Effective feature selection scheme using mutual information," *Neurocomputing*, vol. 63, pp. 325–343, Aug. 2004.

[14] Y. Yang, X. Guan, and J. You, "CLOPE: A fast and effective clustering algorithm for transactional data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2002, pp. 682–687.

[15] C. H. Yun, K. T. Chuang, and M. S. Chen, "An efficient clustering algorithm for market basket data based on small large ratios," in *Proc. 25th Annu. Int. Comput. Softw. Appl. Conf.*, 2001, pp. 505–510.

[16] K. Wang, C. Xu, and B. Liu, "Clustering transactions using large items," in *Proc. 8th Int. Conf. Inf. Knowl. Manage.*, 1999, pp. 483–490.

[17] J. C. Bezdek and N. R. Pal, "Some new indexes for cluster validity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 28, no. 3, pp. 301–315, Jun. 1998.

[18] [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[19] [Online]. Available: http://www.genome.wi.mit.edu/mpr/prostate

[20] [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

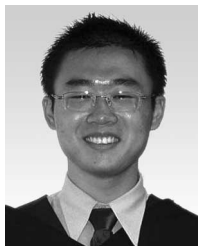
[21] D. Huang and T. W. S. Chow, "Efficiently searching the important input variables using Bayesian discriminant," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 52, no. 4, pp. 785–793, Apr. 2005.

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

Tommy W. S. Chow (M'93–SM'03) received the B.Sc. (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, U.K., where his Ph.D. work was a collaborative project between the International Research and Development, Newcastle Upon Tyne, U.K., and the Ministry of Defense (Navy), London, U.K.

He undertook his training with the Reyrolle Technology, Hebburn, U.K. He then joined the City University of Hong Kong, Kowloon, Hong Kong, where he is currently a Professor with the Department of Electronic Engineering. His main research interests have been in the areas of neural network, learning theory, system identification, and machine fault diagnostics. He is the author or coauthor of numerous published works, including book chapters, and over 90 journal articles related to his research.

Dr. Chow was the Chairman of the Hong Kong Institute of Engineers, Control Automation and Instrumentation Division from 1997 to 1998.



Piyang Wang was an exchange student from Shanghai Jiao Tong University, Shanghai, China. He received the B.E. degree (first honor) from the City University of Hong Kong, Kowloon, Hong Kong, where he is currently working toward the Ph.D. degree in the Department of Electronic Engineering.

His research interest areas are neural networks, pattern recognition, and their applications.

Eden W. M. Ma received the B.E. (first honor) and Ph.D. degrees from the City University of Hong Kong, Kowloon, Hong Kong, in 2001 and 2005, respectively.

Her research interests include pattern recognition and data mining.