JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 951-973 (2023)
DOI: 10.6688/JISE.202307_39(4).0015

Heuristic Feature Selection with Classification Efficiency Using Soft Cluster Analysis for Biological Datasets

HUNG-YI LIN[+] AND RONG-CHANG CHEN
Department of Distribution Management
National Taichung University of Science and Technology
Taichung, 404 Taiwan
E-mail: {linhy; rcchens}@nutc.edu.tw

With a deeper investigation to deciphering the sophisticated relations among input and output variables of multi-class classification problems, the goal of this paper is to propose a new model of variable selection which maximizes the discrimination and minimizes the size of the selected feature subsets. For molecular datasets with a tremendous amount of input variables, the proposed heuristic algorithm is capable of exploring the essential factors of classification problems. Our model devotes to three accomplishments of multi-class classification tasks. Feature discretization using fuzzy clustering analysis for the improvement of feature discrimination is the first. Multivariate analysis for the investigation of information relevance and redundancy is the second achievement in this study. The third is a novel heuristic feature selection algorithm with effectiveness but without overfitting problem. Experimental results convince our model acquires significant discrimination improvement for microarray classification problems.

Keywords: feature discretization, fuzzy c-means, feature selection, feature evaluation, discrimination power

1. INTRODUCTION

The high speed streams and successive growth of high dimensional data together continuously push forward the frontiers of the data volume [16]. For instance, biological data composed of millions or billions of features [24, 29], gazillion of network packets requiring monitoring and classifying for the detection of intruders [33], and the frequent handles of data streams in RFID network and sensor network. In the era of 5G New Radio (5G NR) and artificial intelligence (AI), the massive interconnections of things or objects have an advance announcement that the vastness of IOT information is approaching our life. Having gained omnipresence, tremendous amount of high dimensional data is now pushing the complexities of classification tasks to an even higher level.

In the epochal age of Big data, the characteristics of variety, volume, value, veracity, and velocity frequently occur to high dimensional data [42]. The challenges of high dimensional datasets include data capturing, storage, analysis, search, sharing, transfer, visualization, querying, updating, and so on. Greater and greater computational and analytical tasks are requested for fast various modern applications. Features (or attributes) and instances (or objects) are the basic elements for dataset structure. Feature values are typically expressed by categorical, nominal, or numerical values. Not much ambiguity in categorical and nominal features is worried about. However, the variety of numerical data is much greater due to their continuity or multiplicity, especially when inclusive of high volume data. In seek for verifying data value and promoting data veracity, appropriate preprocess-

Received April 1, 2022; revised August 3 & September 25, 2022; accepted October 9, 2022.
Communicated by Tzung-Pei Hong.
[+] Corresponding author.

ing is usually required and becomes crucial to the success of the subsequent handles. Discretizing real-valued features is one of critical preprocessing issues in classification problems. In addition to data complexity reduction, the concerns of discretization mainly concentrate on the preservation of naïve discrimination capability and data proximity. We will investigate the discretizing effect of soft cluster analysis on single and multiple features in this paper.

It is not feasible for any single variable to ever distinguish multiple classes [7, 21, 25] to their fullest. In general, whenever one class is satisfied, suffering for other classes would in turn happen. Classification problems with multiple classes introduce perplexing interaction among variables, independent or not, thus demanding more efforts in analytical processing. Feature evaluation is the preparation for its selection; however, accurate feature evaluations greatly rely on the preprocessing quality of feature values and even feature vectors. Furthermore, an evaluation criterion should precisely authenticate the discrimination power of features and directs informative features in place. The first goal of this paper is the enhancement of discrimination power for every characterizing feature. Nevertheless, unifying individual decent features do not always lead to the best performance. As a result, the second goal in this paper is to generate a compact subset of features maximizing the discriminative effect for the target decision concept.

"The *m* best features are not the best *m* features" [32] is a famed acknowledgement. More specifically, "the *m* most relevant features are not the most relevant *m* features" could state more explicitly. Relevance analysis in classification problems is to investigate the discriminative capability of individual features to the target class label. In order to promote this capability, a number of distinct selected features are integrated and in turn lead to the phenomenon of redundancy. We note that relevance is the individual property derived from every single feature while redundancy is an ensemble effect originated in a bundle of features. Relevance and redundancy analyses in classification problems are separate issues to be individually discussed. Features holding high relevance to the target class and without serious redundancy in themselves are preferred. Unfortunately, the generation of relevance and redundancy are symbiotic. High relevance usually accompanies huge redundancy.

2. LITERATURE REVIEW AND RELATED WORK

During the past two decades, many literatures and studies [11] dedicate their efforts to balancing the beneficial part of discrimination power and the unfavorable part of data redundancy. In the last decade, many researchers continuously dedicate themselves to improving and promoting feature selection methods. For example, Taguchi Method in Feature Selection (TMFS) [17] overcomes the limitation of MIFS [3]. Quality of Information Feature Selection (QIFS) [23] integrates the concept of maximum-nearest-neighbor into Shannon's information theory. Maximum relevance-minimum multicollinearity (MRmMC) [35] overcomes the problem of correlation characteristics based on conditional variance and achieves redundancy elimination using an orthogonal projection scheme. Dynamic change of selected feature with the class (DCSF) [10] introduced the conditional mutual information between the selected features and the class when considering a candidate feature. The feature selection method based on interaction weight factor and named IWFS is proposed in [40]. IWFS redefined relevance, redundancy and interaction of features in the framework of information theory. The algorithm can deal with irrelevant, redundant and

interactive features. Independent classification information proposed by [38] and its maximization is conducive to achieve a high global discriminative performance. It is a pity that all these studies suffer from a common weakness, in which they assume relevance is repealed simultaneously with redundancy. Integrating relevance and redundancy into one single linear criterion, these studies indicated individual effects are revoked by synthetic effects, which we consider to be faulty. This inappropriate assumption leads to a fatal mistake when searching for next informative features. Possibilistic modeling [6] uses Shapley index paradigm in minimizing the intra-class distance and maximizing the inter-class distance when selecting features. It is a paradigm being able to handle data imperfection or redundancy and is not affected by data variability.

Furthermore, deep learning methods have recently achieved state-of-the-art accuracy for recognition and classification. Convolutional Neural Networks (CNNs) capture both local and global representations in the input samples to learn and reproduce the more important features that helps make better predictions in bioinformatics [2, 41]. Lately, a deep neural network (DNN) has been trained to discriminate between cancer and normal samples using various gene selection strategies [1, 28, 34]. In resolving high dimensional data, deep learning has outstanding performance and becomes an integral component in the outlook for specific patterns within massive datasets. As depicted in [31], firefly search (FFS), elephant search algorithm (ESA), and deep neural network (DNN) were compared when analyzing expression of genes. Clustering with deep learning [26] is proposed to learn a better data representation. We pose the challenges of the current deep learning techniques when conducting microarray gene expression data analysis. First, DNN has at least three hidden layers apart from input and output layer. It constructs the feature hierarchy which combines and aggregates the features from one layer to the next, and it easily increases complexity and level of abstraction. As a result, feature discrimination power tends to be twisted by multilayer perceptron and causes the overfitting problem. Second, DNN needs many hyper-parameters to be set for implementation and finding optimal values for hyper-parameter may involve a great deal of time. Third, a mix of both real and discrete feature values easily results in the infeasible optimal set of hyper-parameters when using gradient descent algorithm. Although DNN is a good choice for handling the very large and high dimensional complex dataset, many problems keep lingering on the feature selection of classification tasks.

This paper is organized as follows: next section the fuzzy c-means algorithm and PBMF-index are sketched. For a convincing argument of applying fuzzy cluster analysis, Section 3 demonstrates the effectiveness of single and multiple features discretization. Mechanisms of authenticating a set of informative features are proposed in Section 4. The novel heuristic feature selection algorithm and complexity analysis are provided in Section 5. The experimental and analytical results are presented in Section 6. Finally, concluding remarks are given in the last section.

3. FEATURE DISCRETIZATION USING FUZZY CLUSTER ANALYSIS

Gene expression microarray analysis relies on many techniques related to data preprocessing, multivariate analysis, statistical analysis, and data mining [18-20]. The raw gene expression values in microarray datasets are continuously challenging machine learning

studies [12, 27]. The values corresponding to quantitative features are likely to have very high data variation or diversity. This likeliness to be overly classified has always been the main issue for classification handles where quantitative features are concerned. As mentioned above, the similarity and dissimilarity messages among these data are more useful than their measured magnitudes in many practical applications. With an aim to warding off complicated numerical analysis, we herein propose the preprocessing of real-valued features with fuzzy cluster analysis and demonstrate the resulting efficacy.

3.1 Fuzzy *c*-Means and PBMF-Index

Fuzzy *c*-means (FCM) developed by Dunn [8] and improved by Bezdek [5] is a well-known soft clustering algorithm which allows one piece of data to belong to two or more clusters. This method is different from hard clustering algorithm which each object strictly belongs to one cluster. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m \| x_i - c_j \|^2, \tag{1}$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by

$$u_{ij} = \left(\sum_{k=1}^{C} \left(\frac{\|x_j - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}\right)^{-1}, \tag{2}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{3}$$

The size of the membership matrix relies on the order of x_i, an oversize matrix may happen when applying to very high dimensional datasets with many clusters. In [14], a proposed technique eliminates the storage of the data structure of membership matrix by combining the two updates into a single update of the cluster centers. This improves the asymptotic runtime from quadratic to linear with respect to the number of clusters. Namely, $O(NC^2)$ is reduced to $O(NC)$, where N objects and C clusters are considered.

FCM approximates and infers the degree of belonging to the cluster, as opposed to only relying on binary choices as well as incomplete or ambiguous data. For highly diverse data, soft clustering algorithm has better discriminative effect than hard clustering. However, two typical clustering questions are frequently addressed: (i) how many clusters are actually present in the data; and (ii) how real or good is the clustering itself. The problem in finding an optimal number of clusters is called the *cluster validity problem* [39]. A number of clustering methods and validation indices [13, 36] have been proposed and successfully employed to solve this problem. The fuzzy version of the PBM-index (abbreviated from the names of the authors), denoted as PBMF-index [30], is employed to verify the quality of FCM cluster analysis in this paper. Briefly elaborated on in the following is the design of PBMF-index.

The PBMF-index is defined as a product of three factors. The product with maximization ensures that the partition has a small number of compact clusters with large separation between at least two clusters. Mathematically, the PBMF-index is defined as follows:

$$V_{PBMF}(K) = (\frac{1}{K} \cdot \frac{E_1}{J_m} \cdot D_K)^2, \tag{4}$$

where K is the number of clusters. The factor E_1 is the sum of the respective distances of each sample to the whole geometric center c_0. This factor does not depend on the number of clusters and is computed as

$$E_1 = \sum_{i=1}^{N} \| x_i - c_0 \|. \tag{5}$$

The factor J_m is the sum of within cluster distances of K clusters, weighted by the corresponding membership value and the same as that in the FCM algorithm. D_K represents the maximum separation of each pair of clusters and computed as

$$D_K = \max_{1 \le i, j \le K} \| c_i - c_j \|. \tag{6}$$

The optimizing of PBMF-index relies on the fewer cluster number, the lower measure of J_m, and the higher estimation of D_K. The calculation procedure is described as follows:

Step 1. Select the maximum number of clusters M_C;
Step 2. Compute the PBMF factor E_1 as Eq. (4)
Step 3. For $K = 2$ to M_C, do
    3.1. Run the FCM algorithm;
    3.2. Compute the PBMF factors J_m and D_K as Eqs. (1) and (5);
    3.3. Compute the $V_{PBMF}(K)$ index as Eq. (3)
Step 4. Select the best number of clusters K^* as:
    $K^* = \arg\max(V_{PBMF}(K))$

When it comes to dealing with multi-class classification problems, FCM algorithm with cluster validity PBMF-index brings many benefits. Feature discretization comes first especially when faced with ambiguous, unknown, or incomplete data. Additionally, fuzzy clustering allows a single feature value or a feature vector to belong to more than a single cluster and it promotes the identification of features that are conditionally co-regulated. Namely, one feature or subset of features may be acted on and concurrently affect or determine more than one specific class. This is why fuzzy clustering is herein preferred than hard clustering. Moreover, PBMF-index searches for the adequate and refined cluster number for single features and feature combinations, which tackles the subjective judgement and avoids the over-fitting problem possibly generated from a high cluster number. Although such discretization indeed requires computational costs before progressing feature selection, its last edge is that it explicitly reduces data complexity and also preserves the inherent information of data proximity or difference, which boosts all the computational and analytical handles in terms of speed in the subsequent procedures of feature selection. We will verify these benefits in the next subsections.

3.2 Single Feature Discretization

The motivation of such preprocessing is to probe all features in order to filter out a reduced collection of informative features from a massive population. Attribute values corresponding to every single feature are first discretized into distinct categorizations, and then the discrimination powers of discretized features are evaluated. Most important of all, feature discretization is expected to retain the original characteristics of data distribution and data proximity so that discrimination power can be preserved. Note the rareness of studies in the precedent researches related to the feature discretization without sacrificing discrimination power. In machine learning, *discretization* refers to the process of converting or partitioning quantitative features to discretized or nominal ones. This serves as a useful method when formulating mass functions of probability. Typically, data is discretized into partitions of h equal lengths/width (equal intervals) or h% of the total data (equal frequencies). Several machine learning algorithms [15, 22, 36] are renowned for yielding better models by discretizing quantitative features. Discrimination power holds predominant magnitude to all other factors as far as classification is concerned. Well discretizing techniques are expected to levitate the power of discrimination. By applying fuzzy c-means on every single dimension, the primary issue herein is to enhance the discrimination power of all quantitative features.

The tumor dataset downloaded from University of Wisconsin [47] is taken to illustrate our design of feature discretization. There are 569 digitized images (357 benign plus 212 malignant) and 30 features are collected from each image. Real-valued features are computed for each cell nucleus without missing value. For being explicit and concentrative on illustrating our designs, 50 images are randomly extracted and 25 features are employed in the following experiments. As shown in the first row of Table 1, the mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image. For instance, f_1 is radius_mean, f_2 is texture_mean, and f_{25} is fractal_dimension_worst (excluding perimeter_mean, concave points_mean, radius_worst, perimeter_worst, concave points_worst). The second row is the resulting cluster number for each feature using PBMF-index. For a thorough study, the soft cluster analysis-FCM and the hard clustering method-expectation maximum (EM) are both employed over these 25 features. Then, two criteria: information gain (*IG*) and gain ratio (*GR*) are used for discrimination power evaluation.

Table 1. Comparison of feature evaluation based on FCM and EM methods.

Feature ID		f_1 f_2 f_3 f_4 f_5	f_6 f_7 f_8 f_9 f_{10}	f_{11} f_{12} f_{13} f_{14} f_{15}	f_{16} f_{17} f_{18} f_{19} f_{20}	f_{21} f_{22} f_{23} f_{24} f_{25}	average
Cluster number		2 2 3 2 4	3 5 5 5 6	8 3 4 3 7	5 4 3 2 4	5 4 3 3 3	
FCM	*IG*	.54 .25 .50 .30 .32	.39 .12 .06 .31 .01	.39 .42 .09 .12 .12	.18 .31 .07 .38 .54	.03 .10 .38 .17 .06	0.235
	GR	.59 .14 .55 .20 .14	.31 .06 .04 .26 .01	.20 .26 .07 .06 .10	.12 .11 .03 .18 .45	.02 .16 .17 .07 .03	0.165
EM	*IG*	.60 .17 .60 .02 .26	.33 .02 .07 .30 .06	.36 .40 .03 .16 .12	.07 .24 .07 .27 .55	.10 .10 .30 .17 .03	0.215
	GR	.63 .12 .63 .01 .11	.28 .01 .06 .20 .04	.18 .22 .02 .09 .10	.06 .08 .33 .13 .46	.10 .16 .14 .07 .02	0.158

Features f_1, f_3 and f_{20} (marked with shadows and red fond) are consistently evaluated as the top three informative features whether FCM or EM algorithms are employed. Since the data complexity contained in every single feature is distinct and diverse, no clustering method guarantees its maximal efficacy for all features. We note that there are alternative leadings for individual features between FCM and EM methods. However, the averaged *IG* yield of FCM collectively outperforms that of EM by 9.3%, thus verifying the capability

of fuzzy clustering algorithm. For further illustration, the three leading features will be sown as seeds for further searching. Since the average *GR* yields do not appear obvious difference (4.4%), this part is omitted in the subsequent experiments of this example.

3.3 Multiple Features Discretization

Based on one of the selected features (f_1, f_3 or f_{20}), the second feature is investigated through FCM and PBMF-index. The discretization handles were applied on the feature vectors. Table 2 depicts the selected results of paired features with preferable IG performance. The parts with stay and worse *IG* values are not listed in Table 2. We note that there are different numbers of second collaborators explored based on f_1, f_3 and f_{20}. Among these all combinations, (f_{20}, f_{25}) owns the best *IG* result and improve the first selected feature with the best discrimination power. The Spearman's rank correlation coefficients (ρ) between the paired features were also calculated in the fourth column of Table 2. Worth mentioning, the correlations all stay at a very low level so that resulting selected feature subsets are free of redundant information.

Table 2. First selection round based on f_1, f_3 and f_{20}.

Pair	Cluster number	*IG* via FCM	ρ
(f_1, f_{20})	6	0.574	0.198
(f_3, f_{10})	3	0.522	0.120
(f_3, f_{12})	4	0.522	0.120
(f_3, f_{16})	2	0.522	0.120
(f_{20}, f_{16})	6	0.574	0.162
(f_{20}, f_{25})	10	0.660	0.237

Table 3. Second selection round based on Table 2.

Triplet	Cluster number	*IG* via FCM	ρ
(f_1, f_{20}, f_2)	8	0.610	0.092
(f_1, f_{20}, f_{23})	7	0.679	0.068
(f_{20}, f_{16}, f_{15})	8	0.610	0.315
(f_{20}, f_{16}, f_{21})	7	0.679	0.267
(f_3, f_{12}, f_{16})	2	0.562	0.063
(f_3, f_{12}, f_{18})	2	0.562	0.063
(f_3, f_{12}, f_{24})	2	0.562	0.063

To gain more discrimination power from other unselected features, the six paired features of Table 2 were matched up with other features and re-discretized by FCM and validated by PBMF-index. As shown in Table 3, only (f_1, f_{20}), (f_{20}, f_{16}) and (f_3, f_{12}) were capable of obtaining further IG improvement. The fourth column becomes to average the Spearman's rank correlation coefficients of each feature pairs in every triplet item. The *IG* yields from subsets (f_1, f_{20}, f_{23}) and (f_{20}, f_{16}, f_{21}) outperform others and (f_1, f_{20}, f_{23}) have a very low information redundancy inside its combination.

4. APPROACH FOR APPROVING INFORMATIVE FEATURES

Relevance is usually characterized in terms of mutual information [4, 9, 37] or correlation, of which the former is one of the widely used measures to define dependency of variables and the latter is the most popular statistical relationship between two variables. To compare their affection to feature selection, we focus on two mutual information based and one statistical method. We briefly review three methods for feature evaluation and explicitly explain how they can work with fuzzy cluster analysis for boosting feature discrimination power. Suppose features *A* and *B* respectively have *n* and *m* distinct feature values, *i.e.*, a_i

and b_j, where $1 \le i \le n$ and $1 \le j \le m$. The dataset with a target class label C classified by feature A will result in n subsets. In other words, $C = C_{A=a_1} = C_{A=a_2} \cup \ldots \cup C_{A=a_n}$. Similarly, C classified by feature B results in m subsets, *i.e.*, $C = C_{B=b_1} = C_{B=b_2} \cup \ldots \cup C_{B=b_m}$. The information entropy from C is denoted as $H(C)$. The information entropy from C classified by A is denoted as $H(C|A)$ and formulated as:

$$H(C \mid A) = \sum_{i=1}^{n} \frac{|C_{A=a_i}|}{|C|} \times H(C_{A=a_i}). \tag{7}$$

We assume feature A is already in use and then feature B is going to join the classification. The corresponding information entropy when A and B are taken will become

$$H(C \mid A \cup B) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|C_{A=a_i, B=b_j}|}{|C|} \times H(C_{A=a_i, B=b_j}). \tag{8}$$

We note information gains $IG(C; A)$ and $IG(C; A \cup B)$ are respectively calculated by $H(C) - H(C|A)$ and $H(C) - H(C|A \cup B)$. Hence, the improvement of relevance derived from feature B can be formulated as following deduction:

$$\begin{aligned} IG(C; A \cup B) - IG(C; A) &= H(C) - H(C|A \cup B) - [H(C) - H(C|A)] \\ &= H(C|A) - H(C|A \cup B). \end{aligned} \tag{9}$$

For simplicity, such measurement is denoted as $\Delta IG(B|A)$. Extensively, in the case of a collection of features (denoted as S) has already been taken, the relevance improvement caused by a newly added feature α can be measured by $IG(C; FCM(S \cup \alpha)) - IG(C; FCM(S))$, where $S \cup \alpha$ and S are respectively discretized into categorical data by fuzzy c-means. As a result, the higher $\Delta IG(\alpha|S)$ boosts the discrimination power of $S \cup \alpha$ to a further level and validates the necessity of α.

Secondly, if multi-class datasets are dealt with, information gain ratio could be another choice.

$$GR(C; A) = H(C|A) / H(A), \tag{10}$$

where $H(A)$ is the *intrinsic information entropy* from A. Hence, $GR(C|A \cup B)$ measures the normalized information gain from A unifying B. Similarly, $\Delta GR(B|A)$ measures the difference of information gain ratio between $GR(C; A \cup B)$ and $GR(C; A)$, and $\Delta GR(\alpha|S)$ becomes another criterion when validating the necessity of α. That is,

$$\Delta GR(\alpha|S) = GR(C; FCM(S \cup \alpha)) - GR(C; FCM(S)). \tag{11}$$

Thirdly, chi-squared test (χ^2-test) investigates statistical relationships between two variables. A χ^2-test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. In resemblance with previous methods, $\Delta\chi^2(B|A)$ measures the additive dependency from feature B (*i.e.*, $\chi^2(C; A \cup B) - \chi^2(C; A)$), and $\Delta\chi^2(\alpha|S)$ becomes our third criterion when validating the additive dependency caused by the α. The last is formulated as:

$$\Delta\chi^2(\alpha|S) = \chi^2(C; FCM(S\cup\alpha)) - \chi^2(C; FCM(S)). \quad (12)$$

In fact, the additive effect of the new variable α is twofold. The proposed three criteria focus on the detection of distinct relevance information to the target class variable. On the other hand, the join of a new variable in a system cannot be rid of redundant information. Many past studies try to balance the measurements of relevance and redundancy information. For example, MIFS algorithm [3], MIFS-U algorithm [17], mRMR criterion [32] and NMIFS algorithm [11] are respectively formulated as following

$$I(C; f_i) - \beta\sum_{f_s\in S} I(f_s;f_i), \quad (13)$$

$$I(C; f_i) - \beta\sum_{f_s\in S} I(f_s; f_i), \quad (14)$$

$$I(C; f_i) - \frac{1}{|S|}\sum_{f_s\in S} I(f_s; f_i), \quad (15)$$

$$I(C; f_i) - \frac{1}{|S|}\sum_{f_s\in S} NI(f_s; f_i). \quad (16)$$

Unfortunately, integrating the relevance of the feature to be added (*i.e.*, $I(C; f_i)$) and the redundancy of the *i*th feature with respect to the subset previously selected features (*i.e.*, $\sum I(f_s; f_i)$ in a linear formulation is problematic in two aspects. The first aspect relates to the consistency of the measured information. Although the asymmetric selection weight between the left side and right side in Eqs. (13) and (14) is solved by dividing the sum with the cardinality of the set *S* proposed in Eqs. (15) and (16). Their common problem is that they trade the profit of relevant information off against the risk of redundant information. The second aspect relates to arithmetic problem. Even though the designs of $I(C; f_i)$ and $\sum I(f_s; f_i)$ are both based on entropy theory, $I(C; f_i)$ is focusing on the target class label while $\sum I(f_s; f_i)$ is on various input features. Different referred target information may cause different quantitative merits and scales. The assumption of compensatory relation between them is highly risky and it is quite inappropriate to integrate them in a linear formulation.

Generally, features are selected one-by-one so that selection criteria should particularly regulate the supplementary effect led by the upcoming feature rather than focus on the whole effect brought by the already selected features plus the new one. In this study, the success of approving α's enhanced effectiveness for *S* relies on two factors. One is the greater $\Delta GR(\alpha|S)$, or $\Delta\chi^2(\alpha|S)$. As to redundancy information, since the collected variables in a dataset are derived from a specific theme, it is quite hard to avoid correlation between variables (*i.e.*, features). Such situation has driven us to take only the level of the redundancy into account rather than their definite magnitude of redundancy. As schematic in Fig. 1, the low and medium overlaps of redundancy as depicted in the grey area of the second and third panel are preferred. The high overlap of the fourth case is unfavorable. Redundancy with 0% and 100% overlaps are rare. The percentage of redundancy level is proposed to illustrate a general concept, as the exact value depends on the practical application.

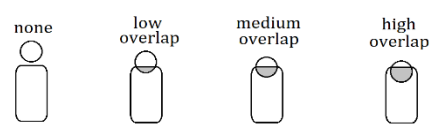


Fig. 1. Redundancy conditions between α and *S*.

In order to precisely define the three redundancy levels, we regulate a simple rating method. As far as the relevance criterion of $\Delta IG(\alpha|S)$ is concerned, the arithmetic mean of mutual information between α and S is denoted as $\overline{MI}$ and formulated as $(1/|S|)\sum_{f_s \in S} I(f_s; \alpha)$. After having measured $\overline{MI}$, a scale is required to determine the degree of data redundancy. $IG(C; FCM(S \cup \alpha))$ herein serves as the scale. According to the magnitude of $\overline{MI}$ and $IG(C; FCM(S \cup \alpha))$ we have observed from our past experiment data [19], values of ratio of $\overline{MI}$ to $IG(C; FCM(S \cup \alpha))$ below 2/3 are regarded as low, as values of ratio of $\overline{MI}$ to $IG(C; FCM(S \cup \alpha))$ over 4/3 are regarded as high. We note that the values of 2/3 and 4/3 are not necessarily suitable for all data, and the number of levels can also be modified dynamically according to the study of interest. As a result, the following three levels, L_i, $i \in \{1, 2, 3\}$, are regulated:

- L_1: $\overline{MI} \le (2/3) \cdot IG(C; FCM(S \cup \alpha))$
- L_2: $(2/3) \cdot IG(C; FCM(S \cup \alpha) < \overline{MI} \le (4/3) \cdot IG(C; FCM(S \cup \alpha))$
- L_3: $\overline{MI} > (4/3) \cdot IG(C; FCM(S \cup \alpha))$

The other two cases $\Delta IG(\alpha|S)$ and $\Delta\chi^2(\alpha|S)$ also adopt similar rating method. The novelty in this study is to moderately modulate supplement and redundancy effects. Hence, one feature is going to be classified as informative if it is capable to maximize its *incremental relevance* to the target class variable and minimize its *redundancy level* to the existent features. We will explain the detailed design in the next section.

5. ALGORITHM AND COMPLEXITY ANALYSES

5.1 Heuristic Feature Selection Algorithm

We propose a two-stage selection process. At the first stage, relevant and irrelevant features are primarily distinguished upon the chosen evaluation criteria (*i.e.*, information gain, gain ratio, and chi-square). Features with better evaluations are collected into the candidate set for the use in the subsequent selection process. Then, the heuristic learning process with the examination of discrimination information is preceded on the candidate features during the second stage. The notations used in our heuristic learning algorithm are initialized as follows:

- U: The set contains all raw features.
- U_c: The set contains the candidate features. The cardinality of U_c is less than U and $|U_c| < |U|$.
- A_1: The set contains those single features which are classified as most characterizing.
- A_1, A_3: The sets contain the paired and tripled features, respectively.
- A_i: The set contains the combinatorial features which are hybrid from i features.

Input: the raw attribute values corresponding to all indigenous features and the target class variable. U and C indicate the raw feature set and the target class variable.
Output: a variety of feature subsets with high discrimination power

1. Discretize every feature in U using FCM and PBMF-index. The discretized features are stored for the handles of subsequent selection and afterward classification works.
2. According to the designated criterion ψ, evaluate all discretized features and sort them in a decreasing order. The feature with the highest evaluation is denoted as a_1 and so forth.

3. Based on a given threshold percentage τ, the features ranking at the front of the sorted U are drew out and collected into U_c. Fetch a few of front features of U_c into A_1.
4. For each element f in A_1,
5. $S \leftarrow f$;
6. For each element α in $U_c - S$,
7. Fetch the native feature values corresponding to $S\&\alpha$ and discretize feature vector $S \cup \alpha$ using FCM and PBMF-index.
8. If $\Delta\psi(\alpha|S) > 0$ and (L_1 or L_2 is probed), then collect α into B.
End for
9. $A_2 \leftarrow S \cup \{\text{argmax}\Delta\psi(\alpha|S)\}$
End for
10. For $i = 2$,
11. $A_1 \leftarrow A_i$; $A_1 \leftarrow A_{i+1}$
12. Repeat Steps 4 to 8 until positive $\Delta\psi(\alpha|S)$ no longer be found.
13. $i = i + 1$
End for
14. Return A_i, $i \geq 1$.

For the better discrimination, all features in U are initially discretized with the soft cluster analysis. Afterwards, the feature evaluation criterion ψ of Step 2 has three choices: information gain, gain ratio, and chi-square test. At Step 3, a threshold percentage τ (for example, $|U_c| = 10\% \times |U|$) is designated for qualifying the relevant features. A given feature amount (for example, $|U_c| = 100$) is an alternative way for this goal. In addition, our algorithm launches multiple search paths starting up with a few of features. These features included in A_1 are the most characterizing and ranking at the top of U_c. From Steps 4-9, heuristic selection mechanism is proposed for exploring the next informative features based on every element of A_1. Hence, $|A_1|$ searching paths are launched due to this setting. Step 7 discretizes the combined features. Step 8 verifies the effectiveness of the test feature according to its supplement effect and dependency rate. Measurements of boosting discrimination and dependency degree will sift a number of informative features into B. Then, the feature α in B which is most contributive to S(*i.e.*, $\arg\max_{\alpha \in B} \Delta\psi(\alpha \mid S)$) will include $S \cup \alpha$ into A_2 as executed by Step 9. For the sake of plentiful results, we can loosen the restriction and consider more α's with positive $\Delta\psi(\alpha|S)$ instead of permitting a unique one with maximal $\Delta\psi(\alpha|S)$.

Activating further heuristic selection rounds as described from Steps 10 to 13 can continuously explore informative feature combinations with longer length. However, more features bundled together incur higher redundancy and greater noisy and in turn gradually accumulates large dependency. Eventually, the positive $\Delta\psi(\beta|S)$ is hard to be explored and this situation terminates the further selection rounds. Step 14 returns the various informative feature subsets gathered in A_i's. Features selected in distinct subsets will be used to train different classifiers and their resulting classification performance will be discussed in Section 5.

5.2 Data and Computational Complexity

We assume that the readers are familiar with FCM cluster analysis validated by PBMF-index as mentioned in Section 3. For complexity analyses in our proposed handles, the total instance and feature amount of one genetic dataset is represented as M and N. The implementation times of PBMF-index for validating the cluster quality is denoted as K and the

generated cluster number when discretizing one feature is C. The number of rounds executed in feature selection is designated as R.

Three processing steps including feature discretization, feature evaluation, and feature selection are respectively highlighted as follows.

(I) Feature discretization (FD): excluding the target class label, real-valued or high variety features are preprocessed by fuzzy c-means algorithm validated by PBMF-index. Nominal data are exempted from this processing. Fuzzy clustering method can group raw data into a non-uniform categorization and assign every cluster a sequencing number. In this step, all features are equally respected without any selective priority. The output of this step is the discrete values corresponding to the processed features. The conventional FCM algorithm taking a cluster number of C is applying on this kind of dataset. Then, the data complexity involved in FD processing is $O((MC)N)$, where $N > M \gg C \geq 2$. Regarding computational complexity, $O((MC^2)NK)$ is required in this processing, where the conventional FCM algorithm is applied for K times validations with PBMF-index. A reduced complexity of $O(MCNK)$ can be achieved when employing the improved FCM [14].

(II) Feature evaluation (FE): All discretized features are separately evaluated by criteria information gain, gain ratio, or chi-square. Every single input feature has to pair with the target class label so that the data amount involved in this processing is $(N + N)M$. This evaluation phase needs at most an asymptotic runtime of $O(N)$.

(III) Feature selection: the heuristic matching process in our algorithm necessitates advanced processing of FD and FE. Fortunately, only a limited number of candidate features are involved in this stage. As stated in Section 5.1, we suggest $10\% \cdot N$ in this paper since most discretized features have minimal IG values and only a few features are evaluated with significant IGs. In respect of FD, discretization of multiple features requires a computational complexity of $10\% \cdot N \cdot O(MC^2K)$ using the conventional FCM algorithm for C clusters and validating by K times PBMF-index. And then, only these discretized feature vectors together with the target label have to be handled by FE. Consequently, the entire heuristic selection needs an execution complexity of $((MC^2)NK) + 10\% \cdot N \cdot O(MC^2K) \cdot R$, where round number R is upon the designated restriction. How the number of selected features affects the classification performance will be investigated in the next section.

The overall data complexity and computation complexity in the execution of our selection scheme are bounded by the processing of feature discretization, *i.e.*, $O(MCN)$ and $O((MC^2)NK)$, which are comparable with that of many other feature selection algorithms. The total computation complexity retains the same economic level as those algorithms classified into the *filter* category.

6. EXPERIMENTAL RESULTS AND ANALYSES

6.1 Dataset Acquisition

Ten genetic datasets, including Lung cancer, Breast cancer, pharmaceutical data (Novartis), and so on, are downloaded from the biomedical and genomic research center such

as Broad Institute, Arizona State University, Knowledge Extraction based on Evolutionary Learning (KEEL), and UCI Machine Learning Repository [43-46]. Except for the last dataset containing integer-valued data, all the other nine datasets are continuous-valued data at different magnitudes. The abstract of these datasets is listed in Table 4.

Table 4. Abstract of ten datasets.

Datasets	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10
# of features	1000	1213	1000	7129	11340	4434	5748	19993	9182	34
data type	continuous									integer
# of samples	197	98	103	72	111	50	171	187	174	366
# of classes	4	3	4	2	3	4	4	2	11	6

For simplicity, ten datasets are sequentially abbreviated from DS1 to DS10. Feature amounts vary from 34 to 19993, and number of target classes varies from 2 to 11. In order to verify the selection effectiveness of our proposed schemes, most datasets we used in this paper have the significantly plentiful feature numbers higher than their sample amounts. Only DS10 has lower feature quantity than its sample number. The data type of all raw features in the first nine datasets, except DS10, is continuous. No nominal features are involved in these ten datasets. Quantitative features with continuous values in DS1~DS9 datasets were discretized by FCM, and the resulting cluster numbers were validated by PBMF-index. As to the discrete features of DS10, such analytical procedure is still applied to attain fewer discretized feature values for the sake of a reduced number of categorizations. However, fuzzy cluster analyses could affect nothing to those features with a small number of discrete values and they retain the original values. For comparison studies, all features corresponding to the 10 datasets are discretized and saved into 10 individual files so that they can be reused for training distinct classifiers. All data preprocessing, feature discretizing, heuristic selections, and classifiers training were implemented in R programming languages executed on a workstation with an AMD Athlon dual core 2.59 GHz processor. To verify our design, four classification methods including C4.5, SVM, NaiveBayes (NB), and k-nearest neighbor (kNN) were used in the comparison experiments. Algorithms C4.5, SVM, and kNN are non-probabilistic while NB depend on the precise nature of the probability model. We take k = 1, 2, and 3 for kNN algorithm and average the individual outcomes.

6.2 Discrimination Power Studies of Selected Features

Because our experiments collect several multi-class datasets, the discrimination power evaluated by gain ratio are taken for investigation. Gain ratio takes class number and the factor of population information amount into account. In order to compare the discretization effects from different clustering methods, three discretization algorithms, FCM, expectation maximum (EM) and K-means (KM), are applied. After one of the discretization algorithms (FCM, EM, or KM) had been applied, features in a given dataset with the best gain ratio values are collected as a "testbed". For DS1~DS9, their respective first 100 features were collected as testbeds. Due to a lower number of features in DS10, all 34 of its features were collected as its testbed. For each testbed, we randomly selected 3 features and averaged their gain ratio. The process was repeated for 20 times to achieve a collective average.

For a more comprehensive observation, random selections in higher numbers (5 and 9) have also been made. Experiments with 3, 5 and 9 features selected were respectively shown

in Figs. 2-4. As shown in Fig. 2, FCM applied in the five datasets including DS1, DS2, DS4, DS7 and DS8 outperforms EM and KM. Obviously, the gain ratios of DS3 and DS5~DS9 using the three methods stay at a level below 0.4. DS8 and DS9 are even lower than 0.1. In fact, similar results can be found in Figs. 3 and 4. Based on the experimental results so far, we infer that the discriminate information stored in DS3 and DS5~DS9 is poor and insufficient. Thus, these datasets are not worthy of further analyses, much less improving classification performance. As to DS10, a high evaluation around 0.8 appears in all three discretization methods. It explicitly reveals that the downloaded data of DS10 possesses sufficient discriminate information. As we pay a higher attention to comparing DS1, DS2, DS4, and DS10, the continuous-valued features in three datasets (*i.e.*, DS1, DS2, and DS4) are successfully boosted by FCM as compared to EM and KM. DS10 shows a tinier difference among the three methods. This should be due to the fact that discretization procedures cannot improve the discrimination power of discrete features. The following figures 3 and 4 will support this declare.

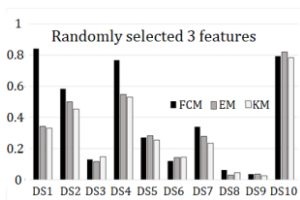


Fig. 2. Three features selected for *GR*s.

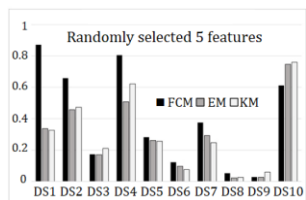


Fig. 3. Five features selected for *GR*s.

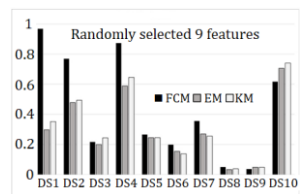


Fig. 4. Nine features selected for *GR*s.

Similar to Fig. 2, Fig. 3 reveals that DS1, DS2, DS4, and DS10 have better evaluations in all three discretization methods. We particularly note that the FCM promotes the gain ratio exceeding 0.6 and even achieving 0.8 in DS1, DS2, and DS4. Besides ensuring better discriminate information, it also reveals that the five selected features boost discrimination power in DS1, DS2, and DS4. As we focus on continuous-valued features in DS1, DS2, and DS4, the FCM still outperforms EM and KM and gains an improvement of 11% as compared to Fig. 2. However, as far as the discrete features of DS10 are concerned, more features cause lower values when applying these discretization methods. This situation is worrisome since the relevance has decreased.

When nine selected features are taken into account, as shown in Fig. 4, the FCM again outperforms EM and KM except DS3 and DS10. The FCM continuously promotes the gain ratios of DS1, DS2, and DS4. Successfully, the FCM pushes the gain ratio of DS1 to a near-optimal level (gain ratio = 0.969). As a summary based on the observation of Figs. 2-4, we conclude that DS1, DS2, DS4, and DS10 possess better data quality of discriminate information than other datasets. Hence, only these four datasets plus DS3 (as a negative comparator) are taken to precede the accuracy studies in Section 6.4.

6.3. Quality Studies of Selected Features

To verify the effectiveness of our proposed feature selection algorithm, we compare our method based on FCM-discretization (abbreviated as FSFCM) with other heuristic algorithms (MIFS and MIFS-U). As mentioned above, relevance and redundancy are two

pivotal factors affecting the quality of selection. Hence, we focus on these two aspects in the following experiment. Gradual evolution with ten rounds of selections (R1~R10) were tested for three heuristic algorithms. First, as to relevance analysis, we observed the change of *IG* values from R1 to R10. On the other hand, the changes of the average *MI*'s (*i.e.*, $\overline{MI}$) were also monitored for redundancy analyses. One binary class dataset (DS4) and three multi-class datasets (DS1, DS5 and DS7) were taken in this experiment.

Fig. 5 demonstrates that DS4 with binary classes gains the significant relevance improvement when using FSFCM, especially during the first 6 rounds. Although the relevance improvement becomes insignificant after the 6th round, FSFCM explicitly outperforms MIFS and MIFS-U over 10 rounds. As to redundancy, MIFS are able to constrain $\overline{MI}$ within a low level below 0.2 while MIFS-U is unable to inhibit the dilation of $\overline{MI}$. Comparing FSFCM and MIFS, FSFCM was found to generate more redundancy than MIFS, but such weakness does not outweigh the strength of FSFCM in relevance. Overall, FSFCM are able to maintain "high" relevance and retain "low" redundancy.

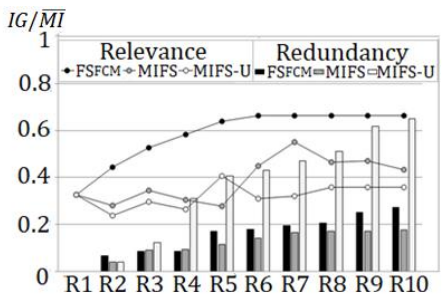
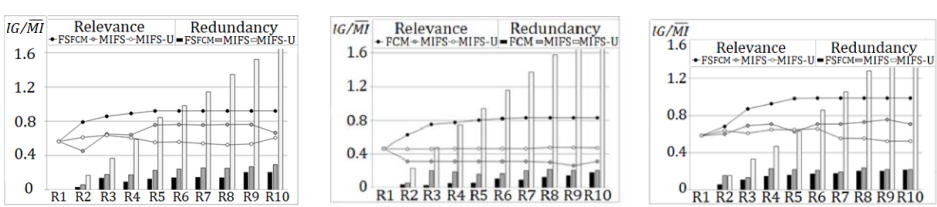


Fig. 5. Feature quality analyses for DS4.

As far as multi-class datasets are concerned, we have several common observations. The first is that an obvious increase of relevance is found in the first five rounds of FSFCM. However, MIFS and MIFS-U failed to push up *IG*s at a steady pace, sometimes even found with decays. The second is that FSFCM and MIFS are able to regulate the redundant effects among the selected features, which MIFS-U is incapable of.

IG and $\overline{MI}$ refer to different target information and then may cause different quantitative merits and scales. The experimental results provided the third observation to verify the inappropriate assumption of compensatory relation between them. We note that the evolutionary trends of *IG* and $\overline{MI}$ are inconsistent. So, it is quite improper to integrate them in one single formulation as proposed by MIFS or MIFS-U. The methods proposed in the previous studies revoked feature redundancy at the cost of relevance.



(a) DS1 (b) DS5 (c) DS7

Fig. 6. Feature quality analyses for three multi-class datasets.

Considering classification tasks, the importance of relevance is higher than redundancy since prediction accuracy is of top concern. Meanwhile, taking computational efficiency into account, reduction of redundancy is necessary for a feature selection algorithm. As far as these four datasets are concerned, FSFCM really fulfills our motivations.

6.4 Accuracy Studies of Selected Features

Let us look at the performance of classifiers built with the features selected in different situations. For simplicity, S1, S2, and S3 respectively stand for the non-heuristic algorithm, the heuristic algorithm without redundancy checking, and the heuristic algorithm with redundancy checking. S1 collects those discretized features whose relevancies to the target class label have high assessment without processing heuristic selection. S2 heuristically selects those discretized features with high relevance assessment but without processing redundancy checking and S3 truly conducts redundancy checking. The first classifier adopted in this study is decision tree C4.5. The reason why CART (classification and regression tree) is not considered is that all continuous-valued features are discretized. SVM is the second classifier taken which is capable of building non-linear decisive boundary. NaiveBayes (NB) assumes strong independence between the features so that a set of low-dependent features will exonerate NB classifier from ineffectiveness. The *k*-nearest neighbors (kNN) algorithm is a non-parametric and supervised learning classifier in which new instances can be joined to the dataset without remodeling. Suppose E and F are the number of training examples and the numbers of involved features, the time complexity of four classifiers are $O(EF^2)$, $O(E^2)$, $O(EF)$, and $O(EF)$, respectively. All experimental results in this study were assessed using 10-fold cross-validation. Only four datasets (DS1~DS4) are carefully analyzed in this part. Figs. 7-21 respectively depict the classification accuracy (*y*-axis) using C4.5, SVM, NB, and kNN classifiers from DS1 to DS4. The number of selected features is labeled in the *x*-axis. As shown in Fig. 7, S2 and S3 have the same results at all cases excluding C4.5. This means data redundancy is not serious among the features selected by *IG* so that it cannot bring the design of S3 into full play. We find S2 and S3 have gained the satisfactory results at earlier rounds than S1 when using NB and kNN classifiers. In case of using *IG*, NB is voted as the best classifier for DS1. Overall, FCM have efficiently improved their own discrimination powers so that the average accurate rates for all cases go beyond 85%. Furthermore, a high rate exceeding 92% is detected for SVM, NB, and kNN since 3 features are involved in their training processes.

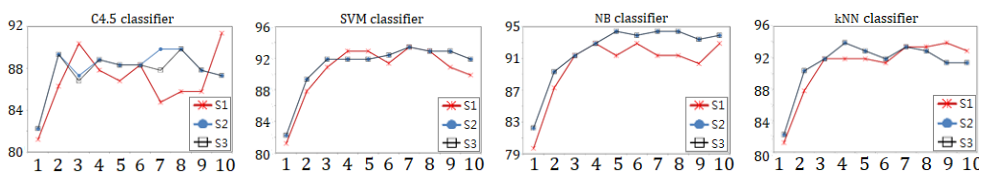
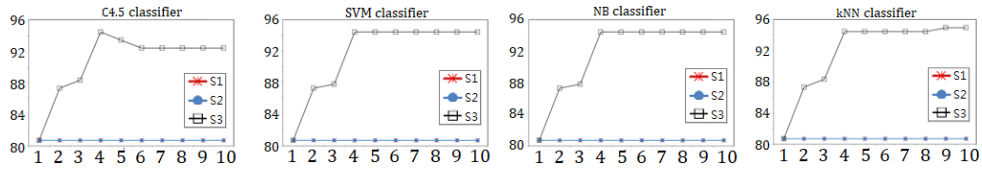


Fig. 7. Classification accuracies (%) for DS1 based on *IG*.

Completely dissimilar to *IG*, the assessment of redundancy seems to be apparent since S2 and S3 have the distinct experimental results. As shown in Fig. 8, S1 and S2 have the same accuracy of 80.71% in all cases. However, S3 keeps promoting classification accuracy

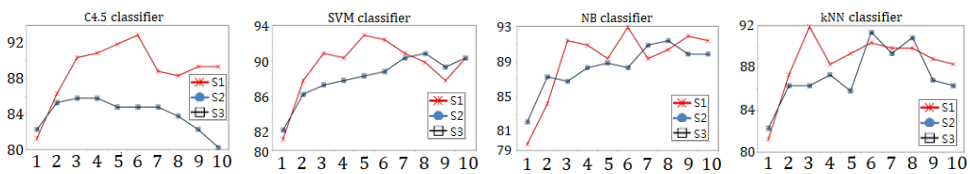
until the 4th feature is selected. As compared with *IG*, FCM cooperated with *GR* and our proposed heuristic selection algorithm successfully pushes the classification accuracy of DS1 to a high rate over 94% at very early stage.



(a) (b) (c) (d)

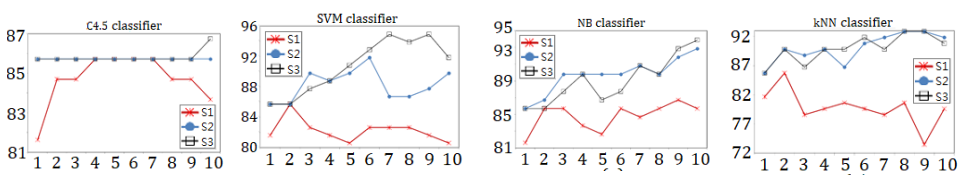
Fig. 8. Classification accuracies (%) for DS1 based on *GR*.

Chi-squared test determines whether there is a significant difference between the expected frequencies and the observed frequencies. Relevancy analysis via χ^2 test has the weakness of ignoring data pair correspondence. As depicted in Fig. 9, the advantages and disadvantages of implementing S1, S2, and S3 are ambiguous. As a short summary from Figs. 7-9, FCM cooperated with SVM and NB could make our proposed selection algorithm generate the stable performance. Hence, we conclude that feature discretized by FCM, evaluated by *GR*, selected by our method, then trained by NB acquires the high classification performance of DS1.



(a) (b) (c) (d)

Fig. 9. Classification accuracies (%) for DS1 based on χ^2.



(a) (b) (c) (d)

Fig. 10. Classification accuracies (%) for DS2 based on *IG*.

Figs. 10-12 show the experimental results based on DS2. With respects to *IG* and *GR*, S2 and S3 explicitly have the superior performances over S1. As to the parts of χ^2, Fig. 12 shows S3 performs better than S2 which is still better than S1 during the 3rd, 4th, 5th, and 6th rounds. From each panel in Fig. 12, the best performance happens around the 4th, 5th, or 6th rounds. We note that the strict mechanism of S3 terminates the selecting process after the 7th round. The best accuracy of 94.90% happens to Fig. 10 (b) where the features were discretized by FCM, evaluated by *IG*, then selected by our proposed method, and finally trained by SVM in the 7th and 9th rounds. To verify the effectiveness of our method, we

take an overall observation over these experimental results and obtain the averaged accuracies of S1, S2, and S3 as 84.37%, 87.62%, and 88.65%.

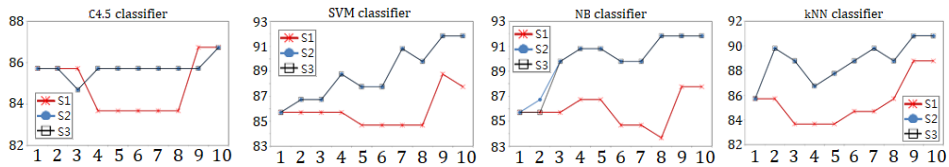


Fig. 11. Classification accuracies (%) for DS2 based on *GR*.

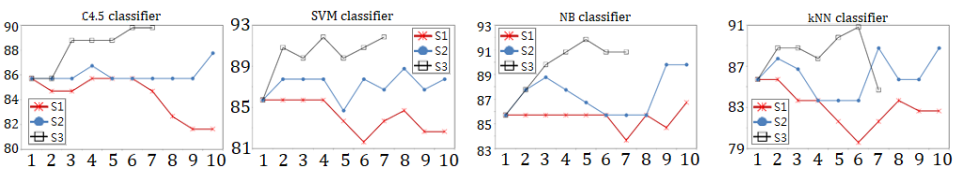


Fig. 12. Classification accuracies (%) for DS2 based on χ^2.

As far as DS3 is concerned, it possesses poorer gain ratio evaluation less than 0.2 as shown in Figs. 2-4. This reveals DS3 contains insufficient classification information. More selected features only cause higher redundant situation and could not benefit classification performance. Obviously, three selection strategies applied over DS3 have different experimental results. Although the best accuracy approximated to 80% as *GR* criterion and NB classifier were applied in S3 during the 10th round, most experimental results only stay at a fair level around 65%. As shown in the four subplots of Fig. 13, no matter which criterion was used, S3 reaches and stays at better accuracy levels in the early selection stages. In substance, S3 gains the higher performance and overtakes S1 and S2 when C4.5, SVM, and NB classifiers are trained in the early selection round. As the *GR* criterion is implemented as shown in Fig. 14, S2 and S3 completely outperform S1 in general. We learn a superiority of S3 in the case of C4.5. An average accuracy level exceeding 70% derived from S2 and

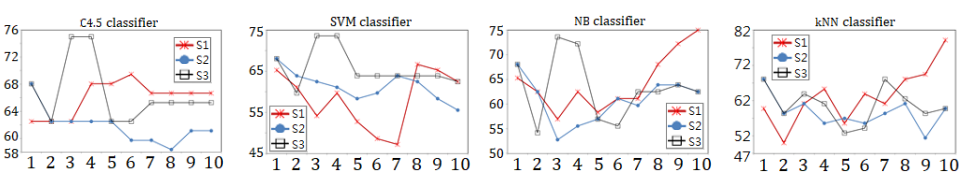


Fig. 13. Classification accuracies (%) for DS3 based on *IG*.

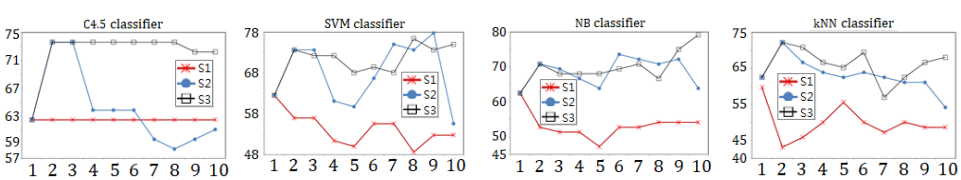


Fig. 14. Classification accuracies (%) for DS3 based on *GR*.

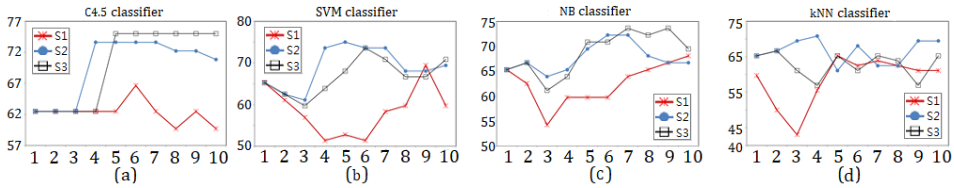


Fig. 15. Classification accuracies (%) for DS3 based on χ^2.

S3 is explicitly superior to that from S1 (around 50%). In Fig. 15, once again, S2 and S3 completely outperform S1 when implementing C4.5. Similarly, this phenomenon happens to SVM and NB during 2nd~8th selection rounds.

IG criterion collaborated with the method S2 could be very appropriate for DS4 with four target classes. They obtain a high accuracy over 92% with only four selected features involved in training classifiers C4.5, SVM, and NB. Redundancy checking in S3 is not helpful in this dataset so its performance is the same as S2 in Fig. 16. The steadier performance growth is not found when applying *GR* and χ^2 criteria. As shown in Fig. 17, the six best features evaluated by *GR* and selected by S1 only acquire the accuracy of 50%, embodying the situation that "the *m* best features are not the best *m* features". As applying *IG* leads to higher classification quality in DS4, we conclude that the features selected by S2 and S3 in DS4 possess more informative but less redundant message. Only the 5th, 6th, and 7th rounds observed in Fig. 18 (b) to which χ^2 criterion and SVM classifier were applied can be competitive to the peak of Fig. 16.

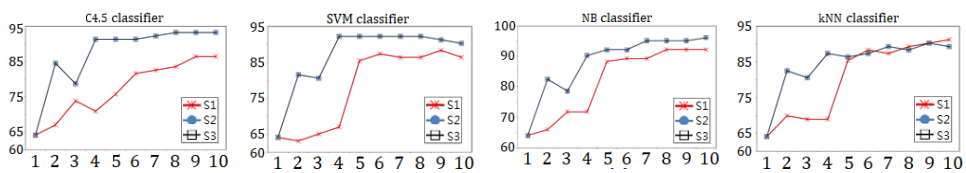


Fig. 16. Classification accuracies (%) for DS4 based on *IG*.

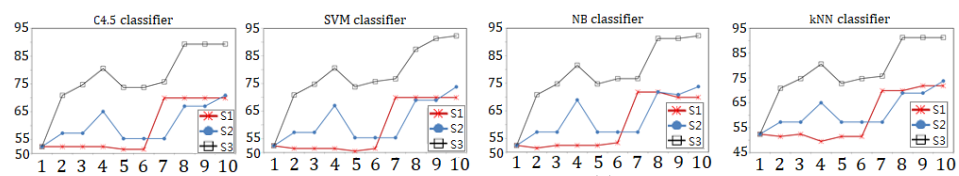


Fig. 17. Classification accuracies (%) for DS4 based on *GR*.

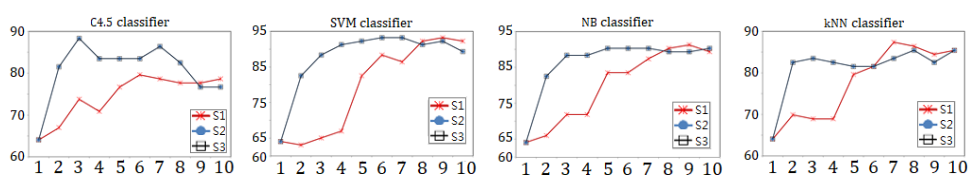
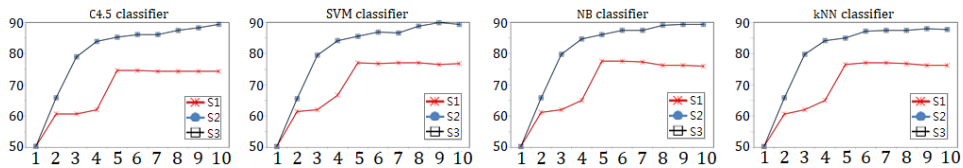


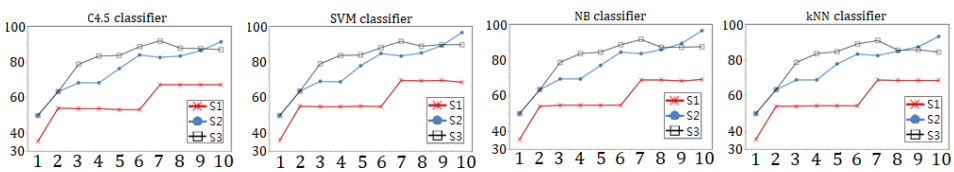
Fig. 18. Classification accuracies (%) for DS4 based on χ^2.

As the experimental results shown in Fig. 19, when DS10 is conducted using *IG* criterion, more features selected by S2 and S3 consistently promote the accuracy rate up to 90%, while S1 tends to encounter bottlenecks and stays at a rather low level around 75%. In Fig. 20, relevancy hybrid with redundancy analyses taken in S3 has better performance than S2 and S1, as *GR* criterion is employed, during first 8 selection rounds. However, relevancy analysis taken in S2 achieves the level exceeding 90% at the 10th round when SVM, NB, and kNN. Since gain ratio criterion takes the class number into account, DS10 with a higher class number of 6 could be the causes of this phenomenon. Similar to Fig. 19, Fig. 21 shows more consistent results of S2 and S3 than S1 when χ^2 criterion is adopted.



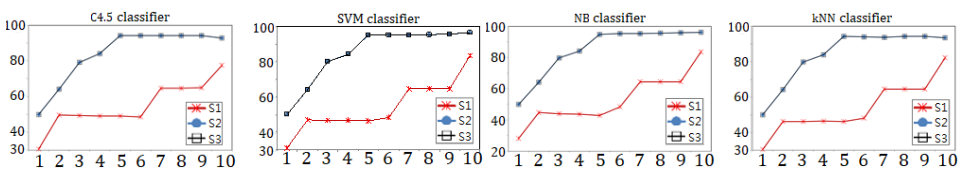
(a) (b) (c) (d)

Fig. 19. Classification accuracies (%) for DS10 based on *IG*.



(a) (b) (c) (d)

Fig. 20. Classification accuracies (%) for DS10 based on *GR*.



(a) (b) (c) (d)

Fig. 21. Classification accuracies (%) for DS10 based on χ^2.

7. CONCLUSIONS

This study is dedicated to extracting the entangled but useful information for classification tasks. The main contributions of this paper are threefold. First of all, feature discretization using fuzzy c-means cluster analysis facilitates the enhancement of discrimination power of characterizing features. Secondly, a novel feature evaluation design capable of balancing the positive and negative factors is presented. Thirdly, the pivotal classification information is excavated by our heuristic selection algorithm.

Although our method might not perform well in the case when non-continuous and independent features are taken into considered, it explicitly provides two helpful concepts for advancing classification tasks. First, the positive effect of relevant features and the negative effect of feature redundancy necessitate a more precise regulation for harmonizing their impacts. Second, variable selection principle should maximize the discrimination capa-

bility and minimize the size of the selected variable subsets. These two concepts will tailor classification techniques toward a new preferable boundary.

REFERENCES

1. T. Ahn *et al.*, "Deep learning-based identification of cancer or normal tissue using gene expression data," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, 2018, pp. 1748-1752.
2. B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature Biotechnology*, Vol. 33, 2015, pp. 831-838.
3. R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, Vol. 5, 1994, pp. 537-550.
4. M. Beraha *et al.*, "Feature selection via mutual information: New theoretical insights," in *Proceedings of International Joint Conference on Neural Networks*, 2019, pp. 1-9.
5. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, NY, 1981.
6. S. A. Bouhamed, I. K. Kallel, D. S. Masmoudi, and B. Solaiman, "Feature selection in possibilistic modeling," *Pattern Recognition*, Vol. 48, 2015, pp. 3627-3640.
7. A. Daniely, S. Sabato, and S. S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," *Advances in Neural Information Processing Systems*, Vol. 25, 2012, pp. 485-493.
8. J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, Vol. 3, 1973, pp. 32-57.
9. J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognition*, Vol. 51, 2016, pp. 295-309.
10. W. Gao, L. Hu, and P. Zhang, "Class-specific mutual information variation for feature selection," *Pattern Recognition*, Vol. 79, 2018, pp. 328-339.
11. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1157-1182.
12. K. Kianmehr, M. Alshalalfa, and R. Alhajj, "Fuzzy clustering-based discretization for gene expression classification," *Knowledge and Information Systems*, Vol. 24, 2010, pp. 441-465.
13. D. W. Kim, K. H. Lee, and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, Vol. 37, 2004, pp. 2009-2025.
14. J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy c-means algorithm," *IEEE Transactions on Fuzzy Systems*, Vol. 10, 2002, pp. 263-267.
15. S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering*, Vol. 32, 2006, pp. 47-58.
16. A. Kumar, L. Bhargava, and Z. Fatima, "Big data analytics and algorithms," in *Big Data Analytics*, 2021, Auerbach Publications, NW, pp. 19-39.
17. N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, Vol. 3, 2002, pp. 143-159.

18. J. Lee, I. Y. Choi, and C. H. Jun, "An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data," *Expert Systems with Applications*, Vol. 166, 2021, No. 113971.
19. H. Y. Lin, "Reduced gene subset selection based on discrimination power boosting for molecular classification," *Knowledge-Based Systems*, Vol. 142, 2018, pp. 181-191.
20. H. Y. Lin, "Gene discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification," *Applied Soft Computing*, Vol. 48, 2016, pp. 683-690.
21. H. Y. Lin, "Efficient classifiers for multi-class classification problems," *Decision Support Systems*, Vol. 53, 2012, pp. 473-481.
22. W. C. Lin, C. F. Tsai, and J. R. Zhong, "Deep learning for missing value imputation of continuous data and the effect of data discretization," *Knowledge-Based Systems*, Vol. 239, 2022, No. 108079.
23. J. Liu, Y. Lin, M. Lin, S. Wu, and J. Zhang, "Feature selection based on quality of information," *Neurocomputing*, Vol. 225, 2017, pp. 11-22.
24. T. R. Mahesh *et al.*, "Early predictive model for breast cancer classification using blended ensemble learning," *International Journal of System Assurance Engineering and Management*, 2022, pp. 1-10.
25. N. Mehra and S. Gupta, "Survey on multiclass classification methods," *International Journal of Computer Science and Information Technologies*, Vol. 4, 2013, pp. 572-576.
26. E. Min, *et al.*, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, Vol. 6, 2018, pp. 39501-39514.
27. L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, "Low-precision feature selection on microarray data: an information theoretic approach," *Medical & Biological Engineering & Computing*, Vol. 60, 2022, pp. 1333-1345.
28. M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Medical Genomics*, Vol. 13, 2020, pp. 1-13.
29. K. S. Myers, M. Place, D. R. Noguera, and T. J. Donohue, "COnTORT: COmprehensive Transcriptomic ORganizational Tool for simultaneously retrieving and organizing numerous gene expression data sets from the NCBI gene expression omnibus database," *Microbiology Resource Announcements*, Vol. 9, 2020, e00587-20.
30. M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, Vol. 37, 2004, pp. 487-501.
31. M. Panda, "Elephant search optimization combined with deep neural network for microarray data analysis," *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, 2020, pp. 940-948.
32. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, 2005, pp. 1226-1238.
33. A. Ponmalar, *et al.*, "Analysis of spam detection using Integration of logistic regression and PSO algorithm," in *Proceedings of the 4th IEEE International Conference on Computing and Communications Technologies*, 2021, pp. 396-402.
34. R. Ramirez, *et al.*, "Classification of cancer types using graph convolutional neural networks," *Frontiers in Physics*, Vol. 8, 2020, p. 203.

35. A. Senawi, H. L. Wei, and S. A. Billings, "A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking," *Pattern Recognition*, Vol. 67, 2017, pp. 47-61.
36. K. Shehzad, "EDISC: a class-tailored discretization technique for rule-based classification," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, 2012, pp. 1435-1447.
37. J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, Vol. 24, 2014, pp. 175-186.
38. J. Wang, J. M. Wei, Z. Yang, and S. Q. Wang, "Feature selection by maximizing independent classification information," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, 2017, pp. 828-841.
39. K. L. Wu and M. S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognition Letters*, Vol. 26, 2005, pp. 1275-1291.
40. Z. Zeng, H. Zhang, R. Zhang, and C. Yin, "A novel feature selection method considering feature interaction," *Pattern Recognition*, Vol. 48, 2015, pp. 2656-2666.
41. J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Method*s, Vol. 12, 2015, pp. 931-934.
42. P. Zhou, X. Hu, P. Li, and X. Wu, "Online feature selection for high-dimensional class-imbalanced data," *Knowledge-Based Systems*, Vol. 136, 2017, pp. 187-199.
43. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.
44. https://jundongl.github.io/scikit-feature/datasets.html.
45. https://sci2s.ugr.es/keel/category.php?cat=clas.
46. https://archive.ics.uci.edu/ml/datasets/dermatology.
47. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data#data.csv.

Hung-Yi Lin (林泓毅) is a Professor of the Department of Distribution Management of National Taichung University of Science and Technology. He received his Ph.D. degree at the Department of Applied Mathematics from National Chung Hsing University in 2005 winter. His current research interests include machine learning, reinforcement learning, and high dimensional feature engineering.

Rong-Chang Chen (陳榮昌) is a Professor of the Department of Distribution Management of National Taichung University of Science and Technology. He received his Ph.D. degree from National Chiao Tung University at Hsinchu, Taiwan in 1994. His research interests include supply chain management, logistics, data mining, artificial intelligence, and electronic commerce.