Knowledge-Based Systems 000 (2017) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Reduced gene subset selection based on discrimination power boosting for molecular classification

Hung-Yi Lin

Department of Distribution Management/National Taichung University of Science and Technology, Taichung 404, Taiwan, ROC

ARTICLE INFO

Article history:
Received 3 July 2017
Revised 27 November 2017
Accepted 29 November 2017
Available online xxx

Keyword:
Discrimination power
Feature selection
Molecular classification
Information gain
Cluster analyses

ABSTRACT

Traditional feature selection methods have two major inappropriate designs in their criterion. Firstly, they trade the profit of relevant information off against the risk of redundant information. Secondly, they cannot get rid of the well-known trap that "the *m* best features are not the best *m* features". There is no necessary inheritance between two consecutive selection rounds. As a remedy for the first problem, we propose a new selection criterion, which concentrates on verifying discrimination boosting effect. A novel feature selection scheme is also proposed in this paper as a mend on the second problem and it can generate multiple subsets with variable feature combinations supporting classification tasks. Our experimental results show that different subsets composed of variable selected features can have so quite similar discrimination power that they might achieve resembled classification quality. These experimental results also verify that our proposed method can successfully explore simple reduced subsets of genes for several genetic datasets with both efficacy and efficiency.

1. Introduction

Feature selection is one of the most decisive issues and a significant challenge in classification problems, pattern recognition, machine learning, and data mining. According to the possible use of output information (e.g., target class label) feature selection methods can be broadly classified as either supervised or unsupervised. Supervised methods [4,23,28] select features based on two principles: one is the relevancy to the prediction of some output information and the other is the redundancy to one another feature. In contrast, unsupervised methods [11,31,32,35] do not necessitate any a priori knowledge involving the output information. There are also two doctrines for unsupervised methods. The first is that they aim to find the smallest subset of features in such a way that all the information content of a dataset is preserved as much as possible. The second is to eliminate all the redundant information between the selected features as far as possible.

Dealing with *irrelevant* or *relevant* features to the target class label has been proposed in most existing supervised feature selection algorithms [8,10,17,18,20–22,25]. Redundancy minimization has also attracted many mature methods [8,9,11,20,22,25,32,33]. Remarkably, relevancy analyses require the input and output information in a dataset and necessitate supervised learning while redundancy analyses only depend on the input information and merely appeal to unsupervised learning. Analytical results attained from two different learning systems (i.e., supervised and unsupervised) need simple yet reasonable integration so as to distinguish the most informative feature subsets from the useless ones. A number of selection criterions and selection methods have achieved success in relevancy analyses or redundancy analyses separately. However, how to assemble these two items in order to obtain a more convincing argument is rarely found in the past studies.

As far as one feature subset is concerned, inclusion of an additional feature can subsequently trigger many influences. Most important of all, whether the improvement of discrimination power becomes apparent? And then, the interaction between all features becomes helpful or not. Is the gained efficacy worthy of the costly execution of selection? In practical applications, rule generalization, overfitting problem, method performance, and classification error may also be highly concerned when evaluating the effectiveness of one newly selected feature.

The motivation of feature selection is to reduce classification errors and the core of this task is to improve and promote the discrimination power of a collection of selected features. Yet considering the fact that authentication of discrimination power of selected features is the essential and fundamental cause, it should be completed in advance of processing feature selection. The discrimination power derived from a collection of features is theoretically expected to classify the target class variable as far as possible. Processing relevancy and redundancy analyses [33] using conventional multivariate analytical procedures will be faced with challenges.

E-mail address: linhy@nutc.edu.tw

https://doi.org/10.1016/j.knosys.2017.11.036
0950-7051/

For instance, the high relevance gains great attraction but suffers from the well-known problem that "the *m* best features are not the best *m* features" [24]. In other words, the collection of the best feature selected in every round does not guarantee to form the best feature subset. As independence but not compensation, redundancy analyses among all the selected features have no less importance than the relevancy ones. A critical issue is that large new information does not signify little redundancy, and vice versa [30]. Relevance and redundancy analyses are based on different information sources (i.e., interaction of inter- and intra-features), whose resulting effects should be taken into account separately and they cannot compensate the insufficiencies for each other. In order to illustrate such misunderstanding presented in many past studies, more detailed discussions are given in the following sections.

The aim of feature selection [12] is to select subsets of variables from the input data, which can efficiently and effectively describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results. In fact, the *necessary* variables generally imply those fundamental or primary factors chosen according to a special criterion or designated manually. Necessary variables are responsible for rough classification or initial categorization. Moreover, additional variables to these necessary ones can refine classification or execute detailed categorization. Nevertheless, all selected variables generally collaborate with one another, and it is meaningless to distinguish between the necessary or extra features, which should be merely regarded as a strategic practice. The number of such extra variables is generally expected to be within a reasonable but *sufficient* quantity. Hence, we assert in this paper the *necessary* and *sufficient* condition should be considered as a top principle of feature selection. The main design issue of this paper is to plan a novel selection scheme that is capable of supporting *necessary* and *sufficient* discrimination power for molecular classification problems.

Modern classification tasks such as text categorization [5], gene microarray analysis [3,19], bioinformatics applications [8,26], combinatorial chemistry, and semantic annotation of media [2,14] necessitate the exploration of useful and relevant patterns and knowledge from Big Data [34]. Patterns and knowledge hide deeply behind diverse and complex data [7]. Unfortunately, data diversity in real-valued features can complicate analytical processes considerably, in turn increasing the computational costs of exploration. Practically, the dissimilarity message among continuous values is more significant than their measured magnitudes. In the initial stage of microarray analyses, the preprocessing of massive amount of continuous gene expression levels is so critical that it could largely influence the succeeding procedures. Clustering continuous values into similar and dissimilar groups can reduce complex data and retain the discriminatory information. Cluster analysis (CA) is used in this paper and responsible for the discretization task of feature values or feature vectors. The discriminative effects of clustered data explicitly outperform those of non-clustered data has been verified in [21]. Since different CAs result in diverse distinction and discretized outcomes, three CAs are employed in our comparing experiments.

Feature selection techniques are categorized into three types: filter methods, wrapper methods and embedded methods [27]. Most of them dedicate their efforts to the heuristic methods and speed up the process of finding a satisfactory solution. Filter methods ignore feature dependencies and the interaction with the used classifiers so that the search in the feature subset space is separated from the search in the hypothesis space. Wrapper methods demand constantly referring to the classification results generated by the engaged classifiers. Embedded methods alter and adjust feature subsets on the strength of the resulting classification quality. To the extent of high dimensional data, the suffering of trial-and-error still challenges these heuristic procedures. This paper attempts to improve the selection efficiency when handling the high dimensional data of bioinformatics. Molecular classification tasks rely on powerful discrimination capability. Our design issue is twofold. The first concentrates on boosting the discrimination power of feature subsets. The second is to develop a new feature selection scheme which explores necessary and sufficient features for molecular classification tasks. The contribution of our method is to retain the advantage of simplicity and lower the time spent selecting in filter methods, while the disadvantage of high computational cost due to the exponential quantity of feature subsets in wrapper methods can be moderately reduced.

2. The weakness and drawbacks of traditional methods

Information gain (IG) and gain ratio (GR) [6,13] are two typical filter criteria commonly used methods in feature ranking [23] and relevance analyses. IG is entropy-based method based on information theory. GR is the variation of IG, which takes class number into account and is normalized. Mutual information [29] is the most popular method when measuring the redundant information in a subset of features. Subtracting $\Sigma I(f_s;f_i)$ from $I(C;f_i)$ is a frequent design adopted in many past feature selection criteria. The MIFS algorithm [1] MIFS-U algorithm [18] mRMR criterion [25] and NMIFS algorithm [12] are respectively formulated as following:

$$I(C;f_i) - \beta \sum_{f_s \epsilon S} I(f_s;f_i) \tag{1}$$

$$I(C;f_i) - \beta \sum_{f_s \epsilon S} \frac{I(C;f_s)}{H(f_s)} I(f_s;f_i) \tag{2}$$

$$I(C;f_i) - \frac{1}{|S|} \sum_{f_s \epsilon S} I(f_s;f_i) \tag{3}$$

$$I(C;f_i) - \frac{1}{|S|} \sum_{f_s \epsilon S} NI(f_s;f_i) \tag{4}$$

Unfortunately, integrating $I(C;f_i)$ and $\Sigma I(f_s;f_i)$ in a linear formulation is problematic in two aspects. The first aspect relates to the consistency of the measured information. $I(C;f_i)$ measures the *relevance* of the feature to be added, and $\Sigma I(f_s;f_i)$ estimates the *redundancy* of the *i*-th feature with respect to the subset previously selected features. Although the asymmetric selection weight between the left side and right side in (1) and (2) is solved by dividing the sum with the cardinality of the set *S* proposed in (3) and (4). 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 $\Sigma I(f_s;f_i)$ are both based on entropy theory, $I(C;f_i)$ is focusing on the target class label while $\Sigma I(f_s;f_i)$ 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.

Now that features are selected one-by-one, 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. Therefore, we redefine one feature with good discrimination power so that it can maximize its incremental relevance to the target class variable and minimize its incremental redundancy to the already selected features when comparing with other features.

In addition, one-by-one selection strategy easily falls into the problem of local optimization. Suppose S_n denotes the feature subset with *best* discrimination power generated in the *n*-th selection round. We note that features in S_n do not necessarily constitute

Please cite this article as: H.-Y. Lin, Reduced gene subset selection based on discrimination power boosting for molecular classification, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.11.036