Chapter 105
An Evolutionary Approach for Balancing Effectiveness and Representation Level in Gene Selection

Nicoletta Dessì
Università degli Studi di Cagliari, Italy

Barbara Pes
Università degli Studi di Cagliari, Italy

Laura Maria Cannas
Università degli Studi di Cagliari, Italy

ABSTRACT

As data mining develops and expands to new application areas, feature selection also reveals various aspects to be considered. This paper underlines two aspects that seem to categorize the large body of available feature selection algorithms: the effectiveness and the representation level. The effectiveness deals with selecting the minimum set of variables that maximize the accuracy of a classifier and the representation level concerns discovering how relevant the variables are for the domain of interest. For balancing the above aspects, the paper proposes an evolutionary framework for feature selection that expresses a hybrid method, organized in layers, each of them exploits a specific model of search strategy. Extensive experiments on gene selection from DNA-microarray datasets are presented and discussed. Results indicate that the framework compares well with different hybrid methods proposed in literature as it has the capability of finding well suited subsets of informative features while improving classification accuracy.

DOI: 10.4018/978-1-5225-1759-7.ch105

INTRODUCTION

Feature selection is one of the important and frequently used techniques in data mining (Chandrashekar & Sahin, 2014). It reduces the number of features, removes irrelevant, redundant, or noisy data, and improves mining performance such as predictive accuracy and result comprehensibility.

The goodness of selected features is usually measured by an evaluation criterion that strongly affects results, i.e. an optimal set of features selected using one criterion may not be optimal according to another criterion. Despite the work on developing criteria for evaluating the quality of results in feature selection algorithms (Kumar & Minz, 2014), the choice of the algorithm appropriate for classification problems remains difficult. It has been argued (Liu & Yu, 2005) that the more feature selection algorithms available, the more difficult it is to select a suitable one for a classification task.

While there is no agreement about the definition of the mathematical statement of the problem (Guyon & Elisseeff, 2003), two major factors seem to be particularly important in designing a suitable algorithm for feature selection in a classification task: improving the predictive accuracy and providing better understanding of the underlying concept that generated the data. We denote the above factors as the *effectiveness* and the *representation level* of the feature selection process.

Specifically, the effectiveness deals with selecting the minimum set of variables that maximize the accuracy of a classifier and the representation level concerns discovering how relevant the variables are for the considered domain.

In more detail, the effectiveness attempts to capture the performance aspect of classification. From this point of view, the major challenge is finding a minimum subset of features that are useful to the prediction. Thus, this aspect is central for classification problems in which accuracy is of primary concern: the more effective the feature selection, the better the performance of the resulting classifier.

The representation level reflects the explanatory power of the selected features in representing essential knowledge about the application domain. The focus is on discovering all the variables suited to the reality that we are trying to represent, deciding how relevant and informative they are. Under this paradigm, the feature selection process privileges the usefulness of the features in representing the application domain i.e. the degree of exactness with which the representation fits the reality.

Research efforts have produced methods that place the emphasis at different times on the effectiveness or on the representation level (Tang, Alelyani, & Liu, 2014). Among the broadly used methods, *rankers* evaluate the discriminative power of features with regard to the class labels of samples by looking only at the intrinsic properties of the data. Thus, rankers emphasize the representation level giving as output a list where features are ordered based on their relevance for the classification task at hand.

Leveraging on rankers, *filter* methods strive to improve the effectiveness by selecting a certain number of highest ranked features for the purpose of classification. However, because the number used is somewhat arbitrary, features selected under this approach depend on an "a priori" choice with little support for determining how many features should be chosen for classification. Moreover, filters do not take into account the classifier to be applied.

In contrast to the filter approach, *wrapper* methods adopt a paradigm in which the main emphasis is on selecting features during the process of classification. Different subsets of features are generated by using a search algorithm and then are evaluated by training and testing a specific classification model. As the whole process aims to optimize the accuracy of the particular classifier, the central aspect in selecting features is the effectiveness rather than improving the representation level.