Feature Selection

Noelia Sánchez-Maroño
University of A Coruña, Spain

Amparo Alonso-Betanzos
University of A Coruña, Spain

INTRODUCTION

Many scientific disciplines use modelling and simulation processes and techniques in order to implement non-linear mapping between the input and the output variables for a given system under study. Any variable that helps to solve the problem may be considered as input. Ideally, any classifier or regressor should be able to detect important features and discard irrelevant features, and consequently, a pre-processing step to reduce dimensionality should not be necessary. Nonetheless, in many cases, reducing the dimensionality of a problem has certain advantages (Alpaydin, 2004; Guyon & Elisseeff, 2003), as follows:

- Performance improvement. The complexity of most learning algorithms depends on the number of samples and features (curse of dimensionality). By reducing the number of features, dimensionality is also decreased, and this may save on computational resources—such as memory and time—and shorten training and testing times.
- Data compression. There is no need to retrieve and store a feature that is not required.
- Data comprehension. Dimensionality reduction facilitates the comprehension and visualisation of data.
- Simplicity. Simpler models tend to be more robust when small datasets are used.

There are two main methods for reducing dimensionality: feature extraction and feature selection. In this chapter we propose a review of different feature selection (FS) algorithms, including its main approaches: filter, wrapper and hybrid – a filter/wrapper combination.

BACKGROUND

Feature extraction and feature selection are the main methods for reducing dimensionality. In feature extraction, the aim is to find a new set of r dimensions that are a combination of the n original ones. The best known and most widely used unsupervised feature extraction method is principal component analysis *(PCA)*; commonly used as supervised methods are linear discriminant analysis *(LDA)* and partial least squares *(PLS)*.

In feature selection, a subset of r relevant features is selected from a set n, whose remaining features will be ignored. As for the evaluation function used, FS approaches can be mainly classified as filter or wrapper models (Kohavi & John, 1997). Filter models rely on the general characteristics of the training data to select features, whereas wrapper models require a predetermined learning algorithm to identify the features to be selected. Wrapper models tend to give better results, but when the number of features is large, filter models are usually chosen because of their computational efficiency. In order to combine the advantages of both models, hybrid algorithms have recently been proposed (Guyon et al., 2006).

FEATURE SELECTION

The advantages described in the Introduction section denote the importance of dimensionality reduction. Feature selection is also useful when the following assumptions are made:

- There are inputs that are not required to obtain the output.
- There is a high correlation between some of the input features.

A feature selection algorithm (FSA) looks for an optimal set of features, and consequently, a paradigm that describes the FSA is heuristic search. Since each state of the search space is a subset of features, FSA can be characterised in terms of the following four properties (Blum & Langley, 1997):

- The initial state. This can be the empty set of features, the whole set or any random state.
- The search strategy. Although an exhaustive search leads to an optimal set of features, the associated computational and time costs are high when the number of features is high. Consequently, different search strategies are used so as to identify a good set of features within a reasonable time.
- The evaluation function used to determine the quality of each set of features. The goodness of a feature subset is dependent on measures. According to the literature, the following measures have been employed: information measures, distance measures, dependence measures, consistency measures, and accuracy measures.
- The stop criterion. An end point needs to be established; for example, the process should finish if the evaluation function has not improved after a new feature has been added/removed.

In terms of search method complexity, there are three main sub-groups (Salapa et al., 2007):

- Exponential strategies involving an exhaustive search of all feasible solutions. Exhaustive search guarantees identification of an optimal feature subset but has a high computational cost. Examples are the branch and bound algorithms.
- Sequential strategies based on a local search for solutions defined by the current solution state. Sequential search does not guarantee an optimal result, since the optimal solution could be in a region of the search space that is not searched. However, compared with exponential searching, sequential strategies have a considerably reduced computational cost. The best known strategies are sequential forward selection and sequential backward selection (SFS and SBS, respectively). SFS starts with an empty set of features and adds features one by one, while SBS begins with a full set and removes features one by one. Features are added or removed on the basis of improvements in the evaluation function. These approaches do not consider interactions between features, i.e., a feature may not reduce error by itself, but improvement may be achieved by the feature's link to another feature. Floating search (Pudil et al., 1994) solves this problem partially, in that the number of features included and/or removed at each stage is not fixed. Another approach (Sánchez et al., 2006) uses sensitivity indices (the importance of each feature is given in terms of the variance) to guide a backward elimination process, with several features discarded in one step.
- Random algorithms that employ randomness to avoid local optimal solutions and enable temporary transition to other states with poorer solutions. Examples are simulated annealing and genetic algorithms.

The most popular FSA classification, which refers to the evaluation function, considers the three (Blum & Langley, 1997) or last two (Kohavi & John, 1997) groups, as follows:

- Embedded methods. The induction algorithm is simultaneously an FSA. Examples of this method are decision trees, such as classification and regression trees (CART), and artificial neural networks (ANN).
- Filter methods. Selection is carried out as a preprocessing step with no induction algorithm (Figure 1). The general characteristics of the training data are used to select features (for example, distances between classes or statistical dependencies). This model is faster than the wrapper approach (described below) and results in a better generalisation because it acts independently of the induction algorithm. However, it tends to select subsets with a high number of features (even all the features) and so a threshold is required to choose a subset.
- Wrapper methods. Wrapper models use the induction algorithm to evaluate each subset of features, i.e., the induction algorithm is part of the evaluation function in the wrapper model, which means this model is more precise than the filter model. It also takes account of techniques, such as cross-validation, that avoid over-fitting. However, wrapper models are very time consuming, which