Chapter 1

Up-to-Date Feature Selection Methods for Scalable and Efficient Machine Learning

Amparo Alonso-Betanzos  
University of A Coruña, Spain

Diego Fernández-Francos  
University of A Coruña, Spain

Verónica Bolón-Canedo  
University of A Coruña, Spain

Iago Porto-Díaz  
University of A Coruña, Spain

Noelia Sánchez-Maroño  
University of A Coruña, Spain

ABSTRACT

With the advent of high dimensionality, machine learning researchers are now interested not only in accuracy, but also in scalability of algorithms. When dealing with large databases, pre-processing techniques are required to reduce input dimensionality and machine learning can take advantage of feature selection, which consists of selecting the relevant features and discarding irrelevant ones with a minimum degradation in performance. In this chapter, we will review the most up-to-date feature selection methods, focusing on their scalability properties. Moreover, we will show how these learning methods are enhanced when applied to large scale datasets and, finally, some examples of the application of feature selection in real world databases will be shown.

INTRODUCTION

In recent years, the dimensionality of datasets, as can be seen in (Zhao & Liu, 2012), has increased steadily. As new applications appear, a dataset with a dimensionality above 10,000 is common in applications such as, for example, medical images, text retrieval or genetic data, most of which are now measured in petabytes (PB, $2^{50}$ bytes). A database is considered high dimensional when (a) the number of samples is very high, (b) the number of features is very high, or (c) the number of samples and features is very high. Learning methods become particularly difficult when dealing with datasets with around 1,000,000 data (samples x features being data).
When data dimensionality is high, many of the features can be redundant or irrelevant, and when high dimensionality is based upon the number of features and not only on the sample size, can even be a more complex problem, more so if the data sample is small. As a typical example, DNA microarray data could contain more than 30,000 features with a sample size of usually less than 100. With datasets of this type, most techniques can become unreliable.

Theoretically, it seems logical that having a higher amount of information could lead to better results. However, this is not always the case due to the so-called curse of dimensionality (Bellman, 1957). This phenomenon occurs when dimensionality increases and the time required by the machine learning algorithm to train the data increases exponentially. High dimensionality constitutes a new challenge for data mining, because the performance of learning algorithms can degenerate due to overfitting. Also, learned models decrease their interpretability as they become more complex and consequently the speed and efficiency of the algorithms declines in accordance with size.

In order to deal with these problems, dimensionality reduction techniques are usually applied, so that the set of features needed for describing the problem can be reduced. Moreover, most of the times, the performance of the models can be improved, together with data and model understanding, as well as reducing the need for data storage (Guyon, Gunn, Nikravesh, & Zadeh, 2006). Dimensionality reduction techniques can be broadly classified into feature construction and feature selection methods. Feature construction techniques attempt to generate a set of useful features than can be used to represent the raw data of a problem. A pre-processing transformation could be considered which may alter the space dimensionality of the problem by enlarging or reducing it. The data set generated by feature extraction is represented by a newly generated set of features, different than the original. On the other hand, feature selection is the process of detecting the relevant features and discarding the irrelevant ones in order to obtain a subset of features that correctly describes the given problem with a minimum degradation of performance. In this case, dimensionality reduction is always achieved, and the set of features is a subset of the original set. As feature selection maintains the original features, it is especially interesting when they are needed for interpreting the model obtained. Feature selection techniques can be applied to both the two main areas of machine learning: classification and regression, although the former is the more common, the reason being that the typical classification problem has several variables whilst the common regression problem usually has just one.

Feature selection methods can be divided into two models: individual evaluation and subset evaluation (Yu & Liu, 2004). The former is also known as feature ranking and evaluates individual features by assigning them weights in a ranking according to their degree of relevance, whilst the latter produces candidate feature subsets based on a certain search strategy. Not all the ranker methods provide a weight of the features, so rankers can also be divided into score methods, i.e. those methods assigning a relevance value to each feature, and “pure rankers” that only return a list of ordered features and the difference in relevance of a feature and the next (or previous) one in the list is unknown.

Aside from this classification, three major approaches (see Figure 1) can be distinguished based upon the relationship between the feature selection algorithm and the inductive learning method used to infer a model (Guyon, Gunn, Nikravesh, & Zadeh, 2006):

- **Filters**: Which rely on the general characteristics of the data and perform the feature selection process as a pre-processing step,
Related Content

The Formal Design Model of a Lift Dispatching System (LDS)
[www.igi-global.com/article/formal-design-model-lift-dispatching/37491?camid=4v1a](www.igi-global.com/article/formal-design-model-lift-dispatching/37491?camid=4v1a)

Algorithms and Principles for Intelligent Design of Flapping Wing Micro Aerial Vehicles
[www.igi-global.com/chapter/algorithms-principles-intelligent-design-flapping/72506?camid=4v1a](www.igi-global.com/chapter/algorithms-principles-intelligent-design-flapping/72506?camid=4v1a)

Toward Automatic Answers in User-Interactive Question Answering Systems
Tianyong Hao, Feifei Xu, Jingsheng Lei, Liu Wenyin and Qing Li (2011). *International Journal of Software Science and Computational Intelligence* (pp. 52-66).
[www.igi-global.com/article/toward-automatic-answers-user-interactive/64179?camid=4v1a](www.igi-global.com/article/toward-automatic-answers-user-interactive/64179?camid=4v1a)

Nonstationary Signal Analysis with Kernel Machines
[www.igi-global.com/chapter/nonstationary-signal-analysis-kernel-machines/36987?camid=4v1a](www.igi-global.com/chapter/nonstationary-signal-analysis-kernel-machines/36987?camid=4v1a)