Best Practices of Feature Selection in Multi-Omics Data

B

Funda Ipekten
https://orcid.org/0000-0002-6916-9563
Erciyes University, Turkey

Gözde Ertürk Zararsız
Erciyes University, Turkey

Halef Okan Doğan
Cumhuriyet University, Turkey

Vahap Eldem
Istanbul University, Turkey

Gökmen Zararsız
Erciyes University, Turkey

INTRODUCTION

Today, there is an increase in data in many areas. With this increase, the number and variety of the variables to be evaluated also increases. The increase in data and variables became a situation that needed to be solved among world problems. In addition, although there is a perception that having too much data in the scientific field, having too much information, correct information, or sufficient information may not be possible. However, it should not be forgotten that there is valuable information in a relatively large amount of data. It should be clear that it can be beneficial to have much data to extract this helpful information. However, performing data analyses to obtain and process this information can be difficult. In addition, its existence is a problem called the curse of data dimensionality (Verkeysen M. and François D., 2005). High-dimensional data sets, where these problems are most common, are used successfully in multiple fields such as genetics, pharmacology, toxicology, nutrition, and genetics. The use of these high-dimensional data allows one to examine biology systems, cellular metabolism, and disease etiologies in more detail. However, the number of samples (n) of these data is considerably lower than the number of variables (p) and the heterogeneity of the data, the missing observations in the data as a result of the use of high-output technology, limits the use of traditional methods that can be used in this field. Therefore, there is a need for the clinical understanding of the biological system based on research and machine learning, and statistical learning methods to analyze this clinical information statistically (Hastie et al., 2009). Several studies are show that machine learning methods are used and applied successfully in studies carried out in this field. Some of these studies are listed in Table 1.

DOI: 10.4018/978-1-7998-9220-5.ch122

Table 1. Some studies using feature selection

Datasets		Methods	References		
Ovarian cancer	Classification	MKL	Wilson et al.	2019	
Breast cancer	Classification	MKL	Tao et al.	2019	
Gene expression	Classification	SVM	Golub et al.	1999	
Gut microbiota	Classification	RF	Franzosa et al.	2019	
Colon cancer	Classification	SVM	Moler et al.	2000	
Ovarian, leukemia, colon	Classification	SVM	Furey et al.	2000	

MKL: Multiple Kernel Learning, SVM: Support Vector Machine, RF:Random Forest

In general, in all of the studies given in Table 1, researchers aim to optimize the classification of disease-related samples, produce models that can be used to predict system behaviors, or properties or provide the most accurate result appropriate in terms of classification performances. However, the large number of variables in the data used can complicate the structure of the models to be created hence reducing their accuracy. In addition, because the number of variables is too large, the investigation of disease-related genes or other omics causes considerable losses in terms of both time and cost (Hastie et al., 2009). Therefore, it is not always possible for researchers to carry out these studies in depth. Therefore, feature selection is made before model training is evaluated to make the models obtained with learning methods more generalizable, predictable, and ineffective against noisy values, with minimum time and minimum cost (Díaz-Uriarte and Alvarez de Andrés, 2006). At this point, the need for feature selection methods increases in developing new technology and bioinformatics applications.

BACKGROUND

The use of feature selection methods has become a prerequisite for models created by using machine learning algorithms and statistical learning methods in the analysis of high-dimensional data sets such as omics data (Metabolomic, transcriptomic, genomic, etc.), next-generation sequencing data, microarray data (Saeys, 2007).

Feature selection purposes;

- It is done to prevent overfitting problems and to improve model performance. It is used for the prediction performance of the relevant situation in classification problems and for better cluster detection in clustering problems.
- It is aimed to create models with the minimum cost at maximum speed.
- It allows more detailed scanning for the investigated situation.

Feature selection methods, which are used for more than one purpose, have an important place for the quality of the studies without compromising the data ‘s originality. This is because the original structure of the data does not change in feature selection. Instead, a subset is selected for the relevant situation (sample or number of features) (Saeys, 2007). Since feature selection methods are independent of the classifier, they are used in supervised, semi-supervised, and unsupervised learning techniques. These methods, which are widely used in machine learning methods, have varieties such as filter, wrapper, and ensemble (Ferreira, 2012). The advantages and disadvantages of these methods are presented in Table 2.