Feature Selection in Bioinformatics

Paulo Leal
Faculdade de Ciências
Universidade do Porto
Email: up201101503@fc.up.pt

António Carvalho
Faculdade de Ciências
Universidade do Porto
Email: up200801471@fc.up.pt

Inês Maia
Faculdade de Ciências
Universidade do Porto
Email: up201101593@fc.up.pt

***Abstract*—In Bioinformatic applications, feature selection techniques provide a solution to the large number of input features, finding which of these are useful in classification of a particular problem. This paper explains these techniques and their possibilities in this area, discussing types of methods and provides a overview of current and future applications.**

I. Introduction

From an illustration to a prerequisite, Feature Selection (FS) has become increasingly important over the last years, having been given rise by developments in various modeling tasks as sequence analysis, spectral analysis, and literature mining. A brief explanation and discussion of FS follows in section II. We also describe the applications in Bioinformatics and the division of the application attribute selection: Sequence, Microarray and Mass Spectra Analysis (Section III). Furthermore, we explain the small sample domains and emphasizes two techniques to solve this problem: evaluation criteria and FS approaches (Section IV). Lastly, there is a overview of Feature Selection in upcoming domains, explaining single nucleotide polymorphism analysis and text and literature mining (Section V) and we provide the existing software packages that implements feature selection methods (Section VI).

II. Feature Selection Techniques (FST)

FST are pattern recognition techniques that serve as a way to avoid overfitting and improve performance (either in a prediction performance as for supervised classification or an improved cluster detection for clustering), to provide more efficient, lower cost models with greater speed, and to better understand how the data was processed.
These advantages also bring added complexity to the modeling task, meaning that instead of optimizing only the model we also need to find adequate parameters for the feature subset as it is not guaranteed that the full feature set relies on the same values that equally satisfy the feature subset.
As for classification of these techniques they can be divided into three categories which relate to how the feature selection techniques

A. Filter Techniques

Analyzing only the basic data features are giving a score and each that scores low is removed, the remaining ones are used to create a subset that is used as input for the classification algorithm. These techniques are easily scaled to very large data sets, they are simple, utilize low computational power, and do not depend on the algorithm they are paired with. Since utilizing only basic features may create issues multivariate techniques (which are discussed in III-B2) can be used instead, although they too only use a degree of dependencies they still provide an improvement.

B. Wrapper Techniques

In contrast to filter techniques, wrapper technique defines a search using a predictive model to score feature subsets. There are some subsets of features that will be created and others will be assessed by classification models and in this evaluation there is a training and model test.

With that, for each new subset will be a training for a model, evaluating this operation and getting a future rating for that subset. As there is this intensive task of testing for each subset, this will take exponential time, using heuristics.

Taking into account this task of exponential time, wrapper technique are usually more computationally intensive than filter. However, there are some important advantages like the capability of perceive feature dependencies and essentially the interaction of model selection in feature subset search.

C. Embedded Techniques

Embedded methods are a group of techniques that perform the feature selection search as part of the classifier construction process. So, this will have an interaction with the classification model as a good part of the wrapper technique but is less computationally intensive.

III. Applications in Bioinformatics

Given the fact that must be a choice of attributes in order to apply this technique in relevant variables, the application attribute selection is divided by:

A. Sequence Analysis

Sequence analysis is based on a selection of attributes, applying FST only in the sequence of variables that really matter (without irrelevant data), consisting of two parts: content analysis and signal analysis.

Before explaining each type, this can be summarized as the interest of applying this technique - if the aim is to verify the general characteristics of the sequence, the suitable type is the content analysis; however, if the important thing is to focus on a specific part of sequence for any detection, it will make more sense to use signal analysis.

1) Content Analysis: This type of sequence analysis focuses on the sequence as a whole, that is, the general characteristics of that and therefore, an important part is having ordered sequence to establish a connection between adjacent features.

With this order, combined with an important application and widely used in bioinformatics - coding potential prediction, comes the Markov models. This model comes with several variations, for handling small sample sizes or extract relevant attributes with filter methods, and are essentially used to extract characteristics and information in a sequence.

Another FST in content analysis is using selective kernel scaling for support vector machines (SVM) and genetic algorithms that access resources weights and therefore removes resources with low weight.

2) Signal Analysis: Signal analysis focuses not in the sequence as a whole but only in small parts of that, having interest in recognize presence of particular elements (like gene structural or regulatory elements), may then consider a search for conserved signals in the sequence.

Therefore, this recognition of important elements it is usually used for finding places of relevant functions and, to find that, usually it is used a regression models, by relating the important motifs in the sequence to the gene expression.

To classify the motifs and genes, it is used the threshold number of misclassification (TNoM) with a calculated p value for each motif, which will sequentially be ordered.

B. Microarray Analysis

In bioinformatics, with the existence of new data sets of microarrays, there is much attention and research in this area due to the great challenge of their large dimensionality of genes and their small sample sizes. To increase the complexity, coupled with that, it comes other problems like noise and variability.

1) Univariate Filter: Univariate filter methods became a feature selection technique widely used, for some reasons such as: understandable and intuitive output provided by univariate feature rankings; less computation time than multivariate filter (faster); have a simpler method to validate the selection by biological lab methods and may be no need in consider genes interactions by the experts.

There are two simple heuristics techniques: setting a threshold on the differences in gene expression between the states and detection of the threshold point in each gene that minimizes TNoM, which can be divided in two parts: parametric and model-free methods;

Parametric methods assume a distribution from which the samples have been generated. In microarray studies, the t-test is one of the most used techniques but some modifications are needed in order to deal with noise and sample size.

The model-free methods estimate the reference distribution of the statistics by using random permutations of data, which enhances the robustness against outliers and alleviates the small sample problem.

2) Multivariate Filter: Since the univariate selection methods do not take into account the gene interactions, other methods have been proposed - multivariate filter methods which essentially have three types:

- Filter methods - Measure feature subset "relevance" and usually order features (individual feature ranking or nested subsets of features). Some of these methods are correlation-based feature selection (CFS) and several variants of the Markov blanket filter method;
- Wrapper methods - Measure feature subset "usefulness" and have two important characteristics: the scoring function used to evaluate each gene subset and crossing a univariately preorder gene ranking with an incrementally augmenting wrapper method. These methods use randomized search heuristics, population-based and sequential search techniques;
- Embedded methods - are similar to wrappers but less computationally expensive. It has focused on discard input features, proposing a subset of discriminative genes. Some of the embedded FS techniques are the Random Forest (which will be explained in section IV) and tecniques using feature weights (give greater or lesser relevance to genes) in linear classifiers, such as SVMs and logistic regression.

C. Mass Spectra Analysis

In recent years, mass spectrometry technology (MS) has become a powerful framework to study proteins on a large-scale, having innovative experimental strategies helping to diagnose diseases. A mass spectrum sample can be characterized by a XY graph: in the x-axis are represented the mass/charge ratio and in the y the corresponding signal itensity and may have more than 15500 data points. [1] [2]

- Filter methods - In order to extract the variables that may be the discriminative features, firstly is important normalize the spectrum and then it can be considered every measured variable or extract the feature, limiting the number of variables;
- Wrapper methods - This method focuses in the feature extraction, reducing the number of variables. To accomplish that, are used population-based randomized heuristics as search engines and may be some variations of this method such as SVM-formulation with weights of variables, neural network classifier and algorithms of decision-trees.

IV. Small Sample Domains

In bioinformatics, a problem that can occur is associated with the existence of small sample sizes given that these samples can have information that may be inaccurate given the small amount of data causing wrong data collection.

In order to solve this problem, two techniques were created: the use of adequate evaluation criteria and the use of FS models that ensure stable and robust.

A. Evaluation Criteria

It was known that some applications could be giving an incorrect value of the reported accuracy percentages, testing the final model with a subset of discriminative features.

Added do this, there is a greater growth of external feature selection process for training the model with the aim of getting a more accurate value. There are evaluation criterion as error rate, which is calculated through a error counting mechanism; and other methods that do not explicity count the errors but their probability. [1] [3]

To deal with these small samples and verify a predictive accuracy, arises bolstered error estimation, which also carries some research to understand the counting based error estimators impact on gene selection. [4]

B. Approaches

Specifying Feature Selection Techniques in Data Mining, it is known that there is an essential point in selecting a subset for evaluating a whole set - the chosen subset will have to be able to infer true values for the set, allowing a good evaluation of all the information and furthermore, the optimal subset may not be unique.

Initially, a particular FS method would be chosen and then the results would be verified. However, it was thought that combining some FS methods instead of choosing only one, could be a great improvement.

Thus, the binding of these models can be combined using ensemble FS approaches, making it easier to cope with the problem mentioned above, of finding a proper subset in the full dataset.

With that, using ensembles methods in feature selection can show better results even for small samples, becoming an integral part of machine learning. A ensemble includes algorithms performing the task of prediction for a given input presented as a set of measurable characteristics, often called features.[5]

An example of a technique that includes an ensemble feature selection is the Random Forest that is a popular choice in the machine learning. RF has a great prediction accuracy for many types of data, is based on a collection of decision trees and can be summarized as three important features:

- offers accurate predictions on many types of applications;
- can use model training to measure the importance of each feature;
- trained model can measure the pairwise proximity between samples.

[6]

V. Upcoming Domains

A. Single Nucleotide Polymorphism Analysis

A single-nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a unique nucleotide in the genome/ shared sequence (A, T, C or G) differs between members of a species or paired chromosomes in an individual. Basically, SNP consists in mutations at a single nucleotide position that occurred during evolution and were passed on through heredity. [2]

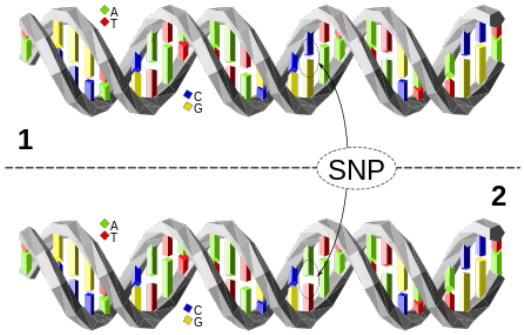

Fig. 1. SNP - Difference in the upper and lower DNA molecule at a single base-pair location

In bioinformatics, genetic association studies test a correlation between disease status and genetic variation with the aim of finding candidate genes (or genome regions) that contribute to a certain disease. Given the fact that SNPS are like 7 million in the human genome, are therefore an important part of this study which also focuses in computational methods for haplotype SNPSs (htSNP), which have several approaches to verify these htSNPs.

B. Text And Literature Mining

Text and literature mining is a new field with great research potential in bioinformatics and data mining, that consists of information extraction applied to biomedical literature. [7]

The text mining may consist essentially in three parts:

- Information retrieval - retrieve relevant documents from large collections (like Corpora) and use query. Example: “which documents refer to antibody A and protein P?”
- Information extraction - extract information from unstructured data to meaningful information, such as entities and relationships. Example: “the antibody A binds to protein P”
- Question answering - Synthesize a coherent answer, scan large document and analyse the question to determine what kind it is. Example: “which antibodies bind to protein P?”.

[8]

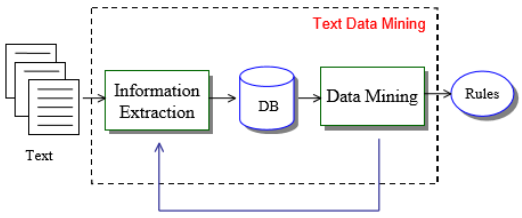

Fig. 2. Overview of IE-based text mining framework

One important representation of text is Bag of Words (BOW) which consists in variables that represents each word

in the text and a value that is the frequency of each word in the text causing the existence of high dimensional datasets and therefore the use of feature selection techniques.

The development of applications in this area are essentially related to the information extraction and use of text mining methods to extract information related to biological processes and diseases, expecting that these can be used for reseachers in this area. [7]

VI. Software Packages

There are some software packages that implements feature selection methods:

- General purpose: WEKA, MLC++, Fast Correlation Based Filter;
- Microarray analysis: SAM, PCP, GALGO, Nudge, Qvalue;
- Mass Spectra analysis: GA-KNN, R-SVM;
- SNP analysis: CHOISS, WCLUSTAG.

[9]

VII. Conclusion

In this paper, we demonstrated the feature selection techniques and their effectiveness in the two main problems already described: the large input dimensionality and the small samples sizes. We explain various types of FS methods in order to find the proper method to a particular problem. Furthermore, we find that this is an area still under study, giving prevalence to univariate methods and increasing variations of multivariate selection algorithms, making this development a promising future in bioinformatics. To conclude, the feature selection techniques have a great power in bioinformatics and there will still be a lot of research, especially in SNPs, text and literature mining and the combination of heterogeneous data sources.

References

[1] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *Journal of Computing*, vol. 1, no. 1, pp. 7–11, 2009.

[2] i.-P. P. Chen, *Current Bioinformatics*. Bentham Science, 2016.

[3] X. Zhou and K. Mao, "The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms," *School of Electrical Electronic Engineering and Bioinformatics Research Centre*, vol. 22, no. 20, p. 2507–2515, 2005.

[4] U. M. Sima, Chao; Braga-Neto and E. R. Dougherty, "High-dimensional bolstered error estimation," *1Computational Biology Division, Department of Electrical and Computer Engineering and Department of Pathology, University of Texas*, vol. 27, no. 21, p. 3056–3064, 2011.

[5] O. Okun, *Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations*. Springer, 2011.

[6] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Springer, 2012.

[7] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," *Department of Computer Sciences University of Texas at Austin*, vol. 7, no. 1, pp. 3–10, 2013.

[8] H.-H. Hsu, *Advanced Data Mining Technologies in Bioinformatics*. Idea Group, 2006.

[9] I. Saeys, Yvan ; Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Department of Plant Systems Biology, Department of Molecular Genetics, Department of Computer Science and Artificial Intelligence*, vol. 23, no. 19, p. 2507–2517, 2007.