Relative Relations in Biomedical Data Classification

Marcin Czajkowski
Bialystok University of Technology, Poland

INTRODUCTION

Advances in data science continue to improve the precision of biomedical research, and machine learning solutions are increasingly enabling the integration and exploration of molecular data (Huang, Chaudhary & Garmire, 2017). To enable a better understanding of cancer and enhance advances in personalized medicine such data need to be converted to knowledge. An interdisciplinary subfield of computer science called data mining (Han, Kamber & Pei, 2012) aims to reveal important and insightful information hidden in data. It requires appropriate tools and algorithms to effectively identify correlations and patterns within the data. However, the overwhelming majority of systems focus almost exclusively on the prediction accuracy of core data mining tasks like classification or regression. Far less effort has gone into the crucial task of extracting meaningful patterns or molecular signatures of biological processes.

Recently, there is a strong need for "white box", comprehensive machine learning models which may actually reveal and evaluate patterns that have diagnostic or prognostic value in biomedical data (McDermott & Wang, 2013). In this chapter, we focus on algorithms for biomedical analysis in the field of eXplainable Artificial Intelligence (XAI) (Angelov, Soares, et. al., 2021). In particular, we present computational methods that address the concept of Relative Expression Analysis (RXA) (Eddy, Sung, et. al., 2010). The algorithms that are based on this idea access the interactions among genes/molecules to study their relative expression, i.e., the ordering among the expression values, rather than their absolute expression values. One then searches for characteristic perturbations in this ordering from one phenotype to another. The simplest form of such an interaction is the ordering of expression among two genes, in which case one seeks to identify typical reversals' pairs of genes (ordering is usually present in one phenotype and rarely present in the other). Such pairs of genes can be viewed as ''biological switches'' which can be directly related to regulatory "motifs" or other properties of transcriptional networks. The classification algorithms based on RXA are often data-driven and due to the comparison between feature relative expression levels within the same sample, the predictor is robust to inter- and intra-platforms variabilities as well as complex analytical and data processing methods like normalization and standardization procedures.

The purpose of this chapter is to illustrate the concept of RXA and the innovations that the use of relative relationship-based algorithms brings. We will also cover the issues and challenges of biomedical data analysis.

DOI: 10.4018/978-1-7998-9220-5.ch161

BACKGROUND

Data mining is an umbrella term covering a broad range of tools and techniques for extracting hidden knowledge from large quantities of data. Biomedical data can be very challenging due to the enormous dimensionality, biological and experimental noise as well as other perturbations. In the literature, we will find that nearly all standard, off-the-shelf techniques were initially designed for other purposes than omics data (Bacardit, Widera, et. al. 2014), such as neural networks, random forests, SVMs, and linear discriminant analysis. When applied for omics data, the prediction models usually involve nonlinear functions of hundreds or thousands of features, many parameters, and are therefore constrain the process of uncovering new biological understanding that, after all, is the ultimate goal of data-driven biology. Deep learning approaches have also been getting attention (Min, Lee & Yoon, 2016) as they can better recognize complex features through representation learning with multiple layers and can facilitate the integrative analysis by effectively addressing the challenges discussed above. However, we know very little about how such results are derived internally. Such lack of knowledge discovery itself in those 'black box' systems impedes biological understanding and are obstacles to mature applications.

In contrast to data mining systems, statistical methods for analyzing high-dimensional biomolecular data generated with high-throughput technologies permeate the literature in computational biology. Those analyses have uncovered a great deal of information about biological processes (Zhao, Shi, et. al., 2015) such as important mutations and lists of "marker genes" associated with common diseases and key interactions in transcriptional regulation. Statistical methods can enhance our understanding by detecting the presence of disease (e.g., "tumor" vs "normal"), discriminating among cancer subtypes (e.g., "GIST" vs "LMS" or "BRCA1 mutation" vs "no BRCA1 mutation") and predicting clinical outcomes (e.g., "poor prognosis" vs "good prognosis"). The statistical analysis is often based on a relatively small number of features thus a small set of informative variables needs to be identified out of a large number (or dimension) of candidates. Therefore, using popular variable selection methods like LASSO and/or dimension reduction methods like PCA is crucial but still, it may limit the prediction model performance.

High-throughput measurements in cell biology (e.g., transcriptomics, proteomics, metabolomics) generate an enormous amount of information, but only implicitly, in the form of raw e.g., expression values. Extracting such knowledge from this single or multi-omics data is a key to future success in the biomarker field (Huang, Chaudhary & Garmire, 2017). However, there stand several challenges that computational approaches need to face. First of all, the dimensions of the dataset can grow into hundreds or thousands of variables, while the number of observations or biological samples remains limited. This disparity is called the curse of dimensionality or the $p >> n$ problem, where p is the number of variables and n is the number of samples. Moreover, missing values and the class imbalance in the data can also lead to results that are biased or less accurate. A class imbalance problem arises when rare events are analyzed and compared against events that happen much more frequently, a common occurrence in omics datasets. Furthermore, standard data mining solutions may not be suitable e.g. large-scale multi-omics analysis due to computational limitations. Finally, studied data often inherent biological and experimental errors and/or rely on capturing a snapshot of complex and dynamic biological systems. Consequently, an incorrect experimental design, erroneous integrational analysis, or randomness may lead to high noise and the risk of wrong scientific conclusions due to false-positive results.