Computational and Structural Biotechnology Journal 13 (2015) 8–17

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Machine learning applications in cancer prognosis and prediction

Konstantina Kourou [a], Themis P. Exarchos [a,b], Konstantinos P. Exarchos [a], Michalis V. Karamouzis [c], Dimitrios I. Fotiadis [a,b,*]

[a] Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
[b] IMBB — FORTH, Dept. of Biomedical Research, Ioannina, Greece
[c] Molecular Oncology Unit, Department of Biological Chemistry, Medical School, University of Athens, Athens, Greece

ARTICLE INFO

Available online 15 November 2014

Keywords:
Machine learning
Cancer susceptibility
Predictive models
Cancer recurrence
Cancer survival

ABSTRACT

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. In this work, we present a review of recent ML approaches employed in the modeling of cancer progression. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples. Given the growing trend on the application of ML methods in cancer research, we present here the most recent publications that employ these techniques as an aim to model cancer risk or patient outcomes.

Contents

Abbreviations: ML, Machine Learning; ANN, Artificial Neural Network; SVM, Support Vector Machine; DT, Decision Tree; BN, Bayesian Network; SSL, Semi-supervised Learning; TCGA, The Cancer Genome Atlas Research Network; HTT, High-throughput Technologies; OSCC, Oral Squamous Cell Carcinoma; CFS, Correlation based Feature Selection; AUC, Area Under Curve; ROC, Receiver Operating Characteristic; BCRSVM, Breast Cancer Support Vector Machine; PPI, Protein–Protein Interaction; GEO, Gene Expression Omnibus; LCS, Learning Classifying Systems; ES, Early Stopping algorithm; SEER, Surveillance, Epidemiology and End results Database; NSCLC, Non-small Cell Lung Cancer; NCI caArray, National Cancer Institute Array Data Management System.

* Corresponding author at: Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece.
E-mail addresses: konstadina.kourou@googlemail.com (K. Kourou), themis.exarchos@gmail.com (T.P. Exarchos), kexarcho@gmail.com (K.P. Exarchos), mkaramouz@med.uoa.gr (M.V. Karamouzis), fotiadis@cc.uoi.gr (D.I. Fotiadis).

http://dx.doi.org/10.1016/j.csbj.2014.11.005

1. Introduction

Over the past decades, a continuous evolution related to cancer research has been performed [1]. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

Given the significance of personalized medicine and the growing trend on the application of ML techniques, we here present a review of studies that make use of these methods regarding the cancer prediction and prognosis. In these studies prognostic and predictive features are considered which may be independent of a certain treatment or are integrated in order to guide therapy for cancer patients, respectively [2]. In addition, we discuss the types of ML methods being used, the types of data they integrate, the overall performance of each proposed scheme while we also discuss their pros and cons.

An obvious trend in the proposed works includes the integration of mixed data, such as clinical and genomic. However, a common problem that we noticed in several works is the lack of external validation or testing regarding the predictive performance of their models. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. Based on [3], the accuracy of cancer prediction outcome has significantly improved by 15%–20% the last years, with the application of ML techniques.

Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis and prognosis [4–7]. Specifically, these studies describe approaches related to the profiling of circulating miRNAs that have been proven a promising class for cancer detection and identification. However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed in [8,9]. These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed.

In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

2. ML techniques

ML, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference [10–12]. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an *n-dimensional* space for a given set of biological samples, using different techniques and algorithms [13]. There are two main common types of ML methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme/model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share.

Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malignant or benign based on its size. The ML question would be referred to the estimation of the probability that the tumor is malignant or no (1 = Yes, 0 = No). Fig. 1 depicts the classification process of a tumor being malignant or not. The circled records depict any misclassification of the type of a tumor produced by the procedure.

Another type of ML methods that have been widely applied is semi-supervised learning, which is a combination of supervised and unsupervised learning. It combines labeled and unlabeled data in order to construct an accurate learning model. Usually, this type of learning is used when there are more unlabeled datasets than labeled.

When applying a ML method, data samples constitute the basic components. Every sample is described with several features and every feature consists of different types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be used for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for ML. Data quality issues include the presence of noise, outliers, missing or duplicate data and data that is biased-unrepresentative. When improving the data quality, typically the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, preprocessing steps should be applied that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data preprocessing that focus on modifying the data for better fitting in a specific ML method. Among these techniques some of the most important approaches include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction. There are many benefits regarding the dimensionality reduction when the datasets have a large number of features. ML algorithms work better when the dimensionality is lower [14]. Additionally, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of fewer features. In general, the dimensionality reduction by selecting new features which are a subset of the old ones is known as feature selection. Three main approaches exist for feature selection namely embedded, filter and wrapper approaches [14]. In the case of feature extraction, a new set of features can be

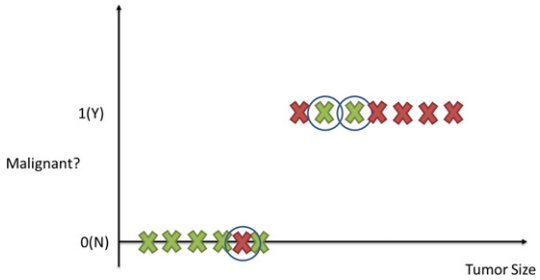

Fig. 1. Classification task in supervised learning. Tumors are represented as X and classified as benign or malignant. The circled examples depict those tumors that have been misclassified.