Machine learning applications in cancer prognosis and prediction

1. Introduction

2. ML techniques

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.

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.

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.

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.

Three main approaches exist for feature selection namely embedded, filter and wrapper approaches

The main objective of ML techniques is to produce a model which can be used to perform classification, prediction, estimation or any other similar task.

A good classification model should fit the training set well and accurately classify all the instances. If the test error rates of a model begin to increase even though the training error rates decrease then the phenomenon of model overfitting occurs. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias–variance decomposition.

The performance analysis of each proposed model is measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC).

Among the most commonly used methods for evaluating the performance of a classifier by splitting the initial labeled data into subsets are: (i) Holdout Method, (ii) Random Sampling, (iii) Cross-Validation and (iv) Bootstrap.

When the data are preprocessed and we have defined the kind of learning task, a list of ML methods including (i) ANNs, (ii) DTs, (iii) SVMs and (iv) BNs is available.

Even though ANNs serve as a gold standard method in several classification tasks [19] they suffer from certain drawbacks. Their generic layered structure proves to be time-consuming while it can lead to very poor performance. Additionally, this specific technique is characterized as a “black-box” technology. Trying to find out how it performs the classification process or why an ANN did not work is almost impossible to detect.

DTs follow a tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision out comes

Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes.

BN classifiers produce probability estimations rather than predictions.

3. ML and cancer prediction/prognosis

The success of a disease prognosis is undoubtedly dependent on the quality of a medical diagnosis; however, a prognostic prediction should take into account more than a simple diagnostic decision. When dealing with cancer prognosis/prediction one is concerned with three predictive tasks: (i) the prediction of cancer susceptibility (risk assessment), (ii) the prediction of cancer recurrence/local control and (iii) the prediction of cancer survival. In the first two cases one is trying to find (i) the likelihood of developing a type of cancer and (ii) the likelihood of redeveloping a type of cancer after complete or partial remission. In the last case, the prediction of a survival outcome such as disease

Specific or overall survival after cancer diagnosis or treatment is the main objective. The prediction of cancer outcome usually refers to the cases of (i) life expectancy, (ii) survivability, (iii) progression and (iv) treatment sensitivity

4. Survey of ML applications in cancer

The majority of these studies use different types of input data: genomic, clinical, histological, imaging, demographic, epidemiological data or combination of these.

4.1. Prediction of cancer susceptibility

4.2. Prediction of cancer recurrence

4.3. Prediction of cancer survival

5. Discussion

One of the most common limitations noted in the studies surveyed in this review is the small amount of data samples.

6. Conclusions

In this review, we discussed the concepts of ML while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.