Expert Systems With Applications 70 (2017) 1–19

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

A survey on computational intelligence approaches for predictive modeling in prostate cancer

Georgina Cosma[a,*], David Brown[a], Matthew Archer[a], Masood Khan[b], A. Graham Pockley[c]

[a] School of Science and Technology, Nottingham Trent University, Nottingham, UK
[b] Department of Urology, University Hospitals Leicester NHS Trust, Leicester, UK
[c] John van Geest Cancer Research Centre, Nottingham Trent University, Nottingham, UK

ARTICLE INFO

Article history:
Received 21 July 2016
Revised 2 November 2016
Accepted 3 November 2016
Available online 9 November 2016

Keywords:
Prostate cancer prediction
Predictive modeling
Computational intelligence
Machine learning
Soft computing
Disease classification
Evolutionary computation
Metaheuristic optimisation

ABSTRACT

Predictive modeling in medicine involves the development of computational models which are capable of analysing large amounts of data in order to predict healthcare outcomes for individual patients. Computational intelligence approaches are suitable when the data to be modelled are too complex for conventional statistical techniques to process quickly and efficiently. These advanced approaches are based on mathematical models that have been especially developed for dealing with the uncertainty and imprecision which is typically found in clinical and biological datasets. This paper provides a survey of recent work on computational intelligence approaches that have been applied to prostate cancer predictive modeling, and considers the challenges which need to be addressed. In particular, the paper considers a broad definition of computational intelligence which includes metaheuristic optimisation algorithms (also known as nature inspired algorithms), Artificial Neural Networks, Deep Learning, Fuzzy based approaches, and hybrids of these, as well as Bayesian based approaches, and Markov models. Metaheuristic optimisation approaches, such as the Ant Colony Optimisation, Particle Swarm Optimisation, and Artificial Immune Network have been utilised for optimising the performance of prostate cancer predictive models, and the suitability of these approaches are discussed.

1. Introduction

The increasing availability of electronic healthcare databases is enhancing opportunities for developing computer-based prediction and decision support models which can be used to improve the management of patients by healthcare professionals. An important challenge for clinical teams remains the prediction and assessment of risk, and the development of accurate approaches for diagnosing, and predicting the diagnosis and therapeutic responsiveness and outcomes. The aim of predictive modeling in the context of medicine involves the development of computational models which are capable of predicting future events and/or healthcare-related outcomes for patients using contemporarily-available healthcare data (Shariat, Kattan, Vickers, Karakiewicz, & Scardino, 2009b; Waljee, Higgins, & Singal, 2014). These models can be based on statistical techniques or computational intelligence techniques, with the latter being a relatively new strategy.

Computational intelligence approaches combine metaheuristic optimisation algorithms such as the Genetic Algorithm and Particle Swarm optimisation, with machine learning algorithms such as the Deep Learning, Support Vector Machines, Bayesian models, and hybrids of these (Mumford & Jain, 2009) for optimising the performance of prediction models. Machine learning algorithms have a fundamental role in predictive modeling, as they can be utilised to create the component which learns from existing patient data in order to be able to make predictions on new patient data. Take, for example, a model which has been developed to predict prostate cancer. Given a set of inputs (also called features, predictors, variables, observations) and a set of clinical results (also called targets), a model can be trained to learn the inputs of this dataset. Once the learning process is accomplished, then the model can accept new inputs and predict clinical outcomes.

Clinical prediction systems which consider a profile of variables for predicting an outcome require sophisticated computational methods (Tewari et al., 2001). Computational intelligence approaches have the capability to deal with the imprecision and uncertainty which is typically apparent in clinical and biological data. Furthermore, these approaches are effective when the data to be modelled are too large or complex for conventional statistical techniques. For example, computational intelligence algorithms

* Corresponding author.
E-mail addresses: georgina.cosma@ntu.ac.uk (G. Cosma), david.brown@ntu.ac.uk (D. Brown), m.archer1992@hotmail.co.uk (M. Archer), masood.khan@uhl-tr.nhs.uk (M. Khan), graham.pockley@ntu.ac.uk (A. Graham Pockley).

http://dx.doi.org/10.1016/j.eswa.2016.11.006
0957-4174/

have been used in risk prediction models for breast cancer (Bourdes et al., 2007; Hameed & Bagavandas, 2011), cardiovascular disease (Vijaya, Khanna Nehemiah, Kannan, & Bhuvaneswari, 2010), and lung cancer (Balachandran & Anitha, 2013; Dass, Rasheed, & Ali, 2014; Diaz, Pinon, & Solano, 2014; Kumar, Ramesh, Vanathi, & Gunavathi, 2011; Sun, Tan, & Zhang, 2008).

This paper provides a survey of recent work on prostate cancer predictive modeling using computational intelligence approaches, provides a broader perspective of the area, and considers challenges that remain to be addressed.

The paper is structured as follows: Section 2 introduces the process of creating a classifier, cross validation techniques, and evaluation of classification models. Section 3 discusses the recent advances in computational intelligence algorithms that have been developed for prostate cancer, particularly the application of Artificial Neural Networks, Deep Learning, Fuzzy approaches, Support Vector Machines, metaheuristic optimisation, Ensemble learning algorithms, and Bayesian approaches, including the Bayesian Network, and the Markov model. Section 4 discusses considerations for selecting a suitable metaheuristic optimisation method. Section 5 provides a discussion of advances, challenges and future areas of potential research. A conclusion and future directions are presented in Section 6.

2. Cancer classification and evaluation

Classification is the process of finding a model which is capable of distinguishing data records into classes. Prior to constructing a classification model, the dataset must be prepared using processes that may include the following:

- *Data normalisation* includes filling missing values, identifying and removing outliers, grouping variables, and normalising data.
- *Feature extraction* is the task of representing the original data in a reduced dimensional space. Feature extraction approaches are related to dimensionality reduction methods which include Principal Component Analysis (PCA) and Singular Value Decomposition (SVD).
- *Feature selection* involves selecting the most useful features without altering the original data representation, and hence choosing a subset of the features relevant for the task. In large datasets, metaheuristic optimisation approaches such as the Genetic Algorithm can be used for finding the best subset of features.

Whether a feature extraction or selection approach is required depends on the type of data and the task. For example, feature extraction approaches are commonly applied to tasks involving image processing, for which it is important to represent the original image in a reduced dimensional space from which noise has been removed. On the other hand, feature selection is important for clinical data when the names of the features are important, for example when identifying which symptoms are the best predictors of a cancer; or when trying to identify which combination of features ('biomarkers') would make up the optimum cancer 'fingerprint'.

A classification model can be constructed once the data preparation stage is completed. The process of constructing a classification model comprises of two main phases, as illustrated in Fig. 1:

1. *Learning phase* (or training phase) in which the data are analysed using a classification algorithm, and from which a classifier (i.e. learned model) is created. During the learning phase, the classification algorithm analyses a set of training data which contains data records comprising of a set of inputs (also called features, predictors, variables, observations) and their known class labels (also called targets, known outputs) in order to learn from the data and build the classifier. A class label is the outcome of an event, for example cancer stage after diagnosis. The class labels are only used during the training process in order to enable the classifier to reason with the inputs. It should be noted that, since the class labels are provided to the classifier during the learning process, this phase is also known as supervised learning. Supervised learning is different from unsupervised learning (or clustering), during which the class label of each data record is not known in advance. Once a classifier is trained, a classification model is derived, and that model is used to predict the class labels of data records for which the class labels are unknown. The classification model may be represented in the form of neural networks, classification rules, graphical models such as Bayesian networks, Markov random fields and decision trees, or as a mathematical/statistical formulae (Han, Kamber, & Pei, 2011). Fig. 1 provides an overview of the process of creating a classification model for predicting patient outputs (i.e. class labels); and
2. *Prediction phase* (or testing and validation) of the model is used for classification/prediction tasks. During this phase, the learned model classifies new unknown data (contained in test and validation datasets) which have not been previously seen by the model in order to evaluate the accuracy of the learned model. Initially, the predictive accuracy of the model is evaluated and, if it is acceptable, then the learned model can be used for prediction.

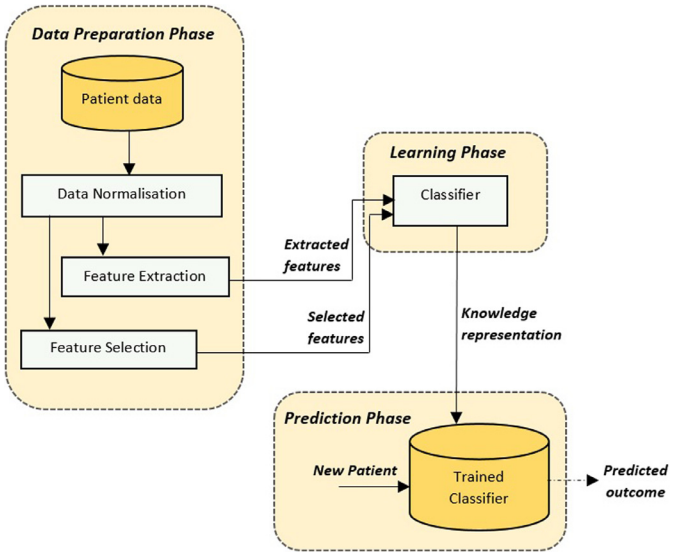

Fig. 1. Process of creating a predictive model: Initially the dataset is prepared and a classification model is constructed. At the *Learning Phase*, the classifier learns from inputs (e.g. patient data records) and known class labels (known result). At the *Prediction Phase*, the classifier takes as input a new previously unseen records and predicts their class labels.

Studies show that the accuracy of the classifiers depends on the specific dataset and the problem to be solved, and that there is no single classifier which outperforms all other classifiers across all tasks. Furthermore, the choice of validation method influences the reported accuracy of the classifier. The concept applies to clinical and biological data, and challenges related to sample size, especially when biological data is concerned (Swan, Mobasheri, Allaway, Liddell, & Bacardit, 2013).

There are a number of methods and metrics which can be used for evaluating the predictive performance of a classifier. Predictive models are most commonly evaluated using the cross-validation method discussed in Section 2.1, and a number of evaluation measures which are discussed in Section 2.2.