Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A survey of machine learning applications in HIV clinical research and care

Kuteesa R. Bisaso [a,b,c,*], Godwin T. Anguzu [a], Susan A. Karungi [b], Agnes Kiragga [a], Barbara Castelnuovo [a]

[a] *Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda*
[b] *Department of Pharmacology and Therapeutics, College of Health Sciences, Makerere University, Kampala, Uganda*
[c] *Breakthrough Analytics Limited, Kampala, Uganda*

ARTICLE INFO

Keywords:
HIV
Machine learning
Clinical research
Application paradigms

ABSTRACT

A wealth of genetic, demographic, clinical and biomarker data is collected from routine clinical care of HIV patients and exists in the form of medical records available among the medical care and research communities. Machine learning (ML) methods have the ability to identify and discover patterns in complex datasets and predict future outcomes of HIV treatment. We survey published studies that make use of ML techniques in HIV clinical research and care. An advanced search relevant to the use of ML in HIV research was conducted in the PubMed biomedical database. The survey outcomes of interest include data sources, ML techniques, ML tasks and ML application paradigms.

A growing trend in application of ML in HIV research was observed. The application paradigm has diversified to include practical clinical application, but statistical analysis remains the most dominant application. There is an increase in the use of genomic sources of data and high performance non-parametric ML methods with a focus on combating resistance to antiretroviral therapy (ART). There is need for improvement in collection of health records data and increased training in ML so as to translate ML research into clinical application in HIV management.

1. Introduction

Machine learning (ML) is generating enormous buzz and gaining importance in many domains. It has given birth to applications such as self-driving cars, speech and language recognition, optical character recognition, e-commerce online recommender systems, fraud detection, email filtering and most recently precision medicine. This trend in popularity of ML is driven by improvements in data collection and storage, and advancement in computing power (processing, memory and storage) over the past decade. Together, these two factors have spurred the use of computers to tackle increasingly complex tasks [1,2].

Machine learning provides computers with the ability to learn without being explicitly programmed. It uses complex algorithms and techniques to recognize patterns in data in order to make predictions. Over the past two decades, there has been a steady increase in the medical research involving ML, progressing dramatically from laboratory curiosity to a practical clinical application. This is largely attributed to growing volumes of clinical, social, epidemiological, genetic and other types of medical data that are overwhelming for humans to infer from and make decisions. Consequently, ML has been envisaged to improve medical practice through “better decision-making, optimized innovation, improved research/clinical trial efficiency, and new tool creation for physicians, consumers, insurers, and regulators” [3–6].

HIV/AIDS remains one of the world's most significant public health and developmental challenges. Despite tens of millions of AIDS related deaths since the beginning of the epidemic in 1981, approximately 36 million people current live with HIV. Approximately 19 million people with HIV are enrolled in routine care programs and receiving treatment [7]. The management of HIV is complicated by the wide variability in both host and viral genetic makeup, dozens of options of ART to chose from, multiple opportunistic infections, demographic differences in disease progression and response to ART as well as socio-cultural differences in acceptability and adherence to treatment, all of which necessitate personalization of care and treatment [8].

A wealth of demographic, clinical and biomarker data is collected from routine clinical care of HIV patients. This data exists in form of medical records available among the medical care and research communities. These large amounts of data make HIV research and treatment

[*] Corresponding author. Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda.
E-mail address: kuteesar@gmail.com (K.R. Bisaso).

https://doi.org/10.1016/j.compbiomed.2017.11.001
Received 7 February 2017; Received in revised form 1 November 2017; Accepted 1 November 2017

a potential beneficiary of ML. The extent of ML application in HIV research and treatment remains unclear. The purpose of this survey is to explore the nature and trends in ML application in HIV/AIDS clinical research and management.

2. Machine learning methods

The key goal of ML is to use an example dataset to map out the characteristics that are most helpful in predicting an outcome of interest, and apply those characteristics to accurately predict outcomes in a new situations not previously encountered [9]. This is referred to as prediction, a process made possible via Bayesian statistics which allows learning a probability distribution from data and utilization of inverse probability to infer the unknowns in future data [10].

A diverse array of machine-learning methods (models and algorithms) has been developed to tackle the wide nature of tasks. These methods are broadly classified into supervised and unsupervised learning [11].

Supervised learning methods search for a function *f(x) that* predicts a target/output variable *(y)* given a set of predictor/input variables*(x)*. The training data is called labeled data because it consists of (*x,y*) pairs of variables. The inputs *x* may be simple vectors or more complex objects such as texts, DNA sequences, molecular structures, images, graphs or videos. Outputs (or labels) may include continuous outcomes or the more common binary yes-or-no outcomes. Regression learning methods predict outcomes in a continuous spectrum while classification learning methods predict outcomes of a categorical (binary) nature [11]. Abundant research has been done on problems such as multiclass classification (where *y* takes on one of *more than 2* labels), multi-label classification (where *y* is labeled simultaneously by several of the *K* labels), ranking problems (where *y* provides a partial order on some set), and general structured prediction problems (where *y* is a combinatorial object such as a graph, whose components may be required to satisfy some set of constraints) [2].

A number of supervised learning methods have been developed. These include multiple linear regression (M-LR), decision trees and forests (DT, DF) [12], logistic regression (LR), support vector machines (SVM) [13–16], artificial neural networks (ANN) [15–17], bayesian classifiers (BC), classification and regression trees (CART) [18,19], K-nearest neighbors (KNN) among others [20]. Ensemble methods combine outputs of multiple independently trained weaker models to make an overall prediction. The selection of the combination of weaker learning methods is made in such a way as to maximize the prediction power of the ensemble algorithm. Ensemble methods include boosting, bootstrap aggregation (bagging), stacking/blending, random forests (RF) [16,21] and their modifications [22].

On the other hand, unsupervised learning involves the analysis of unlabeled (no distinction between input and output) data under assumptions about the structural properties of the data (e.g., algebraic, combinatorial, or probabilistic). Since there are no training examples used in this process, the learning algorithm aims to identify patterns and correlations in the given data. The main applications of these algorithms include clustering and dimensionality reduction. Dimensionality reduction algorithms, including principal components analysis, manifold learning, factor analysis, random projections, and autoencoders, identify and eliminate redundancies in the data so as to remain with only the variables that account for the most variability in the data. Clustering algorithms partition data into coherent clusters and determine the partitioning rule for predicting clusters in future data. The K-means clustering algorithm is the most commonly used method [20]. Computational complexity is a major concern in both clustering and dimension reduction since the datasets to exploit are large and unlabeled [2].

A third major ML paradigm is semi-supervised learning. Here, the data is a mixture of small amounts of labeled and large amounts of unlabeled training data. The algorithm learns the structures of the data from the labeled examples and makes assumptions about the unlabeled data in order to make predictions. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process [8]. Semi-supervised learning is subclassified into inductive learning whereby the goal is to learn from both the labeled and unlabeled dataset to predict labels for future datasets and transductive learning whereby the goal is to predict labels for the unlabeled portion of the data [23].

A more recent category of ML called reinforcement learning involves the algorithm discovering actions that yield the greatest rewards through trial and error. The algorithm is trained to choose actions that maximize reward. It is said to learn from past experiences and capture knowledge to make accurate decisions [24]. The most common example is the markov decision process [25].

In practice, current research and application blends unsupervised and supervised categories of ML. The choice of ML method to use is guided by the objectives of the analysis and the data available. Important data considerations include the number of predictor variables/features available in the data and quality of data. In general, a small but informative feature space results in higher generalizability of the model and avoids overfitting [26] while improving data quality and greatly improves the analysis. Therefore data preprocessing and feature selection, in often using the unsupervised learning methods, are often the initial steps in an ML analysis. The Feature selection process optimizes information gain and minimizes overfitting. Whereas some supervised learning algorithms (SVM) offer the advantage of sparsity and inherently select the most predictive features, explicit feature selection methods exist. These are broadly classified into filter methods (e.g. correlation coefficient scores, chi square tests, information gain), wrapper methods (e.g. step-wise covariate modeling, recursive feature elimination) and embedded methods (regularization algorithms, LASSO, elastic net and ridge regression) [27,28]. The extremely high dimensionality of biological data such as protein and peptide sequences exposes the inadequacy of the above feature selection methods during implementation of ML based approaches. Hence the need for methods that reduce the dimensionality of features so as to decrease computational running times while increasing classification accuracy. For this purpose, feature encoding techniques, which map original representations into new spaces, have been developed. This mapping makes separation of classes easier by condensing complex data patterns into fewer, easily manageable and statistically significant forms which makes the subsequent classification step easier and more accurate. A number of feature extraction methods have been developed and evaluated. These include the orthonormal encoding (OE), frequency based encoding (FE), Taylor's venn-diagram (TVD), residual couple encoding (RC) and a combination of OE and TVD [29,30].

Model training and validation may be done concurrently. A model is trained on one dataset (training set) and its prediction performance tested on another (test set). Prediction performance is a measure of how well a method is able to give correct prediction on unseen data. Model validation techniques include the holdout method, N-fold-cross-validation and bootstrap. With the holdout method, the data is conventionally split into a training (for model building) and test set (for performance evaluation) in a ratio of 2:1 respectively. In comparison, the N-fold-cross-validation method randomly splits the data into k subsets where the k-1 sets of the data are used to train the model while the kth set is used to assess the model's accuracy of prediction. The bootstrap method samples with replacement from the dataset to create n new datasets (bootstrap replicates). The replicates are used to test the model's predictive ability. In both N-fold-cross-validation and bootstrap, the accuracy measures are calculated as the average of all different validation cycles or bootstrap replicates [11,31].

Model performance is gauged using measures of accuracy e.g. root-mean-squared-error (RMSE), mean absolute error (MAR) and percentage prediction error (PPE) for regression with continuous numerical outcomes. With the classification of categorical outcomes, the percentage of correctly predicted observations, sensitivity and specificity, false