Using Machine Learning Algorithms to Improve the Prediction Accuracy in Disease Identification: An Empirical Example

Yong Cai, Dong Dai and Shaowen Hua

Abstract— **In the field of medicine, many diseases are difficult to diagnose. Patients may be free of symptoms for a long time before being discovered illness. This delay can lead to either life threatening incidence or high cost treatment such as liver transplant. Fortunately, technology development enables us to capture and store larger amount of diagnosis and treatment information. Also, various new machine learning algorithms have been developed and shown excellent performance in many circumstances. We use an empirical example to demonstrate how machine learning algorithms can help to improve prediction accuracy in identifying primary biliary cholangitis (PBC) patients. The models have been developed on real world patient diagnosis and treatment history data to predict PBC indication. We find in the example that Random Forest has outperformed AdaBoost or generalized linear models such as Logistic Regression and LASSO. By cross-validation, Random Forest has 0.942 compared to benchmark Logistic Regression 0.759 in prediction accuracy. This example shows about 24% of accuracy improvement over traditional method by selecting an appropriate algorithm. The result provides extra data point to cross-validate the claim that Random Forest is one of the general purposed algorithms that shall be considered in the analysis toolkit.**

Index Terms— **machine learning, Random Forest, Adaboost, algorithm comparison**

I. Introduction

Machine learning (ML) has been gaining lots of popularity in the recent decade. It's evolved from the field of artificial intelligence such as pattern recognition and computational learning. The recent popularity in ML is largely due to fast gaining computational power and ability of cost-effectively collecting and storing large amount of data. ML has been applied in many fields such as fraud detection, image and voice recognition etc. In this paper, we explore using machine learning algorithms to improve the prediction accuracy of disease identification. In an empirical example, we build predictive models to find primary biliary cholangitis (PBC) patients who haven't been diagnosed yet or miss PBC diagnose code.

PBC is a chronic, slow progressive disease that destroys the medium sized bile ducts in the liver [1], [2]. It is also known as primary biliary cirrhosis. PBC is primarily resulted from autoimmune destruction of the bile ducts. This causes bile acids (BA) to remain in the liver, where persistent toxics build up, and eventually damages the liver cell and causes cirrhosis. Disease progression in PBC is rather slow. Many patients with PBC may be completely free of symptoms, especially in the early stage of the disease. Because of this, the disease is often discovered through abnormal results on routine blood tests. In the United States, PBC is the one of the top ten causes of liver transplant [3], [4].

PBC is a rare disease. Boostra et. al. [5] did a systematic review by search 2286 abstracts. They concluded prevalence rates range from 19.1 to 402 per million. The prevalence and incidence rates vary greatly by geography. But the literatures they selected are based on small sample size ranging from 10 to 770 patients. Given the small observational sample, it is difficult to conduct epidemiology research at subnational level, such as finding the regional difference of prevalence and incidence. Predictive models can help identify more patients and derive a larger patient sample for further healthcare related research. In such cases, it is important to have a model provides accurate predictions to reduce the study bias.

Using patient age, gender, treatment pathway and dosage information in the data, we predict the likelihood that a patient has PBC disease. We applied Random Forest, Adaboost, LASSO, Naive Bayes algorithms as well as benchmark logistic regression model to test prediction accuracy.

The remaining of the paper will be arranged in following ways. In the next section, we describe the patient treatment and diagnosis database. In section 3, variable construction and modeling methodologies are presented. The following section applies and compares selected machine learning algorithms for predicting PBC patients. In the last section, the results and empirical findings are discussed.

II. Data Source Description

We extract the PBC training data from IMS medical claims database and longitudinal prescription (Rx) database. The claims data have diverse representation of geography, employers, payers, providers and therapy areas, with coverage of data from 90% of US hospitals, 80% of all US doctors. The data are also longitudinal, with 4 or more years of continuous enrollment.

The Rx data contains patient longitudinal prescription history. It is derived from electronic data received from pharmacies, payers, software providers and transactional clearinghouses. This information represents activities that take place during the prescription transaction and contains information regarding the product, provider, payer and geography. The Rx data is longitudinally linked back to an anonymous patient token and can be linked to events within the data set itself and across other patient data assets. Common attributes

Yong Cai is Director with IMS Health, Plymouth Meeting, PA 19462, USA (email: ycai@imscg.com).

Dong Dai is Senior Manger with IMS Health, Plymouth Meeting, PA 19462, USA (email: ddai@us.imshealth.com).

Shaowen Hua is Assistant Professor with School of Business, La Salle University, Philadelphia, PA, USA (email: hua@lasalle.edu).

and metrics within the Rx data include payer, payer types, product information, age, gender, 3-digit zip as well as the scripts relevant information including date of service, refill number, quantity dispensed and day supply. The Rx data covers up to 88% for the retail channel, 48% for traditional mail order, and 40% for specialty mail order.

All data are anonymous and HIPAA compliant to protect patient privacy.

III. Methodology

In the medical claims database, PBC can be identified by diagnosis ICD-9 code 577.6. However, medical claims database has limited longitudinal coverage or many PBC patients haven't been diagnosed yet. Our goal is to construct a training data including patients with complete diagnosis information and build a predictive model using patient past treatment history to identify PBC. Later on, we can use the model results to score and identify patients in the larger prescription database that doesn't contain ICD-9 flag.

Because Ursodiol is the only FDA approved drug to treat PBC, we restrict the study patients to Ursodiol prescribers only, in this way we capture the majority of the PBC patients who are not with 577.6 ICD-9 codes by only losing those who don't have either 577.6 ICD-9 codes or Ursodiol prescriptions.

A. Training Data

To build the training data, we select patients from medical claims database who have complete history of diagnosis and prescriptions. For those Ursodiol patients, we are able to identify whether they are PBC patients (with ICD-9 code 577.6) or not. The Ursodiol prescribers are selected if they have any prescription of Ursodiol during the one year selection window from August 2013 to July 2014. Index date is defined as their earliest date of prescribing Ursodiol in the selection window. The lookback period is one year back from the day before index date. The PBC indication (Y/N) is derived as whether the patient has ICD-9 code 577.6 during the lookback period. A total sample of 48,472 Ursidiol prescribers are included from claims database, among which there are 12,700 (26%) patients are with PBC disease.

B. Rx predictors for PBC

Since we are going to implement the predictive model trained from claims database to prescription (Rx) data, the predictors are derived only based on prescription activities, no diagnosis history will be included for prediction.

There are three types of predictors: 1) whether (Y/N) the patient has the treatment and 2) the number of fills of the prescriptions during the lookback period for a list of drug of interest, 3) other predictors include such as age, gender, Ursodiol's (on the index date) strength, dose (=strength x quantity/days supply), NDC (National Drug Code) and product name, specialty of physician who prescribed the Ursodiol on the index date. In total, we have 62 predictors.

TABLE I
Confusion Matrix

	Actual (Yes)	Actual (No)
Predicted (Yes)	True Positive (TP)	False Positive (FP)
Predicted (No)	False Negative (FN)	True Negative (TN)

TABLE II
Model Performance Metrics

Metric	Definition
AUC	Area under the ROC curve
Sensitivity	TP/(TP+FN)
Specificity	TN/(FP+TN)
Positive predictive value (PPV)	TP/(TP+FP)
Negative predictive value (NPV)	TN/(FN+TN)
Accuracy	(TP+TN)/(TP+FP+FN+TN)

C. Model Performance Evaluation

For binary classification problem, we have the confusion matrix (Table I) where "Yes" denotes positive class and "No" denotes negative class with any given machine learning algorithm.

With the confusion matrix, we use the following metrics (Table II) to evaluate model performance of predicting PBC with the derived Rx predictors, where receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied and ROC curve is created by plotting the true positive rate (TPR=TP/(TP+FN)) against the false positive rate (FPR=FP/(FP+TN)) at various threshold settings. And the area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

D. Machine Learning Algorithms

For this binary classification (whether the patient has PBC disease or not) problem, Logistic Regression is applied as benchmark of model performance. The machine learning algorithms we consider in this paper include LASSO [6], Random Forest (also denoted by "RF" hereafter, [7]), Naive Bayes (also denoted by "NB" hereafter, [8]) and AdaBoost (also denoted by "AdaB" hereafter, [9]). We split the data into 80% (N=38,778) for training and 20% (N=9,694) for testing. Firstly, on the training dataset, we performed ten-fold cross-validation with all the machine learning algorithms to compare model performance and to choose a best algorithm; secondly, we train a predictive model by applying the best algorithm to the whole training dataset; lastly, we apply the predictive model to classify those patients prescribing Ursoldiol and without ICD-9 Code 577.6, where the cut-off probability is determined from training dataset to maximize the sum of Sensitivity and Specificity (graphically, with respective to the point closest to (0,1) on the ROC curve).

IV. Results and Conclusion

For all five algorithms, the averaged performance metrics (Table III, where cut-off probabilities are values that maximizing the sum of Specificity and Sensitivity on each training fold) and ROC curve (Figure 1) averaged (vertically) over ten-fold cross-validation on the training dataset is as following:

TABLE III
Model Performance (Ten-fold Cross-Validation)

	AUC	Sensitivity	Specificity
Logit	0.835	0.802	0.744
LASSO	0.836	0.793	0.752
NB	0.814	0.796	0.702
RF	0.983	0.939	0.943
AdaB	0.812	0.750	0.775

	PPV	NPV	Accuracy
Logit	0.527	0.914	0.759
LASSO	0.531	0.911	0.762
NB	0.487	0.906	0.727
RF	0.855	0.978	0.942
AdaB	0.542	0.897	0.768

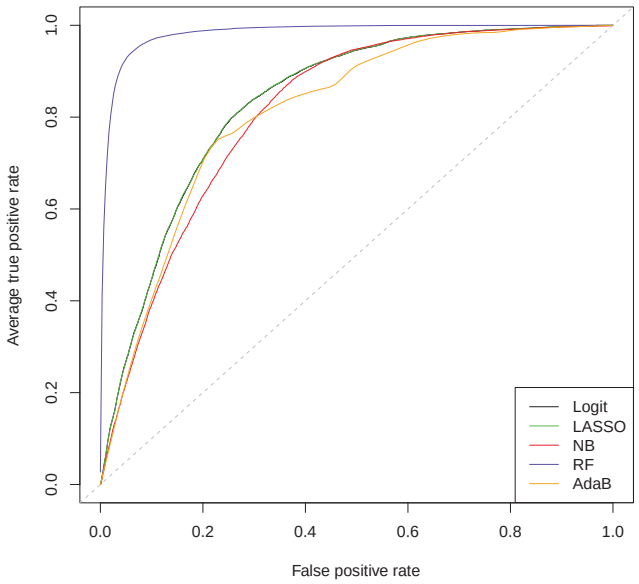

Fig. 1. Ten-fold Cross-Validation ROC curve

We can see that the Random Forest is the best one among all five algorithms, with an average AUC of 0.983 while that of Logistic Regression is 0.835, and other three algorithms of Naive Bayes, AdaBoost and LASSO are all with AUC less than 0.84. Note that on the above ROC curve figure, LASSO (green line) and Logistic Regression (black line) are almost overlapped, meaning that the variable selection process of LASSO mostly tends to select all the predictors resulting the similar performance as Logistic Regression, in other words, nearly each feature is contributing to the PBC disease prediction. And the significant performance improvement of Random Forest over Logistic Regression indicates the interactions among predictors, and in fact, for each treatment, the number of fills and flag (Y/N) are highly correlated, and the dose and strength are highly correlated to product name and NDC.

TABLE IV
Confusion Matrix and Performance Metrics of Classification on Testing Dataset

	Actual (Yes)	Actual (No)
Predicted (Yes)	True Positive (TP) $N = 2,321$	False Positive (FP) $N = 356$
Predicted (No)	False Negative (FN) $N = 219$	True Negative (TN) $N = 6,798$

Sensitivity	Specificity	PPV	NPV	Accuracy
0.914	0.950	0.867	0.969	0.941

Since Random Forest has the best performance among all five algorithms, we use it to train a predictive model with all the training dataset, then we use the cut-off probability that maximize the sum of Specificity and Sensitivity as cut-off value to classify whether a patient has PBC or not in the testing dataset. The confusion matrix ('Yes'=with PBC, 'No'=without PBC) and a list of performance metrics are as Table IV.

We also find that the performance of Random Forest on the testing dataset is consistent with the results of cross-validation on training dataset.

Our findings are also consistent with that of Caruana and Niculescu-Mizil (2006) [10] where Random Forest was one of the highest performers in their empirical example. As a general purpose ensemble method, Random Forest can pick complex variable dependencies and account for variable interactions. By averaging multiple decision trees and training on different parts of within the same training data, Random Forest was able to utilize the complex patient variables to a better extent and yield excellent accuracy (0.941) in cross-validation. Through the PBC empirical example, we provide one more data point to show that Random Forest is one of the most accurate general purposed learning techniques [11].

Acknowledgment

References

[1] R. Poupon, "Primary biliary cirrhosis: a 2010 update." *Journal of hepatology*, vol. 52, no. 5, pp. 745–758, 2010.
[2] G. M. Hirschfield and M. E. Gershwin, "The immunobiology and pathophysiology of primary biliary cirrhosis." *Annual review of pathology: Mechanisms of disease*, no. 8, pp. 303–330.
[3] K. D. Lindor, M. E. Gershwin, R. Poupon, M. Kaplan, N. V. Bergasa, and E. J. Heathcote, "Primary biliary cirrhosis." *Hepatology*, vol. 50, no. 1, pp. 291–308, 2009.
[4] C. Selmi, C. L. Bowlus, M. E. Gershwin, and R. L. Coppel, "Primary biliary cirrhosis." *The Lancet*, vol. 377, no. 9777, pp. 1600–1609, 2011.
[5] K. Boonstra, U. Beuers, and C. Y. Ponsioen, "Epidemiology of primary sclerosing cholangitis and primary biliary cirrhosis: a systematic review." *Journal of hepatology*, vol. 56, no. 5, pp. 1181–1188, 2012.
[6] R. Tibshirani, "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society., Series B (Methodological)*, pp. 267–288, 1996.

[7] L. Breiman, "Random forests." *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[8] T. Mitchell, "Generative and discriminative classifiers: naive bayes and logistic regression," 2005. [Online]. Available: https://www.cs.cmu.edu/ tom/mlbook/NBayesLogReg.pdf
[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[10] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms." in *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168.
[11] R. Genuer, J. M. Poggi, and C. Tuleau, "Random forests: some methodological insights." *arXiv preprint*, 2008.