Predictive Modeling for Refractory Epilepsy Population

Huang Li

***Abstract*—Drug resistant epilepsy (DRE), whose population is also known as refractory epilepsy patients, is a major clinical and societal problem for epilepsy patients. In this project, we define refractory epilepsy population and non-refractory population, extract features from clinical record data and implement predictive model for identifying the refractory population in an early stage.**

I. Introduction and motivation

Drug-resistant epilepsy (DRE) is a major clinical and societal problem for one in three epilepsy patients. Patients with epilepsy are experiencing inadequate control of seizures even with anti-epilepsy drugs (AED) therapy. We call this drug-resistant epilepsy population as refractory epilepsy patients. The reason of the occurrence of the refractory population can be quite diverse and have not been studied thoroughly in clinical domain. However, we can make use of the patients' previous clinical data to predict the DRE in a early time stage for the patient. The problem can be solved by different machine learning classification algorithms and a predictive model. Thus, in this project we extract refractory epilepsy population and non-refractory population from a cohort of epilepsy patients, and implement predictive model for identifying the refractory population in an early stage.

II. literature search

There are many studies concerning the factor of DRE inside the population of refractory epilepsy patients. [1] held an observation study in patients with epilepsy and analyzed factors that influence the seizure outcome in anti-epileptic drug treatment of epilepsy. [3] studied the response of newly diagnosed epilepsy patient to anti-epileptic drugs to identify the factors that associated with subsequent poor control of seizures. [4] defined the role of hypotheses in the clinical spectrum of drug-resistant epilepsy by reviewing the laboratory and clinical evidence supporting the drug transport and the drug-target hypotheses. Particularly, in [2] and [5], the risk factors of DRE and AED resistance have been extracted. [5] identified risk factors for development of DRE using an adult cohort of patients with generalized epilepsy. [2] recognized the risk factors for AED resistance in a cohort of patients with focal epilepsies by making classification of the grade of AED resistance, and defined "new AEDs" to estimate their helpfulness.

III. Problem formulation

This problem is a predictive modeling problem where we aim to classify the refractory epilepsy patients as early as possible using different machine learning classification algorithms. We call this early stage time point as index date, and we define the index date as the date when patient has been given the first AED. We generate an observation window before the index date to filter data and avoid bias. Then, we define the refractory epilepsy population as epilepsy patients who failed 4 or more AED, and non-refractory epilepsy population as patients who never failed a single AED (only one AED given). In this process, we regards an addition of AED as a failure of the previous AED. Next, we extract features base on clinical record of the patients. Then we use classification machine learning algorithms to create the predictive model and use accuracy, AUC, precision, recall, F1 score, ROC curve to evaluate the model.

IV. Approach and implementation

In this section, we introduce the approaches we used in all steps of solving the problem. Hadoop pig and Amazon AWS are used when implementing below approaches and doing experiments.

A. Data filtering

In data filtering phase, we extract patients who are epilepsy patients from the dataset. Also, the patients extracted would be given at least one AED in the record. We use three criteria to filter the cohort. First, we make use of the ICD-9 code to look at the patients' diagnostic events. We filter patients who have at least one 345.* (epilepsy diagnostic events) or at least two 780.39 (other convulsions) at any time in the data. Next, we gather 23 known AED to filter patients who al least have one given AED, and we find the first AED given date as the index date for each patient. Then we extract all the patients who is older than 16 at index date to focus on the adult group. In the former section we define the refractory epilepsy population as epilepsy patients who failed 4 or more AEDs, and non-refractory epilepsy population as patients who never failed a single AED (only one AED given). We use this method to generate case group as refractory epilepsy patients and control group as non-refractory epilepsy patients. The flow chart of the above steps can be shown in fig 1.

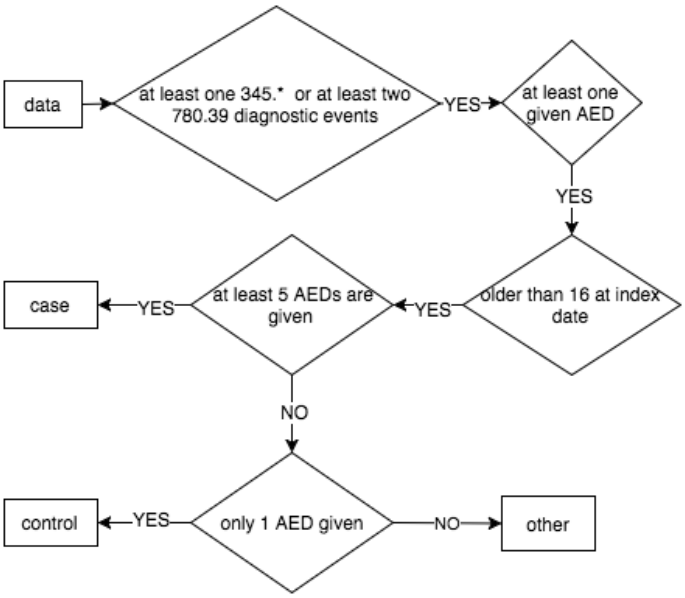

Fig. 1. Flow chart of data filtering

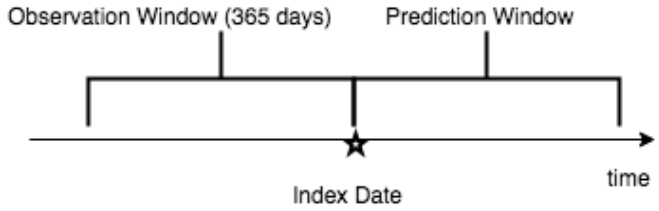

Fig. 2. Observation window and prediction window

B. Feature construction

In order to construct the feature, we first create an observation window which starts from one year before index date and ends at index date to avoid bias. We only consider the data inside the observation window to create the features. The observation can be shown in fig 2.

For prediction, we did not bound the prediction window, because we want to know whether the patient will eventually becomes refractory, which means whether patient failed 4 or more AEDs.

The clinical record data is used to construct features. It should be noted that the data contains 4 fields include:

1) PATIENT-ID - A number representing the patient ID.
2) EVENT - A name of the medical event. It consists of some prefix describing which category the event belongs to, and event name such as some medical code.
3) DATE - A date of the event in the format of 'yyyy-MM-dd'. This date is artificially spun. Do not be confused with the dates from the future.
4) VALUE - A possible value from the event. Some events have a meaningful value such as days of supply for a drug. On the other hand, some events have a meaningless value such as 1.0 for diagnosis code.

Also noted that the event field contains YOB, GENDER, INPATIENT, OUTPATIENT, ER, HOSPDAYS, DIAG, PROC, MOLECULE, CLASS, SUPPLYDAYS, QUANTITY. We are not going to elaborate on the meaning of each field here and we only describe how we generate the features.

We count the number of each patient's INPATIENT event (patient encounters event as an inpatient), OUTPATIENT event (patient encounters event as an outpatient), ER event (patient is admitted through emergency room), DIAG event (patient is diagnosed with an ICD-9 code), PROC event (patient is provided a procedure of a code), MOLECULE event (patient is given a drug), CLASS event (the class of the drug given to patient), and we sum the event VALUE for each patient's GENDER event (VALUE is 1.0 for male and 0.0 for female), HOSPDAYS event (The patient stayed in a hospital for the VALUE days), SUPPLYDAYS event (patient is given a drug for VALUE days of supply), QUANTITY event (patient is given a drug with VALUE of dosage). Note that DIAG, PROC, MOLECULE, CLASS, CLASS, SUPPLYDAYS, QUANTITY events are not single event, they contains different code for drug, procedure, diagnostic, class, etc. We only compute the count and summation on distinct event for each patient.

Because of the dimensionality of feature vector is large and sparse, sparse representation should be employed. In this project, we normalized the feature and rank distinct event to generate SVMLight format of the feature to do training and testing.

C. Model design

Different machine learning algorithms are implemented to create the model, including generalized linear models such as logistic regression, stochastic gradient descent classifier; ensemble method such as adaboost and random forest classifier. Also, we used support vector machines classifier, and decision tree classifier. To get a more reliable results, we use k-fold cross validation and randomized cross validation to test the model.

V. Experiment Design and evaluation

In this section we present the experiments we conducted and the evaluation of different models.

A. Data statistics

Here is a table of the initial data statistics, the number is the distinct patients per event type:

Event	events1	events2	events3
GENDER	12556	255021	1228652
YOB	12556	255021	1228652
DIAG	12664	256957	1230947
PROC	12663	256953	1230936
OUTPATIENT	1451	29257	160537
INPATIENT	1009	20302	106651
ER	1271	25226	136185
HOSPDAYS	1009	20302	106651
MOLECULE	8496	173768	1230947
QUANTITY	8496	173768	1230947
SUPPLYDAYS	8496	173768	1230947
CLASS	8496	173768	1230947

Here we also show the distinct event count for the patients important events including DIAG, PROC, MOLECULE, CLASS.

Event	events1	events2	events3
DIAG	7969	14202	18082
PROC	6819	13071	17407
MOLECULE	1504	2524	3194
CLASS	269	324	353

The events1 dataset is a small dataset that used for initial experiments. In this project we mainly use events2 dataset to construct feature and evaluate the model. The events3 (20G) dataset is 6 times larger than events2. We didn't use this dataset in the early stage of the project because it was provided during the middle stage, and we will be using this data set for more experiments.

B. Model evaluation

We used python scikit-learn package and implemented five different classification algorithms and used k-fold cross validation where k is 5 and randomized CV where the test percentage is 0.2. Here are all the evaluation results for the six algorithms.

Logistic Regression	k-foldCV	randomizedCV
Accuracy	0.929	0.930
AUC	0.610	0.611
Precision	0.910	0.903
Recall	0.221	0.225
F1-score	0.356	0.360

SVM	k-foldCV	randomizedCV
Accuracy	0.928	0.925
AUC	0.625	0.619
Precision	0.754	0.734
Recall	0.259	0.247
F1-score	0.385	0.370

Decision Tree	k-foldCV	randomizedCV
Accuracy	0.911	0.905
AUC	0.709	0.683
Precision	0.487	0.468
Recall	0.463	0.412
F1-score	0.472	0.438

SGDClassifier	k-foldCV	randomizedCV
Accuracy	0.929	0.931
AUC	0.599	0.605
Precision	0.997	0.992
Recall	0.200	0.211
F1-score	0.333	0.349

Adaboost	k-foldCV	randomizedCV
Accuracy	0.916	0.915
AUC	0.532	0.535
Precision	0.734	0.747
Recall	0.066	0.074
F1-score	0.122	0.135

Random Forest	k-foldCV	randomizedCV
Accuracy	0.934	0.929
AUC	0.659	0.639
Precision	0.778	0.770
Recall	0.325	0.286
F1-score	0.446	0.417

As we can see from the above results, random forest classifier and linear models like logistic regression and SGD classifier and SVM model gives high accuracy. Decision tree has the best AUC. Ensemble method adaboost gives worst performance compare to other algorithms.

We also compared their ROC curves. Here, we only show the ROC curve of logistic regression, decision tree and adaboost in fig 3, 4, 5. Noted that the ROC curve is present using probability estimates which makes the AUC differ from the AUC we previous calculated (using predicted label).

We found that for decision tree the number of pair of true positive rate point and false positive rate is significantly lower than the ROC curve for logistic regression and adaboost. This is because we often do binary splits in decision tree which makes the prediction discrete rather than a continuous, which may lead to bad performance when enlarge the dataset. We would try to use this model again in a bigger data set to see if the model still present the high AUC.

For trade-off of precision and recall, decision tree has a balance but low score, but the rest of the algorithms all have high precision and low recall, which means the true positive prediction label has low portion of the positive label. Since the recall is low the f1-score turns out to be low. However, based on the high precision we can know that the true positive label has a large portion of the sum of true positive and false positive, which means whenever the model gives out a positive prediction, the accuracy of this prediction is very high. This is very important in predicting of a disease because we are not willing to give false prediction, which often leads to inappropriate treatment for patient. Thus, we could make use of these algorithms in different perspectives.

Knowing that the random forest classifier gives a decent accuracy, we tried to tune the parameter of this classifier by enlarging the number of trees in the forest, enlarging the number of features to consider when looking for the best split, changing the function to measure the quality of a split, and we managed to improve its general performance to a higher level. Here is the improved result. We also show the ROC

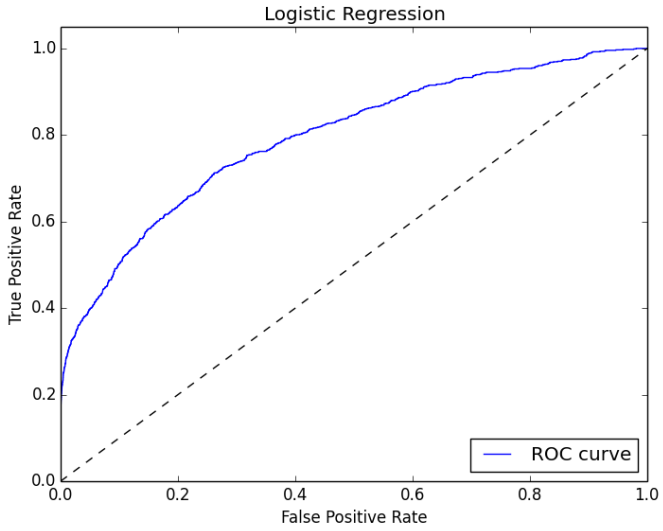

Fig. 3. logistic regression ROC curve

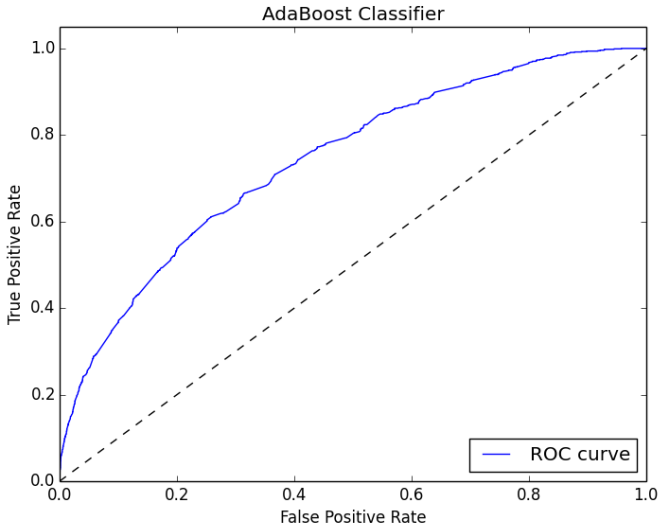

Fig. 5. adaboost ROC curve

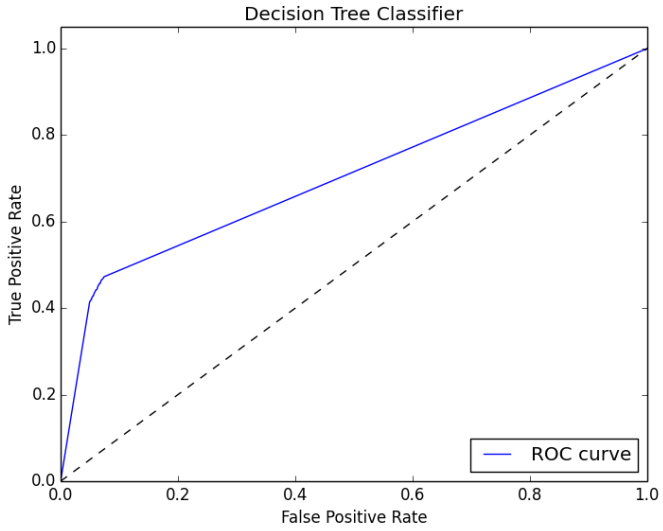

Fig. 4. decision tree ROC curve

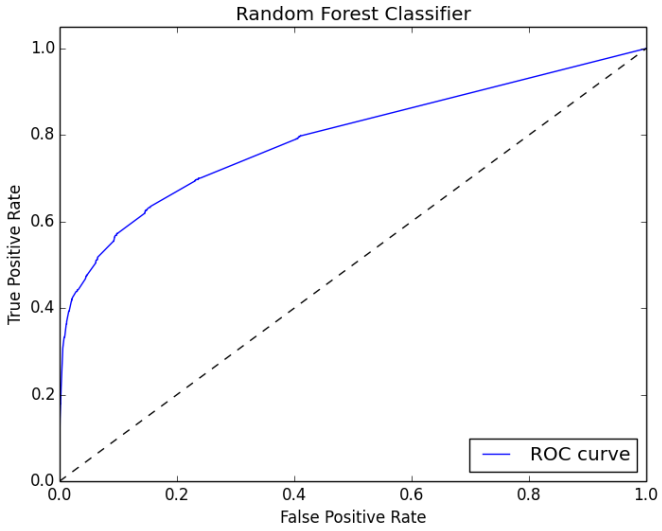

Fig. 6. Random Forest (tuned) ROC curve

curve for the random forest classifier in fig 6. We can see that the curve is much smoother than the decision tree, meaning there are more number of pair of true positive rate point and false positive rate.

Random Forest (tuned)	k-foldCV	randomizedCV
Accuracy	0.937	0.934
AUC	0.685	0.663
Precision	0.785	0.778
Recall	0.380	0.336
F1-score	0.501	0.470

Here we show the evaluation result we get using events3 dataset following the same feature construction and model implementation method. From the below charts we can see the performance of the models remain almost the same except for decision tree and random forest classifier. For decision tree the AUC drop significantly, which is led by our discussion when we see the ROC curve of this algorithm in the smaller dataset. However although the AUC for random forest classifier dropped as well, it beats decision tree with this larger data set, which means the property shown when discussing the ROC curve in the smaller dataset is essential when analyzing the classification algorithm.

Logistic Regression	k-foldCV	randomizedCV
Accuracy	0.930	0.930
AUC	0.609	0.610
Precision	0.920	0.916
Recall	0.217	0.220
F1-score	0.352	0.362

SVM	k-foldCV	randomizedCV
Accuracy	0.928	0.928
AUC	0.609	0.616
Precision	0.910	0.919
Recall	0.203	0.205
F1-score	0.332	0.335

Decision Tree	k-foldCV	randomizedCV
Accuracy	0.900	0.884
AUC	0.630	0.621
Precision	0.440	0.421
Recall	0.360	0.352
F1-score	0.407	0.384

SGDClassifier	k-foldCV	randomizedCV
Accuracy	0.929	0.929
AUC	0.597	0.600
Precision	1.0	1.0
Recall	0.197	0.199
F1-score	0.327	0.332

Adaboost	k-foldCV	randomizedCV
Accuracy	0.913	0.913
AUC	0.520	0.618
Precision	0.721	0.722
Recall	0.061	0.073
F1-score	0.100	0.105

Random Forest (tuned)	k-foldCV	randomizedCV
Accuracy	0.912	0.911
AUC	0.639	0.633
Precision	0.733	0.737
Recall	0.061	0.073
F1-score	0.489	0.485

VI. Conclusion

In this project, we make use of epilepsy patient medical record data to predict whether the patient will become drug-resistant, which also called refractory. We filter the data to extract adult epilepsy patient who is given AED, and use a one-year observation window to create the features. Then, we implement different machine learning algorithms to create the predictive model. We found feature engineering is important in this process. Creating decent features could leading to better prediction results even with simple model. Also, for the predictive model, we evaluate all the models, and found out we could make use of the random forest classifier model to make accurate prediction to the patients on large patient dataset.

There are two things that can be done further for this project. First, dimensional reduction method could be applied for generating more decent feature. Second, the random forest classifier model runs slow on large dataset such as event3, we could optimize the inner steps to make this algorithm runs faster. For future work, we would look into these unfinished work, and we hope the predictive model could be used in real world to predict the epilepsy patients to help them getting more appropriate instructions.

VII. Supplement Material

The video link of the presentation:
https://www.youtube.com/watch?v=i1HzxmMZeAQ

References

[1] J. Bauer, L. Buchmüller, M. Reuber, and W. Burr. Which patients become seizure free with antiepileptic drugs? an observational study in 821 patients with epilepsy. *Acta Neurologica Scandinavica*, 117(1):55–59, 2008.

[2] I. Gilioli, A. Vignoli, E. Visani, M. Casazza, L. Canafoglia, V. Chiesa, E. Gardella, F. La Briola, F. Panzica, G. Avanzini, et al. Focal epilepsies in adult patients attending two epilepsy centers: Classification of drug-resistance, assessment of risk factors, and usefulness of new antiepileptic drugs. *Epilepsia*, 53(4):733–740, 2012.

[3] P. Kwan and M. J. Brodie. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319, 2000.

[4] D. Schmidt and W. Löscher. Drug resistance in epilepsy: putative neurobiologic and clinical mechanisms. *Epilepsia*, 46(6):858–877, 2005.

[5] A. Voll, L. Hernández-Ronquillo, S. Buckley, and J. F. Téllez-Zenteno. Predicting drug resistance in adult patients with generalized epilepsy: A case–control study. *Epilepsy & Behavior*, 53:126–130, 2015.