Classification of Different Phenotypes of Refractory Epilepsy using Big Data Analysis

Xiaobing Li

Georgia Institute of Technology

Abstract- **Refractory epilepsy happens in around 30% epilepsy patients and the phenotypes of this population have not been studied thoroughly. This project classified different phenotypes using big data tools, that will allow us to predict drug resistance of patients therefore give them earlier treatments and more specified treatment plans.**

We studied the data of 12664 patients in an epilepsy database. Using big data tools, I extracted the refractory/ drug-resistant epilepsy (DRE) population and possible features related to DRE, then applied a RandomForest model to classify the DRE and non-DRE patients. Finally, I used K-means clustering to identify different phenotypes and compared the characteristics among two different phenotypes.

We found that the RandomForest model can successfully classify DRE and non-DRE patients based on only the events happened before patients were taking their second AED. The classify accuracy was 99.49%. The most important predicting features were the durations that a patient took his/her second and first AED and the dose of the second AED. K-means clustering divided the DRE population into two groups which have significantly different age, sick_age, delay_time and AED history.

Index Terms **-** Refractory Epilepsy, Phenotype, Big data, Health analytics, Data mining, Machine learning

I. Introduction and motivation

Drug-resistant epilepsy(DRE) is a major clinical and societal problem for one in three epilepsy patients[1,2]. We call this drug resistant epilepsy population as refractory epilepsy patients. Previous studies have attempted to identify phenotypic markers that can be used to predict refractoriness in patients with epilepsy. These include the type of syndrome, underlying etiology, patient history of seizure frequency and density, and EEG findings[1-4]. Other studies suggested genetic analysis, neuro-imaging techniques and epidemiological data analysis for predicting DRE[1-6]. Voll et al. showed developmental delay and more than one seizure type had significant correlations with DRE using logistic regression analysis[5]. Therefore the phenotypes of the refractory population can be quite diverse and related to many factors.

II. Problem formulation

To answer the above question, traditional studies need to collect a lot of historical, clinical, neuro-physiological and even genetic information. These strategies will always cost huge amount of time and effort and can only include limited patients. In modern days, we have data science tools to deal with huge scales of data, which might provide a quick and effective way to study the phenotypes of DRE. The specific data science question we are asking in this project is: giving limited information of a huge population of epilepsy patients, whether we can develop models/strategies to identify DRE patients and classify different phenotypes of DRE. The results can be useful to predict the drug resistance of patients.

III. Approach and implementation

Data source The dataset used in this project is prepared by CSE8803 TAs. I downloaded it from sunlab. cncuaxfkq7hp.us-east-1.rds.amazonaws.com through PostgreSQL GUI tool pgAdminIII. Based on the introduction in the description file, this dataset is a small subset (about 12k patients) sampled from entire dataset. The original source information is not provided.

The main dataset file event.csv consists of four fields: PATIENT_ID, EVENT_ID, EVENT_TIME and EVENT_VALUE. All medical /clinical related terms (e.g. ICD9, CPT, Generic name etc.) are hashed. A reference file aed_list.txt contains the hash strings of a list of Anti-Epileptic Drugs (AED).

Analytic infrastructure I used a local Hadoop cluster for the analysis. The cluster is consist of three computers. One computer has 8 cores at 3.3GHz and 8G memory. Other two computers have 4 cores at 3.0GHz and 4G memory. Cloudera Hadoop VMs are installed on these computers. Hive, Pig, Spark and R are supported. This sampled dataset is only 167M and contains ~2.5M records, so I used R studio for a quick preliminary analysis and data visualization, then deployed the strategy in Spark.

Xiaobing Li is a student in Online Master Program of Computer Science at Georgia Institute of Technology. Email: xli605@gatech.edu.

Implementation The approach and implementation consisted of a list of steps: First, initial statistics for the whole dataset and extract all the epilepsy patients. The data integrity was checked in this step. Since we didn't have the diagnosis information, I assumed that any patient who was taking any AED was an epilepsy patient; Second, extract the features related to DRE; Third, feature selection and data visualization. Here I used the findCorrelation in the Caret R package to calculate the correlation matrix. Features with an absolute correlation of 0.75 or higher were removed. To visualize the DRE and non-DRE patients, PCA method in the Caret package was applied; Fourth, DRE prediction. A RandomForest model was built based on the events before first two AEDs to predict if a patient would be DRE positive in the future; Fifth, k-means clustering to classify the phenotypes of DRE. These modelings were first tested on R packages and eventually were implemented in Spark.

In the project instruction, the mentor suggested to define the refractory epilepsy population as epilepsy patients who failed 4 or more anti-epilepsy drugs (AED), so that I tagged patients with 4 or more AEDs as DRE positive. Those with 1-3 AEDs and still had events after the last AED were considered DRE negative. Therefore the above feature selection and PCA can be done based on this tag. The RandomForest prediction can be estimated by a confusion matrix, also sensitivity and specificity.

K-means methods for classifying the phenotypes of DRE is an unsupervised methods, therefore there was no absolute criterion for the outputs. Based on the PCA visualization, I decided 2-3 phenotypes would be reasonable.

IV. Experiment Design and evaluation

Data Preparation

The initial statistics were done in R. The two data files events.csv and aed_list.txt were loaded. The field event_id was splitted into prefix(DIAG, PROC, MOLECULE etc.) and hashed medical/clinical terms, so that I can easily search AED in the table.

This dataset contains 2,469,485 records from 12664 patients, in which 5786 patients were taking AED. 283 of them have taken more than 4 AEDs (Figure1). By our definition, these patients were DRE patients. Because the event dates were masked, I can only approximately estimate the ages of patients by calculating the difference between their YOB and the latest date in the dataset ("2113-12-14").

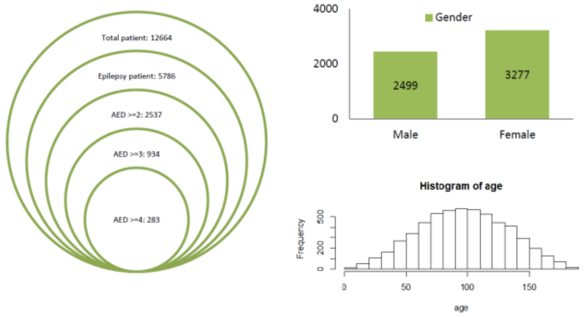

Figure 1. Left: patient numbers in different categories. AED>=4 are defined as DRE patients. Right upper: gender distribution. Right bottom: histogram of age. The ages showing here are estimations. Clearly most of patients have passed away.

Feature Extraction

The following features were extracted from the dataset: the approximate patient age estimated by the maximum event date in the dataset (Age), the patient gender(Gender, Male=1, Female=0), the age of the patient when he/she took his/her first AED(Sick_Age), the date between his/her first non-AED drug and the first AED(Delay_time), total dosage for the 1st AED(AED1_dose), time between the 1st AED and the 2nd AED(AED1_time), total dosage for the 2nd AED (AED2_dose), time between the 2nd and 3rd AEDs (AED2_time), dosage for the 3rd AED (AED3_dose), time between the 3rd AED and the 4th AED (AED3_time), total dosage for the 4th AED (AED4_dose), time between the 4th AED and the 5th AED (AED4_time), total dosage for the 5th AED (AED5_dose), total number of AEDs (AED_num), total dosages(Total_dose). If patients did not take so many AEDs, the related fields would be filled with 0. The AEDX_time for the last AED would be the event_value for SUPPLYDAYS. The AED_num counted only the distinct AED numbers. The patients with 4 or more AEDs would be tagged as DRE(1), otherwise DRE(0). During the feature extraction, 10 patients without enough information were removed.

All the features were plotted based on DRE-negative(value=0) and DRE-positive(value=1) in Figure 2. For visualization purpose, DRE values are jittered around 0 or 1. The data points around y=1 belong to the DRE population and the data points around y=0 belong to the non-DRE population. In this figure, no single feature had a significant correlation with DRE except AED_NUM. However, because the AED_NUM was what I used to define DRE, it should not be included as a predictive feature. In the following analysis, AED_NUM was removed.

Furthermore, the feature data had two issues. First, the values of different features were widely ranged.

Therefore I centered and scaled the feature data to 0-1 for the following analysis. Second, the DRE patients were only a small portion of the epilepsy population (283/5776=4.90%). To make the dataset balanced for modeling, I randomly re-sampled the dataset to make the number of DRE patients as same as the number of non-DRE patients. That made cross-validation and result estimation much easier.

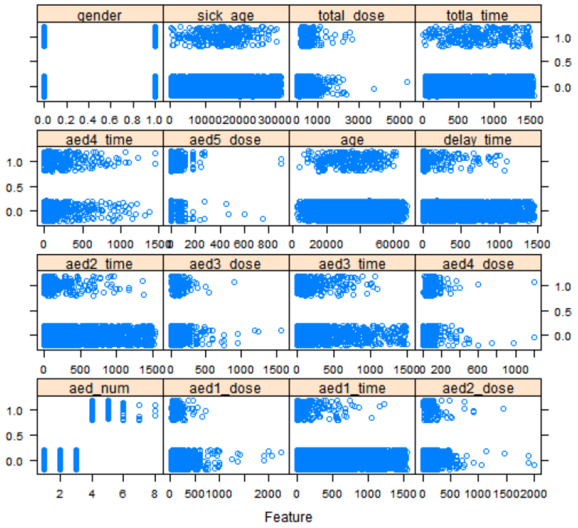

Figure 2. All the features are plotted to DRE. The Y axis represents the DRE classes.

Feature Selection

Some of the features I chose may be highly correlated with each other. They should not be carried into further modeling. The Caret R package provides the findCorrelation which can generate a correlation matrix for all the features. If a feature had a >0.75 correlation with another feature, it would be removed. I found even Total_dose had high correlations with AED2_dose, AED3_dose and AED4_dose (>0.6), the correlations were still smaller than 0.75, so that none of features were removed in this step.

Next I tested the feature importance using a randomForest core function with 5 cross-validations. The features listed based on the importance from the greatest to the smallest were aed2_time, aed3_time, aed1_time, aed4_dose, sick_age, aed5_dose, aed4_time, delay_time, age, total_dose, aed1_dose, aed3_dose, total_time, aed2_dose, gender. The classification accuracy was plotted to the number of top features in Figure 3. The top five features contributed the most part of the classification accuracy. The least related feature was gender.

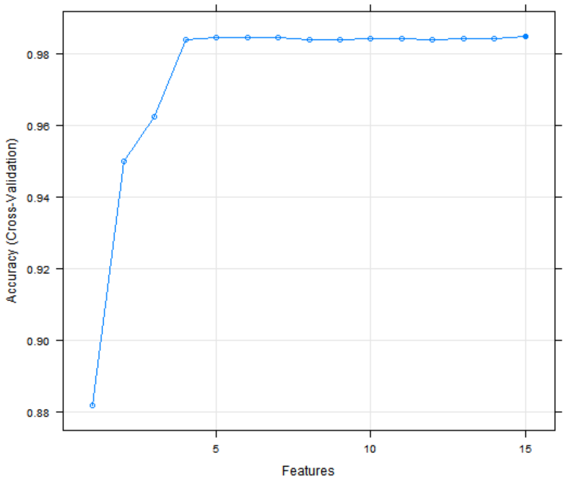

Figure 3. The accuracy-feature plot. The top five features contributed to the most part of the accuracy.

Before diving into the prediction modeling, I applied PCA on the dataset and plotted PC2-PC1 and PC3-PC1 in Figure 4. DRE and non-DRE data points were marked in Blue and Red respectively. For this dataset, PC1 explained 23.6% of the variance, PC2 explained 12.3% of the variance and PC3 explained 10.3% of the variance. Although the dataset cannot be simply represented by a few of PCs, the PCs provided an easy way to visualize the dataset. From the figure, we can tell the DRE population (Blue) did have different characteristics as the non-DRE population. Moreover, we saw the non-DRE population might be able to divide into different clusters.

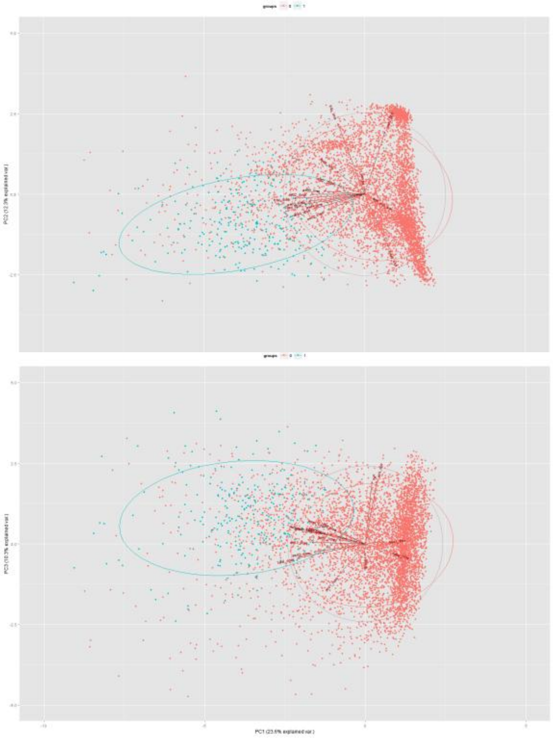

Figure 4. PCA dimension reduction. DRE patients are in RED and non-DRE patients are in BLUE. Top: PC2 vs PC1. Bottom: PC3 vs PC1.

DRE Prediciton

For DRE prediction, I first extracted the event features before the first two AEDs, which were Age, Gender, Sick_Age, Delay_time, AED1_dose, AED1_time, AED2_dose and AED2_time. The patients were tagged as DRE positive (1) or DRE negative(0). The R RandomForest package was used. It was not necessary to split the dataset into a train set and a test set for a RandomForest model because it randomly sampled data. The model with 500 trees classified DRE and non-DRE patients in a 99.49% accuracy. Sensitivity was 100% (defined as DRE patients were correctly classified as DRE). Specificity was 99.0% (defined as non-DRE patients were correctly classified as non-DRE). Note I have re-sampled the dataset to make the DRE:non_DRE ratio at 1:1, so the model prediction was quite successful. The confusion matrix was shown in Table 1.

Table 1. Confusion matrix

Confusion matrix	*estimate of error rate: 0.51%*		
	Non-DRE	*DRE*	
Non-DRE	*5437*	*56*	*Specificity = 98.98%*
DRE	*0*	*5493*	*Sensitivity = 100.00%*

The RF model also provided the feature importance (Figure 5). The importance was estimated by the mean decrease of Gini coefficient, which is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). A high value of MeanDecreasedGini represented a high importance. Figure 5 indicated that AED2_dose, AED2_time and AED1_time were the top three importance features to classify DRE and non-DRE. Gender had the least influence.

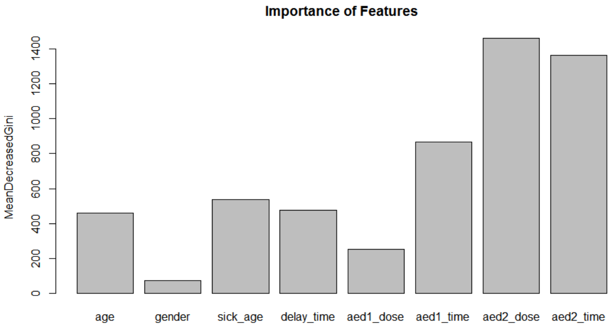

Figure 5. Feature importance in the meaning of MeanDecreasedGini. A high value of MeanDecreasedGini represents a high importance.

K-means clustering

For classify the phenotypes of DRE patients, I used K-means clustering. Different k numbers were tested. Based on the PCA visualization (Figure not shown), I chose K=3 as a reasonable cluster number. However, when choosing K=3, one cluster only included 2 patients (Figure 6, left panel). Therefore, I did another clustering at K=2, which was shown in the right panel of Figure 6. The average distance in K=3 clustering was 0.48 while the average distance in K=2 clustering was 0.19, so that K=2 clustering was better than K=3.

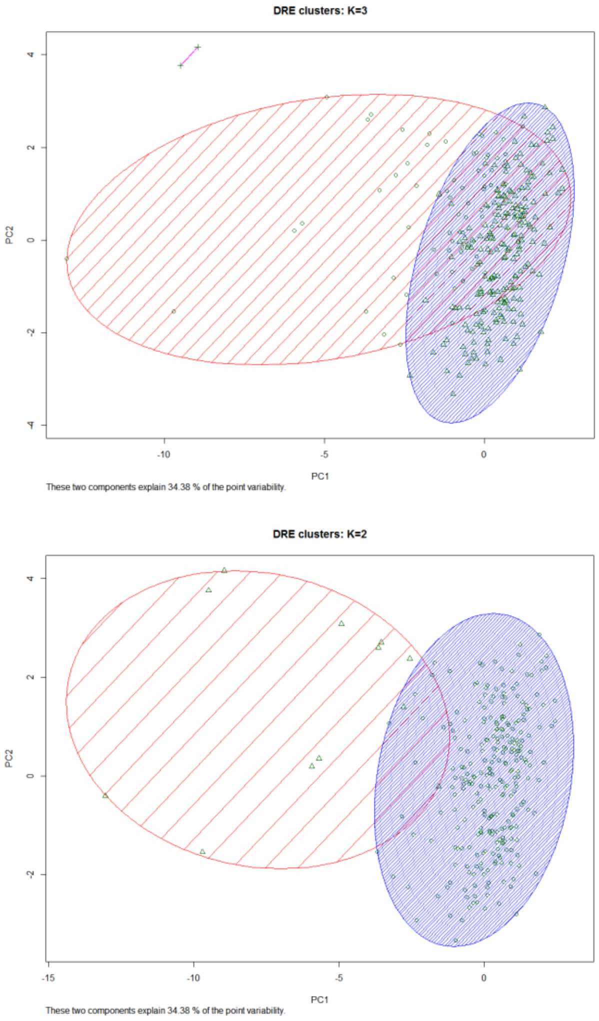

Figure 6. K-means clustering for DRE patients. Top: K=3. Bottom: K=2.

Table 2 compares different features in these two clusters. The DRE patients in Cluster 1 had significantly older age (p=0.008), older sick_age (p=0.010), longer delay_time (p=0.003), lower AED1_dose (p=0.040), lower AED2_dose (p=0.009), lower AED3_dose (p=0.012), lower AED5_dose (p=0.015) and lower total_dose(p=0.000) than Cluster 2.

Table 2. Comparison between two clusters

	Age	Gender	Sick_age	Delay_time	AED1_dose	AED1_time	AED2_dose	AED2_time
P	**0.008**	**0.365**	**0.010**	**0.003**	**0.040**	**0.813**	**0.009**	**0.250**
	AED3_dose	**AED3_time**	**AED4_dose**	**AED4_time**	**AED5_dose**	**Total_dose**	**Total_time**	
	0.012	**0.434**	**0.053**	**0.884**	**0.015**	**0.000**	**0.423**	

V. Conclusion

The RandomForest model can successfully classify DRE and non-DRE patients based on only the events happened before patients were taking their second AED. The classify accuracy was 99.49%. The most important predicting features were the durations that a patient took his/her second and first AED and the dose of the second AED.

K-means clustering divided the DRE population into two groups which have significantly different age, sick_age, delay_time, AED1_dose, AED2_dose, AED3_dose, AED5_dose and Total_dose.

Our results proved that with limited information of epilepsy patients, we can successfully predict whether a patient would be a DRE patient in the future in very high accuracy. That will provide evidence for doctors to give a patient earlier treatments and more specified treatment plans. Moreover, we suggested that there were different phenotypes in the DRE population. The phenotypes have significantly different characteristics between each other. More studies are required for those different phenotypes.

Supplement Materials:

You can find my oral presentation video at

https://youtu.be/B3VuU6B7UDo

ACKNOWLEDGMENT

Thanks all the TAs and professor for their time and effort in creating the course and project materials together.

References

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

[2] Kwan, P, Arzimanoglou, A., Berg, A.T. et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia 51: 1069-1077, 2010.

[3] I. Gilioli, A. Vignoli, E. Visani, M. Casazza, L. Canafoglia, V. Chiesa, E. Gardella, F. La Briola, F. Panzica, G. Avanzini, M. P. Canevini, S. Franceschetti, and S. Binelli. 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, Apr. 2012.

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

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

[6] French, J. A., Refractory Epilepsy: Clinical Overview. Epilepsia, 48: 3–7, 2007.