Automated feature selection of predictors in electronic medical records data

Jessica Gronsbell, Jessica Minnier, Sheng Yu, Katherine Liao, and Tianxi Cai

The report summarizes the above mentioned paper

It is the report part of the STA2112 final project

Prepared by Yi Han and Yaqi Shi

Department of Statistical Sciences

University of Toronto

Toronto, Ontario, Canada

December 6, 2022

1 Introduction & Motivation

Electronic Health Record (EHR) contains many clinical data that are useful for identifying disease phenotypes. Linking the phenotype data from EHR with omics (biological) data from biobanks is useful for translational research in omics studies. Hence, it is important to accurately and efficiently identify disease phenotypes from EHR.

One historical approach for disease phenotyping is to determine the patient's binary disease status according to the International Classification of Diseases Ninth Edition (ICD-9) billing code. A drawback of this approach is that the billing codes for many disease phenotypes are vague, which can lead to incorrect prediction of disease phenotypes. Another historical approach is to create gold standard labels of disease phenotypes by manually reviewing the medical chart, and applying a supervised machine-learning model to associate the gold standard label with some features in EHR such as medication prescriptions and laboratory results. However, since this algorithm usually uses a lot of features, and thus requires a significant amount of gold standard labels for training. The process of creating those gold standard labels relies on heavy labor work, hence this approach is not efficient and may not be stable across different institutions due to the dependence on labor work.

This paper discusses a novel procedure to accurately and efficiently identify disease phenotypes based on EHR data. Specifically, the authors developed an unsupervised procedure based on unlabeled observations to automatically identify features that are predictive of the disease phenotypes, which solves the inconsistency and inefficiency issues of previous approaches.

2 Methods

In this paper, the authors proposed a new method that performs automated feature selection on EHR data based on the unlabeled sets. The proposed method contains two main procedures: unsupervised feature selection procedure and variable selection via resampling procedure.

In this paper, the authors used $\mathbf{X}_{p\times 1}$ to represent candidate EHR features, $\mathbf{S}_{q\times 1}$ to represent the surrogate features and Y to represent the true binary disease status for a subject. The underlying data contained N independent and identically dis-

tributed random vector $\mathscr{F} = \left\{\left(Y_i, \mathbf{X}_i^\top, \mathbf{S}_i^\top\right)^\top, i = 1, \ldots, N\right\}$. Since the true status of the disease were difficult to obtain, the working dataset would be the observed set $\mathscr{D} = \left\{\mathbf{W}_i = \left(\mathbf{X}_i^\top, \mathbf{S}_i^\top\right)^\top, i = 1, \ldots, N\right\}$. The quantities defined during the demonstration of the method will be introduced as well.

2.1 Key Assumptions

The authors made several assumptions on $\mathbf{S}, \mathbf{X}$ and Y for the proposed method. One basic assumption that carried over from previous work is that $\mathbf{X}$ is elliptical symmetric. For the surrogate variable $\mathbf{S}$, the authors assumed that $\mathbf{S}$ depends on $\mathbf{X}$ only through Y, that is, (Carroll et al., 2006).

$$\mathbf{S} \perp \mathbf{X} \mid Y$$

Then the authors assumed that Y follows a GLM, that is

$$P(Y = 1 \mid \mathbf{X}) = g\left(\alpha_0 + \mathbf{X}^\top \beta_0\right) = g\left(\overrightarrow{\mathbf{X}}^\top \theta_0\right)$$

with $\overrightarrow{\mathbf{X}} = \left(1, \mathbf{X}^\top\right)^\top$, $\theta_0 = \left(\alpha_0, \beta_0^\top\right)^\top$ and $g(\cdot)$ is a known smoothed link function for GLM. The authors further defined $\mathcal{A} = \{j \mid \beta_{0j} \neq 0\}$ and the goal of this method is to find a concise and predictive set $\mathcal{A}$.

2.2 Unsupervised feature selection procedure

Clustering and regularized estimation are the two main steps in the unsupervised feature selection procedure. In the clustering step, the authors estimated $\pi_{\mathbf{S}} = P(Y = 1 \mid \mathbf{S})$ by imposing a working parametric mixture model of $\mathbf{S}$ as

$$\mathbf{S} \sim \tau f_{\boldsymbol{\Theta}_1}(\mathbf{s}) + (1 - \tau) f_{\boldsymbol{\Theta}_0}(\mathbf{s})$$

where $f_{\Theta_y}(\cdot)$ are distribution functions of $\mathbf{S} \mid Y = y$ with unknown parameter Θ_y for $y = 0, 1$ and $\tau = P(Y = 1)$. The authors estimated the model parameters $\Theta_\bullet = \left(\Theta_1^\top, \Theta_0^\top, \tau\right)^\top$ using the maximum likelihood estimator obtained from the Expectation-Maximization algorithm (Dempster et al., 1977). Then the authors approximated $\pi_{\mathbf{S}}$ through

$$\hat{\pi}_{\mathbf{S}} = \Pi_{\mathbf{S}}\left(\hat{\Theta}_\bullet\right) \quad \textit{where} \quad \Pi_{\mathbf{S}}\left(\Theta_\bullet\right) = \frac{\tau f_{\boldsymbol{\Theta}_1}(\mathbf{S})}{\tau f_{\boldsymbol{\Theta}_1}(\mathbf{S}) + (1 - \tau) f_{\boldsymbol{\Theta}_0}(\mathbf{S})}.$$

After estimating $\pi_{\mathbf{S}}$, the authors moved on to the regularized estimation step by fitting a penalized quasi-logistic regression of $\hat{\pi}_{\mathbf{S}_i}$ against $\mathbf{X}_i$. Here the authors denoted the estimate $\widehat{\theta} = \left(\widehat{\alpha}, \widehat{\beta}^\top\right)^\top = argmin_\theta \widehat{\mathcal{L}}(\theta)$ where

$$\widehat{\mathcal{L}}(\theta) = N^{-1} \sum_{i=1}^{N} \ell\left(\theta^\top \overrightarrow{\mathbf{X}}_i, \hat{\pi}_{\mathbf{S}_i}\right) + \lambda_N \sum_{j=1}^{p} |\beta_j| / \left|\widetilde{\beta}_j\right|,$$

with $\ell(x_i, \pi_i)$ being the negative log-likelihood contribution from the i th subject and $\lambda_N \sum_{j=1}^{p} |\beta_j| / \left|\widetilde{\beta}_j\right|$ being the adaptive least absolute shrinkage and selection operator (ALASSO) penalty. Now the authors obtained the active set $\widehat{\mathcal{A}} = \left\{j : \widehat{\beta}_j \neq 0\right\}$ containing the index of the features that are predictive of Y. To simplify the computation, the authors used the method in Wang and Leng (2007) and approximate $\widehat{\mathcal{L}}(\theta)$ by

$$\left\|\widetilde{Y} - \theta^\top \widetilde{X}\right\|_2^2 + \lambda_N \sum_{j=1}^{p} |\beta_j| / \left|\widetilde{\beta}_j\right|$$

where $\widetilde{\theta} = \left(\widetilde{\alpha}, \widetilde{\beta}^\top\right)^\top$ is the minimizer of $\tilde{\mathcal{L}}(\theta) = N^{-1} \sum_{i=1}^{N} \ell\left(\theta^\top \overrightarrow{\mathbf{X}}_i, \hat{\pi}_{\mathbf{S}_i}\right)$, $\widetilde{X}$ is the symmetric half matrix of $\widetilde{I} = N^{-1} \sum_{i=1}^{N} \ddot{\ell}\left(\widetilde{\theta^\top} \overrightarrow{\mathbf{X}}_i, \hat{\pi}_{\mathbf{S}_i}\right)$ such that $\widetilde{I} = \widetilde{X}^T \widetilde{X}$ and $\ddot{\ell}(x, \pi) = \partial^2 \ell(x, \pi) / \partial x^2$ and $\tilde{Y} = \widetilde{\theta}^\top \widetilde{X}$.

2.3 Variable selection via resampling

After the unsupervised feature selection procedure, the next step is to perform variable selection through resampling. The authors have obtained a set of index $\hat{\mathcal{A}}$ that corresponding to the features being predictive to Y. However, due to the uncertainty in $\hat{\mathcal{A}}$, the authors believed that $\mathbf{X}_{\hat{\mathcal{A}}}$ does not perform as well as $\mathbf{X}_{\mathcal{A}}$, especially when the number of true golden standard features is relatively small and $\mathbf{X}_{\hat{\mathcal{A}}}$ may contain some week or null signals, which would influence the model prediction significantly.

The authors proposed a resampling method to deal with this instability in the feature selection. The authors wanted to obtain an estimate of $\widehat{\rho}_{0j} = P\left(\widehat{\beta}_j = 0\right)$ and perform feature selection based on $\widehat{\rho}_{0j}$. Since standard sampling procedure is heavy in computation, the authors decided to use the m out of n bootstrap (Bickel, 1997) and subsampling (Politis and Wolf, 1999). The authors denote the size of subsample as N_b, where N is the size of the EHR dataset.

For each subsample $\mathcal{R}_m$, the authors preformed the regularized estimation described in the first procedure and obtain an estimate $\widehat{\theta}^{(m)} = \left(\widehat{\alpha}^{(m)}, \widehat{\beta}^{(m)\top}\right)^\top$ of $\hat{\theta}$. After repeating

the subsampling M times, the authors obtained:

$$\widehat{\rho}_{0j} = M^{-1} \sum_{m=1}^{M} I\left(\widehat{\beta}_j^{(m)} = 0\right)$$

Continuing with feature selection, the authors decided to keep the feature that has a $\widehat{\rho}_{0j}$ below the cutoff $\rho_{cut} \in (0, 1)$. The authors found that $\rho_{cut} = 0.5$ is a standard and useful cutoff in practice.

2.4 Method Summary and Comparison

The authors developed a two-step procedure to perform automated feature selection using the unlabeled set only. The authors first performed a unsupervised feature selection through clustering and regularized estimation and then performed the variable selection through resampling. Based on the method assumptions, the authors demonstrated that the variables selected through fitting a penalized quasi-logistic regression of $\pi_{\mathbf{S}}$ on $\mathbf{X}$ are consistent up to a scalar with the result by fitting Y on $\mathbf{X}$ using GLM. This implies that people can use the features selected through the unlabelled set to predict the label of the subject. The method is different from previous ones as it does not rely on the existence of the labelled set and it can deal with multiple surrogate variables with non-binary values. This approach is logical as it adapts to the feature of EHR dataset and it can solve the efficiency and accuracy problem when modelling EHR data. Since it is time consuming to get the labelled set, developing a method without labelled set can improve the efficiency of performing data analysis. The proposed method provides an efficient and accurate way of selecting features, which achieved the goal of this paper.

3 Simulation Results

3.1 Simulation Settings

The procedure was demonstrated using simulated data under 3 settings:

- setting 1: the authors assumed the logistic regression model for Y is correct, $P(Y = 1|X) = \frac{1}{1+exp(-(\alpha_0+\beta_0^T X))}$, and $E[b^T X|X^T\beta_0]$ is linear in $X^T\beta_0$

  The distribution of X is elliptically symmetric is a result presented by Li and Duan (1989), on which the authors' method is based. Hence this setting aims to examine the performance of the proposed method when all assumptions hold.

- setting 2: distribution of X is not elliptically symmetric

  Here in setting 2, the authors want to test if the proposed algorithm is robust when breaking the assumption on the distribution of X.

- setting 3: $S \not\perp X \mid Y$

  The assumption $S \perp X \mid Y$ ensures that S relates to X only through Y, and thus establishes the connection between the predicting power of X for Y and the predicting power of X for π_S. Hence, the authors used this setting because they want to determine whether their proposed method works when this assumption does not hold.

3.2 Simulated Data

In all settings, N=5000 samples were simulated with binary response Y and prevalence P(Y=1) = 0.3. The authors considered p=50 or 100 features in each setting and used 2 surrogate variables S_1, S_2. X was generated based on multivariate normal (MVN) distribution which is elliptically symmetric. The covariate vector X is defined as:

- $X_i \sim MVN(0, \sum^X) + \mu^X_{y_i} \quad \mu^X_{y_i} = y_i \sum^X \beta_0 \quad \sum^X_{kk} = 1 \quad \sum^X_{kj} = 0.5^{|k-j|}$

- in setting 2, the authors applied the transformation $log(\lfloor exp(X) \rceil + 1)$, where $\lfloor exp(X) \rceil$ means getting the nearest integer of exp(X)

The surrogate vector S is defined as:

- setting 1 & 2: $S^0_i \sim MVN(0, \sum^S) + \mu^S + y_i \Delta^S_0$

- setting 3: $S^0_i \sim MVN(0, \sum^S) + \mu^S + y_i \Delta^S_0 + [X_1, X_3]$

where $\mu^S = [-1, -2]^T$, $\sum^S_{11} = 0.65$, $\sum^S_{12} = 7.3$, $and \sum^S_{22} = 0.65$. The authors also applied the transformation $S = log\{\lfloor exp(S^0) \rceil + 1\}$. In this report, the authors reported the results under a "strong signal setting" with $\beta_0 = [1.2, -1.2, 0.5, -0.3, 0.3, 0.1, 0.1, 0^T_{(p-7)\times 1}]^T$, $\Delta^S_0 = [0.75, 3]^T$.

This simulation set-up makes sense because it addresses the different assumptions in different simulation settings. For example, the authors applied the transformation $log(\lfloor exp(X) \rceil + 1)$ on covariates X generated in setting 2 to address the assumption that $E[b^T X | X^T \beta_0]$ is not linear in $X^T \beta_0$ according to the simulation setting 2. The

transformation $S = log\{\lfloor exp(S^0)\rceil + 1\}$ ensures that S was generated as count variables, which makes sense because surrogate variables are often count variables in EHR data. We also notice that the way S was generated in setting 3 is different than that in setting 1&2. The authors added $[X_1, X_3]$ to S_i^0 to address the fact that $S \not\perp X \mid Y$ in simulation setting 3.

To verify the assumption that Y follows a logistic regression model in a real data set, we can plot the proposed logistic regression function on the data set and examine whether the plot aligns with the trend in the data set. We can also look at the R^2 in the model output in R, if R^2 of the fitted logistic regression model on Y is small, then this suggests that a logistic regression model for Y is probably not suitable. Although it is hard to check whether $E[b^T X | X^T \beta_0]$ is linear in $X^T \beta_0$ in real data, we can plot the distribution of X to see if it is elliptically symmetric. To verify the assumption that $S \perp X \mid Y$, we can plot S against X under different Y values to see if the plot suggests any correlation between S and X given Y.

3.3 Simulation Methods

The authors compared supervised methods that directly fit Y on X with other existing unsupervised feature selection methods, specifically, they are interested in the methods using the proposed automated clustering procedure with/without resampling ($AutoClust$, $AutoClust_R$). The unsupervised feature selection methods they compared are: $PenReg_{S_1+S_2}, PenReg_S$: select features from a regression model with $S_1 + S_2$ and S as the response respectively, $RankCor_{S_1+S_2}$: feature selection using the rank correlation method proposed by Yu et al. (2015) based on $S_1 + S_2$ and an absolute rank coefficient threshold 0.15, $Extreme$: feature selection using the extreme sampling method proposed by Yu et al. (2016), and $Agarwal_{S_1+S_2 \geq 1}$: use method proposed by Agarwal et al. (2016)

After feature selection, the authors used n=100 or 200 labeled samples to train the final algorithm. The authors repeated the simulations 500 times using Monte Carlo simulations, and calculated the average area under the receiver operating characteristic curve (AUC) obtained on the 100 or 200 labeled samples respectively (AUC_{100}, AUC_{200}) for each model to indicate prediction performance. The authors used proportion of times each feature was selected among the 500 replications in setting 1, and the average model size in setting 2&3 to indicate feature selection results.

3.4 Simulation Results & Comparison of Methods

In all 3 settings, the authors observed that: (1) supervised methods tend to produce overly simple models which have weaker prediction performance, (2) the automated selection procedure has improved the prediction performance compared to directly training a supervised algorithm on the features, (3) $AutoClust_R$ has the highest AUC_{100} and AUC_{200} among all models in both p=50 and p=100 cases.

Some unsupervised methods might have better prediction performance than the supervised methods, but do not perform well on feature selection. For example, in setting 1, $PenReg_{S_1+S_2}$, $PenReg_S$, and $Agarwal_{S_1+S_2\geq 1}$ are more likely to include null features. $RankCor_{S_1+S_2}$ and $Extreme$ have the tendency to exclude features that should have been selected in setting 1. Hence, compared with other models, $AutoClust$ and $AutoClust_R$ are the optimal methods that achieved a balance between accurate feature selection and high prediction performance. We see that in all 3 settings, the average AUC on 100 labeled samples for $AutoClust_R$ is similar to or larger than the average AUC on 200 labeled samples using the supervised methods. This suggests that the proposed automated feature selection and resampling procedure can achieve similar or better prediction performance using only half of the labeled samples required by the supervised methods. In terms of the problem proposed at the beginning, this would mean that we can reduce a significant amount of manual work for gold standard labeling using the proposed approach.

In addition to simulated data, the authors also examined the algorithm performance on real EHR data to identify rheumatoid arthritis (RA) because RA is the most common autoimmune joint disease which affects 1% of the worldwide population. The authors used NLP mentions of RA related terms appearing in online knowledge sources (such as Wikipedia and Mayo Clinic) as candidate features, and considered two surrogate variables: number of recorded RA ICD-9 codes, and count of NLP mentions of RA in patient records. We see that the surrogate variables relate to the candidate features only through the true response Y (RA), hence the conditional independence assumption $S \perp X \,|Y$ is satisfied. The real data study gives the same results as the simulation study: (1) $AutoClust$ and $AutoClust_R$ have larger average AUC than the supervised algorithm trained with all features, (2) $AutoClust_R$ only need half as many labels required by the supervised methods yet achieve the same prediction accuracy, which can reduce the time and labour effort for gold standard labeling in real life.

4 Conclusion

This paper proposed a new automated feature selection method using only the unlabelled dataset. It constructed a silver standard labels using model-based clustering and performed regularized regression on it to select the features. The authors implemented additional resampling procedure to improve the feature selection result from previous steps. Based on the simulation study and real EHR data analysis on RA, it has been shown that the automated feature selection with resampling method performs well on both feature selection and model prediction compared to other supervised and unsupervised learning methods. Another remark from the paper is that people may achieve similar prediction accuracy as the standard supervised learning when using the proposed method for feature selection with half the labelled set size. With such high performance, this proposed method will help reduce the number of medical chart reviews (labelling) thus improve the efficiency of training phenotyping algorithms.

To summarize, the authors believed that the proposed method has higher prediction power (AUC) and provides less but accurate features, especially with resampling. This method can significantly improve the efficiency for EHR phenotyping and provide insights to further develop unsupervised phenotyping methods.

The results presented in this paper are reliable and convincing. The authors performed simulation studies with different scenarios and the results were consistent among scenarios. In all three scenarios, the automated feature selection method performed well in feature selection. It contained most of the important features and the number of features was smaller compared to other methods. Also, the proposed method provided a more accurate model prediction compared to other methods. The authors also validated the method through real EHR data analysis and the conclusion still hold.

Based on the result, it is reasonable to believe that the automated feature selection method with resampling is substantially different to previous work. This paper is different as it developed a method to perform feature selection using only the unlabelled data. It improved the efficiency by eliminating the time-consuming task of medical chart review. It provided a method that deals with multiple non-binary surrogate variable and has the potential of handling high-dimensional predictors. It also developed a silver standard variable based on the surrogate variable.

5 References

Agarwal, V., Podchiyska, T., and Banda, J. M., et al. (2016). Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inf Assoc JAMIA 23, 1166–1173.

Bickel, P. J., Gze F., and Zwet W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. Statistica Sinica 7, 1-31.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective. New York: CRC press.

Dempster, A. P., Laird, N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc, Ser B (methodol) 39, 1–38.

Politis, R. J. P. and Wolf, M. (1999). Subsampling. New York: Springer.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. J Am Stat Assoc 102, 1039–1048.

Yu, S., Liao, K. P., and Shaw, S.Y., et al. (2015). Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. J Am Med Inf Assoc 22, 993–1000.

Yu, S., Chakrabortty A, Liao KP, et al. (2016). Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inf Assoc JAMIA 24, e143–149.

6 Contributions

- Introduction & Motivation: Yi Han
- Methods: Yaqi Shi
- Simulation Results: Yi Han
- Conclusions: Yaqi Shi