Resampled dimensional reduction for feature representation in machine learning

Herdiantri Sufriyana

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Department of Medical Physiology, Faculty of Medicine, Universitas Nahdlatul Ulama Surabaya, 57 Raya Jemursari Road, Surabaya 60237, Indonesia. https://orcid.org/0000-0001-9178-0222

Yu Wei Wu

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0002-5603-1194

Emily Chia-Yu Su (emilysu@tmu.edu.tw)

(1) Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. (2) Clinical Big Data Research Center, Taipei Medical University Hospital, 250 Wu-Xing Street, Taipei 11031, Taiwan. (3) Research Center for Artificial Intelligence in Medicine, Taipei Medical University, 250 Wu-Xing Street, Taipei 11031, Taiwan. https://orcid.org/0000-0003-4801-5159

Method Article

Keywords: resampling method, dimensional reduction, latent variable, machine learning

Posted Date: October 13th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1636/v1

Abstract

We aimed to provide a resampling protocol for dimensional reduction resulting a few latent variables. The applicability focuses on but not limited for developing a machine learning prediction model in order to improve the number of sample size in relative to the number of candidate predictors. By this feature representation technique, one can improve generalization by preventing latent variables to overfit data used to conduct the dimensional reduction. However, this technique may warrant more computational capacity and time to conduct the procedure. The key stages consisted of derivation of latent variables from multiple resampling subsets, parameter estimation of latent variables in population, and selection of latent variables transformed by the estimated parameters.

Introduction

Adoption of insurance-based healthcare is emerging worldwide, which needs better prevention in order to improve both patient outcome and healthcare efficiency. To achieve these goals, machine learning algorithms are widely applied for developing a clinical prediction with satisfying predictive performance.[1-6] However, machine learning models, including those applying multivariable logistic regression, were high risk of bias, especially because of low sample size in relative to the number of candidate predictors.[7]

To deal with this problem, dimensional reduction can be applied to represent many candidate predictors into fewer latent variables. However, most prediction models that used these variables, if not all, conducted a dimensional reduction without either resampling or data partition, which exposed to a risk of optimistic bias, and is not robust for samples beyond the training set, which is one of the main problems in current machine learning practice.[8] This is because resampling or data partition are more well-known in either predictive modeling or supervised machine learning, compared to a dimensional reduction that is typically used for statistical inference and unsupervised machine learning.

We applied this protocol for multiple studies in a human-machine learning project. This compared our human-machine learning algorithm with those applying human learning and other machine learning algorithms to predict a variety of health outcomes. Ethical clearance was waived by the Taipei Medical University Joint Institutional Review Board (TMU-JIRB number: N202106025). We followed two guidelines for developing and reporting machine learning predictive models in biomedical research,[9,10] specific for multivariable prediction models instead of those identifying risk or prognostic factors.[10] This protocol aimed to provide a resampling protocol for dimensional reduction resulting a few latent variables, focusing on but not limited to application for developing a machine learning prediction model.

Reagents

Equipment

The data analysis was conducted using R 4.0.2 programming language (R Foundation, Vienna, Austria) in RStudio 1.3.959 (RStudio PBC, Boston, MA, USA). The R packages were all in sync by utilizing Bioconductor 3.11.[11] Since machine learning predictive modeling was the main context of this protocol, an R package of caret 6.0.86 was used for data partition. To facilitate main steps of this protocol, we created an R package rsdr 0.1.0. We also created medhist 0.1.0 to preprocess the sample dataset. These are available for download from this repository https://github.com/herdiantrisufriyana. Details on other R package versions and all of the source codes (vignette) for the data analysis are available in https://github.com/herdiantrisufriyana/resdimer.

A set of hardware requirements may be needed to reproduce our work. This is a single machine with 8 logical processors for the 3.40 GHz central processing unit (CPU) (Core (TM) i7-4770, Intel®, Santa Clara, CA, USA), and 16 GB RAM. But, if the sample size is smaller than that of dataset we used in this protocol, a machine with only 4 logical processors and 4 GB RAM can also be used.

Procedure

1. Split a dataset randomly for a derivation and validation set

Only the derivation set was used to estimate latent variables at the population level. This set was also used later for training set of a prediction model. This would make the model blinded to the distribution of weights for feature representation of any external validation sets.

2. Choose resampling and dimensional reduction methods

We made the rsdr 0.1.0 (an R package) that allows future investigators to conduct a principal component (PC) analysis or singular value decomposition using resampling methods, as described in this protocol. Instead of computing singular values by bootstrapping, as an example, we computed PCs by *k*-fold cross-validation for reasons of simplicity considering a simpler theoretical framework and an achievable computational capacity. To compute PCs by *k*-fold cross-validation, each of β_l, μ_j, and σ_j was inferred from the derivation set only, of which a $(K\text{-}k_m)/K$ part of n instances for $m=[1,2,\cdots,K]$ (equation in Figure 1) was used each time to compute the variance.

3. Standardize each variable with variable-wise average and standard deviation

For every subset of resampling, an X matrix was constructed of $n \times p$ dimensions for $i=[1,2,\cdots,n]$ instances and $j=[1,2,\cdots,p]$ candidate predictors. Each vector was standardized with a column-wise μ_j mean and σ_j standard deviation of all instances for each candidate predictor.

4. Map from higher to lower dimension by finding weights that maximize variances of new dimensions

For every subset of resampling, we mapped each vector $x_{(i)} \in X$ onto a new vector of PC scores $t_{l(i)} = x_{(i)} \bullet \beta_l$ for $l=[1,2,\cdots,q]$ by a matrix β of weight vectors where q ranged up to p. Mapping was used to find estimates of weight vectors that maximized the variance (equation in Figure 1). The l^{th} PC was calculated by subtracting the l^{th}-1 PC from X, then finding the estimate of the l^{th} PC as l^{th}-1 PC.

5. Estimate variable-wise average and standard deviation and weights of the transformation at population level

An estimate of the weight vector β_l was calculated by averaging β_l, μ_j, and σ_j from all $K=10$ of $(K-k_m)/K$ parts. The eigenvalue of the matrix is commonly known for X^TX, which achieves the maximum variance by β as the eigenvector. For each PC, one can find some original variables that are represented by a PC by filtering those with minimum absolute number of estimated weights of the transformation for that PC.

6. Apply the estimated values to standardize and transform original variables into new dimensions

Each of original variables in either derivation or validation set was standardized by subtracting it with the variable-wise estimated average, and subsequently by dividing it with the variable-wise estimated standard deviation. All of the standardized variables were mapped to each of PCs by multiplying each of these variables with the estimated weights. A dot product, which is a PC, was a sum of all the multiplication results.

7. Select a particular number of new dimensions with highest proportions of variance explained

This step is optional for predictive modeling. The recommended number of sample size in relative to the number of candidate predictors was computed for a specific algorithm (e.g. 200 events per variable for random forest).[12] Maximum number of candidate predictors was calculated by dividing the number of events with that number. The PCs were sorted by proportion of variance explained from the highest to the lowest. We selected top PCs as many as the maximum number of candidate predictors. Only top PCs were used for predictive modeling.

Troubleshooting

Step 1 to 6

Problem

Premature stop of computation

Possible reason

Dataset consists of a large sample size or number of variables.

Solution

Consider to use computer with larger memory size or RAM. Alternatively, split table into ones consisting unique sets of variables.

Step 2 to 4

Problem

Unable to conduct dimensional reduction

Possible reason

Variables with non-zero variances are none or only 1 variable.

Solution

Consider to collect other variables.

Step 7

Problem

No selected latent variable

Possible reason

The sample size is too small in relative to the minimum events per variable.

Solution

Consider to collect data with larger sample size.

Time Taken

All steps

Approximate time: 20 to 60 minutes (pre-computed)

Step 1

Approximate time: 5 to 10 minutes (pre-computed)

Step 2 to 4

Approximate time: 5 to 30 minutes

Step 5 to 6

Approximate time: 5 to 10 minutes

Step 7

Approximate time: 1 to 2 minutes

Anticipated Results

The number of latent variables, i.e. PCs, is the same with that of original predictors at maximum. The composition is different among PCs for the weights of original predictors in each PC. Original predictors with larger weights may describe what a PC represents, semantically. One may assign each PC a term that describes original predictors with larger weights in that PC. By this way, this will also improve our interpretation if a PC is considered important for a prediction. Derivation of PCs with resampling may provide a better estimate for these latent variables in population. A latent variable may also imply a novel factor of a disease.

References

1. Fleuren, L.M., et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46, 383-400 (2020).

2. Lee, Y., et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord 241, 519-532 (2018).

3. Gonem, S., Janssens, W., Das, N. & Topalovic, M. Applications of artificial intelligence and machine learning in respiratory medicine. Thorax 75, 695-701 (2020).

4. Bien, N., et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 15, e1002699 (2018).

5. Hannun, A.Y., et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 25, 65-69 (2019).

6. Rajpurkar, P., et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15, e1002686 (2018).

7. Sufriyana, H., et al. Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Med Inform 8, e16503 (2020).

8. Wilkinson, J., et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2, e677-e680 (2020).

9. Luo, W., et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res 18, e323 (2016).

10. Moons, K.G.M., et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 170, W1-w33 (2019).

11. Huber, W., et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12, 115-121 (2015).

12. van der Ploeg, T., Austin, P.C. & Steyerberg, E.W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 14, 137 (2014).

Acknowledgements

The social security administrator for health or *badan penyelenggara jaminan sosial (BPJS) kesehatan* in Indonesia gave permission to access the sample dataset in this protocol (dataset request approval number: 5064/I.2/0421). This protocol was funded by the Ministry of Science and Technology (MOST) in Taiwan (grant number MOST109-2221-E-038-018 and MOST110-2628-E-038-001) and the Higher Education Sprout Project from the Ministry of Education (MOE) in Taiwan (grant number DP2-110-21121-01-A-13) to Emily Chia-Yu Su.

Figures

$$\hat{\beta}_l = \frac{1}{K}\sum_{i\in k^{th}part} \arg\max_{\|\beta\|=1}\left\{\sum_{i=1}^{n}\left(x_{(i)}\cdot\beta_l\right)\right\}^{-k} = \frac{1}{K}\sum_{i\in k^{th}part} \arg\max_{\|\beta\|=1}\left\{\sum_{i=1}^{n}\left(\beta^T X^T X\beta\right)\right\}^{-k}$$

Figure 1

Equation of principal component derivation by k-fold cross validation

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppprotocolresdimer.pdf