Prediction of Adverse Drug Reaction using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset

Yi xin zhao*
School of Statistics and Data Science, Nankai University, Tianjin, Tj 300071, China and Hsbc Business School, Peking University, Shenzhen, Gd 518055, China

Han Yuan*
School of Statistics and Data Science, Nankai University, Tianjin, TJ 300071, China and T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA

Ying Wu†
School of Statistics and Data Science, Nankai University, Tianjin, TJ 300071, China

ABSTRACT

Early prediction of adverse drug reaction (ADR) is crucial in clinical research. The development of electronic medical record (EMR) provides an excellent resource for retrospective studies to extract samples and establish models that can be used for prediction of clinical deterioration. However, classical statistical models like multivariate logistic regression (LR) may result in unreliable predictions when handling unbalanced datasets. To develop a trustworthy model on unbalanced ADR data, we first transformed the EMR including medical notes into numeric variables. Then we introduced support vector machine (SVM), random forest (RF), AdaBoost, XGBoost, and artificial neural network (ANN) to deal with the challenge of high dimensionality. Furthermore, we utilized the ensembling approach to tackle data imbalance. Finally, we analyzed potential model mechanisms to provide interpretability and compared methods from the perspective of procedure elapsed time. The results showed ensembling contributed considerable improvement in prediction ability of various machine intelligence models. Compared with the baseline, RF, AdaBoost and XGBoost presented superiority, and ANN without fine-tuning showed similar competence. The results of this study demonstrated the great potential of machine learning models in medical domain.

CCS CONCEPTS

• **Applied computing**; • **Life and medical sciences**; • **Health informatics**;

KEYWORDS

Electronic Medical Records, Machine Learning, Natural Language Processing, Support Vector Machine, AdaBoost, Random Forest, XGBoost, Artificial Neural Network

*YZ and HY are co-first authors.
†Correspondence: Ying Wu, Email: ywu@nankai.edu.cn

ICMHI 2021, May 14–16, 2021, Kyoto, Japan

ACM ISBN 978-1-4503-8984-6/21/05…$15.00
https://doi.org/10.1145/3472813.3472817

ACM Reference Format:
Yi xin zhao, Han Yuan, and Ying Wu. 2021. Prediction of Adverse Drug Reaction using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset. In *2021 5th International Conference on Medical and Health Informatics (ICMHI 2021), May 14–16, 2021, Kyoto, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3472813.3472817

1 INTRODUCTION

Electronic medical record (EMR) systems have been implemented as storage platforms of patients' information in many hospitals, which are of great potential for retrospective research [1]. Recently, many researchers have utilized machine learning in various tasks of EMR, such as adverse cardiac events prediction, clinical events prediction, and early detection of sepsis [1–5]. However, the common problem of data imbalance in EMR impairs the capabilities of classifiers [6].

In EMR, datasets regarding adverse drug reaction (ADR), which is defined as a harmful or unpleasant reaction resulting from medicinal products [7], are relatively limited in comparison with other types of data [8–12]. Chinese patent drugs (CPD) are contemporary medicinal products in China, whereas, these drugs are sometimes correlated with uncertain adverse reactions [13, 14]. With the exception of limited datasets and related research, the other difficulty of ADR early detection is the extreme data imbalance. For a ICU-related dataset, percentage of minority class samples can be 10%, but in an ADR dataset, it is common to obtain a percentage less than 1%.

To address these limitations, natural language processing (NLP) and other pre-processing methods were first utilized to transform ADR related clinical notes to usable categorical data. Next, we applied traditional logistic regression (LR), several independent machine learning methods (support vector machine (SVM), random forest (RF), AdaBoost, and XGBoost) and a deep learning neural network to make ADR predictions. Furthermore, researchers integrated ensembling with those models to upgrade their performance. Finally, we analyzed potential mechanisms and interpreted our prediction using the variable importance.

The main contributions of this paper are:

1 Provide a pragmatic reference to deal with relatively esoteric EMR related Chinese clinical notes.

2 Explore potential ADR inducement caused by Chinese patent drugs and analyze possible mechanisms.

3 Present the application of ensembling to enhance classification performance in machine learning and deep learning models.

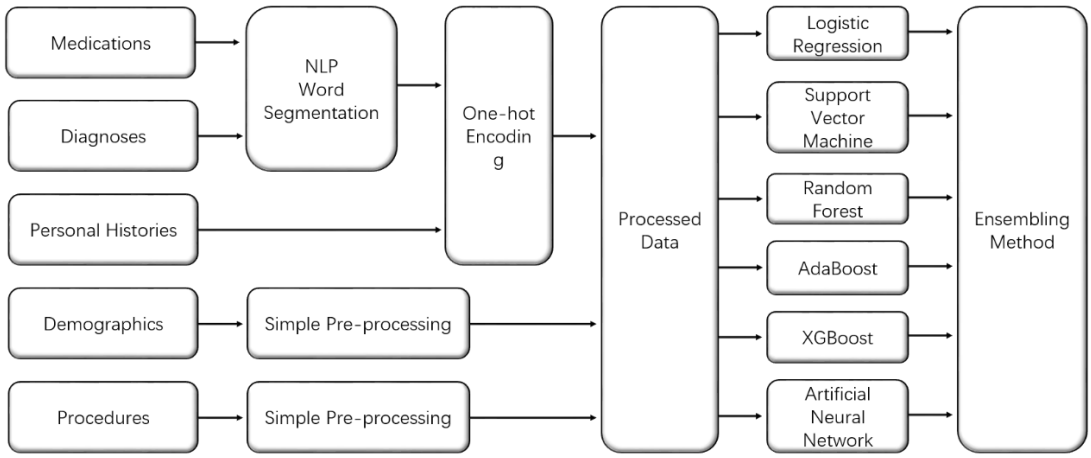

Figure 1: Experimental workflow

Figure 2: Medical Codes Cloud Graph

4 Demonstrate the effectiveness of multiple machine intelligence methods in clinical prediction tasks and give recommendations based on the procedure elapsed time and prediction performance.

2 METHOD

2.1 Word segmentation and one-hot encoding

In the NLP area, word segmentation is performed to transform sentences into individual words, and many scientists contributed a lot to Chinese word segmentation [15–18]. In this project, we took advantage of Jieba, a Chinese word segmentation tool, and a supplementary medical word dictionary from SOGOU, a company targeting input software (https://pinyin.sogou.com/dict/detail/index/15125) [19].

Since the dataset contained categorical data, especially in the medications and diagnoses part, we applied one-hot encoding to process those data [20].

2.2 Support Vector Machine (SVM)

SVM is a binary classification model. Its basic model is defined as the linear classifier with the largest interval in the feature space and learning strategy is to maximize the interval [21].

If x is used to represent data points and y is used to represent categories, the learning goal of a linear SVM model is to find a hyperplane in n-dimensional data space, which can be formulated as

$$w^T * x = 0 \tag{1}$$

Researchers define the support vector as the data point closest to the hyperplane in each class. Specifically, the learning goal is to find the parameter w to maximize the distance of the support vector.

Compared with linear classifiers, SVM maps the input space to a high-dimensional feature space through the kernel function, and constructs the optimal separation superstructure in the high-dimensional feature space.

2.3 Random Forest (RF)

Random forest is a combination of multiple decision tree classifiers, which summarizes the prediction results of each tree through a voting mechanism. Random is reflected in two points. First, each decision tree is trained using only a part of total samples, which is randomly generated with replacement sampling. Second, random split selection means that the features used for each decision tree's split are also from random sampling [22]. Finally, a large quantity of trees is generated and trained independently to determine the output.

2.4 AdaBoost

AdaBoost (Adaptive Boosting) algorithm is a boosting method that combines multiple weak classifiers into a strong classifier, proposed by Yoav Freund et. al. in 1995 [23]. The AdaBoost can be expressed

Table 1: ADR prediction results

Models	Sensitivity	Specificity	F-measure
LR	0.3333	0.9997	0.49996
LR+EB	0.6667	0.7892	0.72276
SVM	0.1667	1.0000	0.28571
SVM+EB	0.4167	0.9630	0.58166
RF	0.4167	1.0000	0.58824
RF+EB	0.6667	0.9899	0.79675
AdaBoost	0.5000	1.0000	0.66667
AdaBoost+EB	0.7500	0.9255	0.82857
XGBoost	0.5000	1.0000	0.66667
XGBoost+EB	0.7500	0.9286	0.82978
ANN	0.0833	1.0000	0.15384
ANN+EB	0.5833	0.8791	0.70133

as follows:

$$f(x) = sign\left(\sum_{k=1}^{K} \alpha_k * G_k(x)\right) \tag{2}$$

where x represents input variables, $G_k(x)$ represents week classifier of round k, α_k represents the weight of the week classifier k.

Its core concept is that the weight of the sample misclassified by the previous weak classifier would be strengthened and used to train the next classifier. In each round of training, the total sample is used to train a new weak classifier to generate a new sample weight. This procedure will iterate until reaching the predetermined error rate or reaching the specified maximum iterations [23].

2.5 XGBoost

XGBoost is the abbreviation of Extreme Gradient Boosting, which is an ensemble machine learning algorithm based on decision trees using Gradient Boost as the framework. XGBoost involves an optimization process that employs additive models and forward stage wise algorithms to achieve learning [24].

The objective function of XGBoost can be expressed as follows:

$$L(\emptyset) = \sum_i I(\widehat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{3}$$

where the first term is the cost function measuring the distance between the true value and the prediction value, and the second term

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{4}$$

is the penalty and γ, λ are parameters.

2.6 Artificial Neural Network (ANN)

ANN refers to a complex network consisting of processing units. Proposed by Minsky, et al., multilayer perceptron is frequently used in classification [25, 26]. A three-layer perceptron with ReLU was utilized here.

2.7 Ensembling models

Ensembling (EB) is a useful algorithm in unbalanced datasets [27, 28]. In this experiment, we fixed the positive training data and sampled randomly 200 times in the negative training data

Table 2: Comparison of machine learning models training time

Models	Training time (seconds)
LR	22.25
LR+EB	41.53
SVM	92.28
SVM+EB	10.44
RF	108.83
RF+EB	32.00
AdaBoost	2433.64
AdaBoost+EB	6414.86
XGBoost	0.84
XGBoost+EB	46.79

to construct subsamples for the training of modules. Datasets for training was denoted as

$$D: \ [d_1, \ d_2, \ d_3, \ \ldots, \ d_{200}]$$

Module m_i was trained on data d_i, and the classification probability vector of the test data j was

$$p_{ij}, \ i = 1, \ 2, \ 3, \ \ldots, \ 200 \tag{5}$$

Then we combined the results of all modules to make final predictions of patient j, which can be calculated as

$$p_j = \frac{\sum_{i=1}^{200} p_{ij}}{200} \tag{6}$$

3 EXPERIMENTAL DETAILS AND RESULTS

3.1 Data Source and Preprocessing

The dataset used was extracted from an EMR Database in China, which has been authorized for research and corresponding publication. It contained 30,703 patients' ADR outcomes, demographic data, procedures, and clinical notes (including personal history, diagnoses, and medications). Since clinical notes included diagnoses and medications in the form of Chinese sentences, so we introduced Natural Language Processing (NLP) methods to process them. We

Table 3: Important variables in models

Variables	XGBoost	Adaboost	RF	LRS
Age	√	√	√	
Course of treatment	√	√	√	
Transfusion stash time	√	√	√	
Infusion speed	√	√	√	
Solvent dosage	√	√	√	
Solvent category	√	√		
Dermatitis	√			√
Xueshuantong	√			
Naodanbai	√	√	√	√
Hypertension	√		√	
Sanqi Xiaozhong Capsule		√		
Fasudil		√		
Chronic Obstructive Pulmonary Disease		√		
Loose-Jointed Pill				√
Enteritis				√
Erysipelas			√	√
Syringomyelia			√	√
Pelvic Infection				√
Aescin			√	

utilized SOGOU Medical Vocabulary as supplementary and applied word segmentation.

Some medications, especially CPD, had multiple expressions or abbreviations. To handle it, several medical specialists joined and made decisions. We employed one-hot encoding in medications and diagnoses. In addition, we calculated the body mass index (BMI) based on weight and height. For missing data issues, we used the Chi-square test and delete unrelated variables whose missing rates larger than 50%. Finally, we matched various types of data through patients' IDs. Ultimate predictors were uploaded online (https://www.researchgate.net/publication/348960256_ADR_APPENDICES_Diagnoses_and_Medications_Chinese-English_Comparison). The training dataset and test dataset was split in the ratio of 7:3.

All research was implemented on PyTorch 1.6.0, Python 3.8, R 4.0.3, and RStudio 1.3. Figure 1 shows our workflow. Figure 2 displays the frequency of medications and diagnoses based on Cloud Graph [29].

3.2 Evaluation Criteria

We utilized sensitivity, specificity, and F-measure [30] as evaluation criteria. Details of prediction results in the test dataset was summarized in Table 1

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$F - measure = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity} \tag{9}$$

Models training time was an important consideration in machine learning. Table 2 shows the procedure elapsed time of models on two Intel (R) Xeon (R) Silver 4210 CPUs (RAM 128GB).

3.3 Significant Variables of Interpretability

We extracted top 10 important features in XGBoost, Adaboost, and Random Forest (RF), and all predictors of logistic regression after the stepwise (LRS) procedure in Table 3

4 DISCUSSION

For all models, descent of specificity was in exchange for increase in sensitivity and F-measure after ensembling, which aligned with our goals of ADR detection. From the perspective of F-measure, ensembling was a simple but powerful method to upgrade prediction capabilities. The increased percentages of F-measure were substantial: LR 44.56%, SVM 103.58%, RF 35.45%, AdaBoost 35.45%, XGBoost 24.28%, and ANN 24.47%. In both raw models and ensembling models, AdaBoost and XGBoost were top models, exhibiting excellent abilities of the boosting family in unbalanced prediction tasks. We can also find that ANN trained with 100 epochs without fine-tuning showed good performance, demonstrating great potentials of neural networks.

Table 2 displays the difference of training time. For SVM and RF, the use of ensembling reduced the procedure time since the time saved in each training modules offset the time consumed by the increases of training modules. In boosting family, XGBoost was more recommended compared with AdaBoost for its efficiency and accuracy in F-measure.

Table 3 describes important variables in different methods, improving the interpretability of this prediction. XGBoost, AdaBoost, and RF obtained relatively similar conclusions. Biologists might be inspired by this result, e.g. Naodanbai was a significant variable in all four models. Some researchers have published some findings in Naodanbai ADR mechanisms such effects of Naodanabi on cerebral blood vessels and peripheral vasodilation caused ADR of the digestive system [31].

5 CONCLUSION

We developed several machine intelligence models for ADR prediction and employed ensembling to upgrade model classification abilities on an extremely unbalanced dataset. Also, multiple interpretable machine learning models provided the variable importance, leaving room for further clinical research. In addition, the pragmatic pipeline of research for electronic medical records including Chinese medical notes was a good reference for other researchers.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant 11701295) and Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, China. YW designed the study. YZ and HY are co-first authors. All authors have no conflict of interest.

REFERENCES

[1] Miao C, Yu A, Yuan H, Gu M, Wang Z. Effect of Enhanced Recovery After Surgery on Postoperative Recovery and Quality of Life in Patients Undergoing Laparoscopic Partial Nephrectomy. Frontiers in Oncology. 2020;10:2165.
[2] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR Workshop Conf Proc. Aug 2016;56:301-318.
[3] Chen P, Dong W, Wang J, Lu X, Kaymak U, Huang Z. Interpretable clinical prediction via attention-based neural network. BMC Med Inform Decis Mak. Jul 9 2020;20(Suppl 3):131. doi:10.1186/s12911-020-1110-7
[4] Huang Z, Dong W, Duan H, Liu J. A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records. IEEE Trans Biomed Eng. May 2018;65(5):956-968. doi:10.1109/tbme.2017.2731158
[5] Lauritsen SM, Kalør ME, Kongsgaard EL, *et al.* Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artif Intell Med. Apr 2020;104:101820. doi:10.1016/j.artmed.2020.101820
[6] Liang Z, Zhang G, Huang JX, Hu QV. Deep learning for healthcare decision making with EMRs. 2014:556-559.
[7] Cho BH, Yu H, Kim K-W, Kim TH, Kim IY, Kim SI. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial intelligence in medicine. 2008;42(1):37-53.
[8] Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. The lancet. 2000;356(9237):1255-1259.
[9] Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.
[10] Beaulieu-Jones BK, Orzechowski P, Moore JH. Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database. Pac Symp Biocomput. 2018;23:123-132.
[11] Wang S, McDermott MBA, Chauhan G, Ghassemi M, Hughes MC, Naumann T. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. presented at: Proceedings of the ACM Conference on Health, Inference, and Learning; 2020; Toronto, Ontario, Canada. https://doi.org/10.1145/3368555.3384469
[12] Peng X, Long G, Shen T, Wang S, Jiang J, Blumenstein M. Temporal Self-Attention Network for Medical Concept Embedding. 2019:498-507.
[13] Qiao Z, Zhang Z, Wu X, Ge S, Fan W. MHM: Multi-modal Clinical Data based Hierarchical Multi-label Diagnosis Prediction. presented at: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020; Virtual Event, China. https://doi.org/10.1145/3397271.3401275
[14] Zheng R, Tao L, Kwong JS, *et al.* Risk factors associated with the severity of adverse drug reactions by Xiyanping injection: A propensity score-matched analysis. Journal of Ethnopharmacology. 2020;250:112424.
[15] Guo X-j, Ye X-f, Wang X-x, *et al.* Reporting patterns of adverse drug reactions over recent years in China: analysis from publications. Expert opinion on drug safety. 2015;14(2):191-198.
[16] Webster JJ, Kit C. Tokenization as the initial phase in NLP. 1992:
[17] Sproat R, Emerson T. The first international Chinese word segmentation bakeoff. 2003:133-143.
[18] Chang P-C, Galley M, Manning CD. Optimizing Chinese word segmentation for machine translation performance. 2008:224-232.
[19] Zhao H, Huang C, Li M. An improved Chinese word segmentation system with conditional random field. 2006:162-165.
[20] Sun J. Jieba Chinese text segmentation. 2014.
[21] Choong ACH, Lee NK. Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. IEEE; 2017:60-65.
[22] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-297.
[23] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
[24] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 1997;55(1):119-139.
[25] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 2016:785-794.
[26] Minsky ML, Papert SA. Perceptrons: expanded edition. 1988;
[27] Ruck DW, Rogers SK, Kabrisky M. Feature selection using a multilayer perceptron. Journal of Neural Network Computing. 1990;2(2):40-48.
[28] Bhowan U, Johnston M, Zhang M, Yao X. Evolving diverse ensembles using genetic programming for classification with unbalanced data. IEEE Transactions on Evolutionary Computation. 2012;17(3):368-386.
[29] Zhang D, Ma J, Yi J, Niu X, Xu X. An ensemble method for unbalanced sentiment classification. IEEE; 2015:440-445.
[30] Lang D, Chien G, LazyData T, *et al.* Package 'wordcloud2'[J]. 2016.
[31] Wong DC, Sweetman C, Ford CM. Annotation of gene function in citrus using gene expression information and co-expression networks. BMC plant biology. 2014;14(1):186.
[32] Xu F, Zhao Z. Literature Analysis of Adverse Reactions of Brain Protein Hydrolysate. Chinese Journal of Hospital Pharmacy. 2007;27(7):1005-1006.