Improving Young Stroke Prediction by Learning with Active Data Augmenter in a Large-Scale Electronic Medical Claims Database

Chen-Ying Hung, Ching-Heng Lin, and Chi-Chun Lee

Abstract—**Electronic medical claims (EMC) database has been successfully used for predicting occurrences of stroke and a variety of other diseases. However, inadequate predictive performances have been observed in cases of rare occurrences due to both insufficient training samples and highly imbalanced class distribution. In this work, our aim is to improve stroke prediction, especially for young age group (25-45 year-old) in a large population-based EMC database (552,898 subjects). We learn a young stroke predictive deep neural network model using a novel active data augmenter. The augmenter selects the most informative EHR data samples from old age stroke patients. This approach achieves 9.3% and 8.2% area under the receiver operating characteristic curve (AUC) value improvements compared to training directly with only young age group data and training all age groups data, respectively. We further provide analyses on the AUC values obtained as a function of the training data size, and the amount and the type of augmented data samples.**

I. Introduction

Making accurate prediction of stroke occurrences can be of great clinical value. The mortality and disability associated with stroke significantly impact lives of patients and their families. An effective data-driven predictive machine learning (ML) algorithm will increase the efficiency of stroke prevention and therefore improve patients' outcomes. As volumes of data grow in healthcare systems, ML technique has been applied to solve several major clinical problems, such as stroke risk assessment [1], detection of heart-failure [2] and death prediction of critical illness patients [3].

Deep learning neural network (DNN) has achieved impressive results in many technical fields [4]. Our recent work also shows that this multi-layered neural network architecture is indeed capable of modeling complex and hidden relationship information recorded in the EHRs to obtain a high general stroke predictive accuracy (with an AUC value more than 0.9) [1]. However, the subgroup stroke prediction performance can severely be degraded, especially for young age group. Most conventional learning algorithms assume a balanced distribution of data classes and adequate samples from the targeted class [5], which makes it difficult to apply in scenarios of young stroke prediction. In the task of young stroke detection, more than 99% of records are from normal patients, and only less than 1% of records are from young patients with stroke. Directly learning an age-specific model is insufficient in obtaining robust predictive results.

Predicting young stroke occurrence, while technically challenging, is an important task for preventing such an acutely ill disease in this age population. In fact, the currently-used clinical stroke risk assessment models, such as the Framingham [6] and the QRISK [7] scoring systems, are developed mainly for high risk patients, especially for old age population. Due to few young stroke records, these models have poor generalizability for young stroke prediction. The ability to accurately predict young stroke occurrence remains limited in current health-care system. In a highly-imbalanced class distribution scenario, while the standard approach in developing a ML model is by up (down)-sampling the training set [8], this approach provides no additional information to the predictive task at hand except at algorithmically smoothing the prediction boundary.

In this work, we propose a novel *active data augmenter* to improve young stroke prediction. The method first performs augmentation to tackle issue of imbalanced distribution by adding data of *related* tasks instead. Since stroke events happen in population with a wide age range, the characteristics across age ranges would likely to possess relevant information to each other. Hence, when focusing on young stroke prediction, we can learn from an augmented data space by including other age range stroke samples. This particular method can be seen as a special case of transfer learning (TL), i.e., TL exploits task-relatedness to improve the performance of each individual task by combining the related information from other tasks [3, 9]. Secondly, a proper mechanism used in selecting the most informative augmented samples play a crucial role, i.e., not all data samples are equally-informative. In this case, instead of conventional data augmentation strategy [10], we propose to use an active-learning procedure in selecting the most informative augmented stroke samples from old age group to train our final young age DNN predictor.

In this study, we demonstrate that the predictive performance of young stroke prediction model can be improved by exploiting stroke-related information from different age groups. We evaluate our active data augmenter approach on a large-scale population-based EHR database (includes a total of 552,898 subjects) for young stroke occurrence prediction. This study details our technical framework. Our active data augmenter is capable of obtaining promising predictive accuracy (AUC values more than 0.8) for young stroke disease prediction. This approach achieves a maximum of 9.3% performance improvement comparing to training with non-augmented young age group data.

CCL is the corresponding author for this work. He is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan (phone: +88635162439. e-mail: cclee@ee.nthu.edu.tw).

CYH is with the Department of Electrical Engineering, National Tsing Hua University, and the Department of Internal Medicine, Taipei Veterans General Hospital, Hsinchu Branch, Hsinchu, Taiwan (e-mail: s881091@ym.edu.tw).

CHL is with the Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan (e-mail: epid@vghtc.gov.tw).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: National Tsing Hua Univ.. Downloaded on October 08,2021 at 02:19:10 UTC from IEEE Xplore. Restrictions apply.

II. METHODS

A. Database and study population

The dataset for this study is extracted from the National Health Insurance Research Database (NHIRD). The National Health Insurance program covers about 99% of Taiwan population. The National Health Research Institute (NHRI) has established a systemic sampling of patient data resulting in the available NHIRD. This database contains de-identified health-care information of over 900,000 patients from 2000 to 2008. The detailed description of the NHIRD was described in [11]. These random samples of patients have been confirmed by the NHRI to be representative of the general population in Taiwan. The NHRI further made data in an anonymous format to protect individual's privacy. The database has been used for several important medical researches [1, 11]. Our study is approved by the Institutional Review Board of Taichung Veterans General Hospital.

We design a task to predict young stroke occurrence using the outpatient department database. Patients aged 25 to 85 years in 2003 are identified from the database. Patients are not eligible for enrollment if they had any types of stroke (International Classification of Diseases, Tenth Revision, Clinical Modification, [ICD-10-CM] code: I60~I69) in the duration of 2000-2003. The selected population is divided into two groups, namely young age group (25-45 year-old) and old age group (45-85 year-old). The outcome event is defined as any ischemic stroke (ICD-10-CM code: I63) recorded in the hospital discharge diagnoses in the inpatient database. Following this exclusion criteria, our final dataset includes a total number of 552,898 patients.

B. Feature engineering

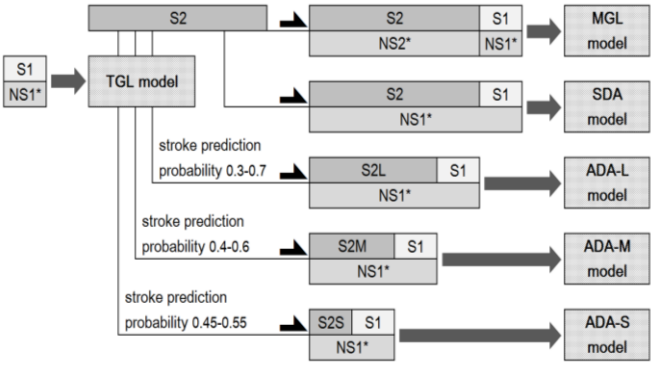
In our previous work, we have established a feature engineering method for deriving predictive analytics using EMCs [1]. We utilize data from the outpatient department within past three years before enrollment to generate features. In brief, we gather the following measurements from the record of an individual patient: demographic information, health insurance related measurements, medication use, and disease diagnosis. The information of medication use is converted to Anatomical Therapeutic Chemical (ATC) code and the disease diagnosis is converted to ICD-10-CM code by using the code-converting sheet provided by the National Health Insurance Bureau of Taiwan.

To generate the final feature vector that can capture both the relevant clinical measurements and temporal information, we utilize time stamp for these measurements. In total, we extract 7,932 features from the dataset. These features can be abstracted as combinations of measurement and temporal dimension. We additionally perform feature selection to identify the most discriminative features as a data preprocessing before training our ML algorithm. We use simple Pearson correlation method to select the most relevant clinical variables. In the current work, we select 200 most important attributes out of 7,932 features.

C. Active data augmenter

We develop a framework of *active data augmenter* with a goal to improve the performance of young age stroke prediction. Firstly, the data augmentation method has been shown to be effective in improving parameter estimation for tasks with imbalanced sample distribution [3, 9]. Since we seek to learn the ML parameters of predicting stroke in young age group, we consider young age population as the *target* task whereas the rest of population as the *source* task. The examples from the source and target tasks can be jointly leveraged to learn the classifier parameter of the target task.

Figure 1. Constructions of expanded augemented datasets for the six learning procedures in this study.



* with random down-sampling process

We additionally include an *active*-learning strategy to select the most informative samples from the source domain. This approach aims to choose the most informative instances and iteratively add them to the training dataset [5]. One of the most commonly used active learning method is an entropy based method [12]. In this work, we utilize a similar entropy-based strategy in selecting source domain stroke samples by finding instances which lie in the most uncertain (high entropy) region based on a trained target-domain DNN probability output on these samples. These data can be augmented to refine classification boundary and achieve a better prediction result. In summary, we propose an active data augmenter in this work to improve our stroke predictive accuracy for the young age population.

D. Experimental procedures

In the young stroke prediction problem, the tasks are primarily defined on the basis of age group. There are two different age groups and each group includes both stroke and non-stroke class, we represent the two age groups as young age (S1, NS1) and old age (S2, NS2), respectively. We design six different learning procedures as the following (as illustrated in Figure 1):

- Target Group Learning (TGL) models directly on young age population: we train a DNN model to perform stroke prediction for the young age group (S1, NS1) dataset. Random down-sampling of NS1 sub-dataset (randomly selects a subset of data with a total amount equals to S1 sub-dataset) is performed to guarantee an almost identical class distribution between stroke and non-stroke cases.
- Multiple Group Learning (MGL) model: we use data from all two age groups (S, NS), where S = S1+S2 and NS = NS1+NS2. Random down-sampling of NS sub-dataset is also performed.

Authorized licensed use limited to: National Tsing Hua Univ.. Downloaded on October 08,2021 at 02:19:10 UTC from IEEE Xplore. Restrictions apply.

TABLE 1. NUMBERS OF PATIENTS AND RECORDS IN EACH GROUP.

	No of all patients	No of patients with stroke	No of records with stroke
Age 25-45 years	288,582	215	3,092
Age 45-85 years	264,316	4,580	100,641
Total population	552,898	4,795	103,733

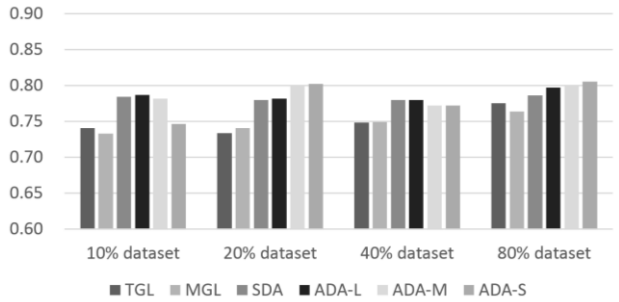
- TGL on young age population with Straightforward Data Augmentation (SDA) approach: this model learns information of stroke patients in all two age groups simultaneously. We train SDA models for young age group (S, NS1), where S = S1+S2. Down-sampling of NS1 sub-dataset is performed.
- Our proposed Active Data Augmenter with Large augmentation (ADA-L): this approach learns information from stroke patients in young age group first and actively select data from stroke patients in other age groups by using this seed model. We train ADA-L models for young age group (S1+S2L, NS1), where S2L are selected from S2 dataset by thresholding the stroke prediction probability from the seed model to be in the range of 0.3-0.7. Down-sampling of NS1 sub-dataset is performed.
- Our proposed Active Data Augmenter with Medium augmentation (ADA-M): this approach train ADA-M models for young age group (S1+S2M, NS1), where S2M were selected from S2 dataset by thresholding the stroke prediction probability from the seed model to be in the range of 0.4-0.6. Down-sampling of NS1 sub-dataset is performed.
- Our proposed Active Data Augmenter with Small augmentation (ADA-S): this approach train ADA-S models for young age group (S1+S2S, NS1), where S2S were selected from S2 dataset by thresholding the stroke prediction probability from the seed model to be in the range of 0.45-0.55. Down-sampling of NS1 sub-dataset is performed.

We target young stroke detection task to perform these six experiments and examine the performance of each model using 5-fold subject independent cross validation. In each cross validation fold, 80% of the total dataset are used as training sets and the rest 20% as testing sets. We also conduct additional subsampling experiments by reducing the training data amount to exam the effect on the amount of available data have on our proposed approach. Full (equals as 80% of total dataset), half (40% of total dataset), quarter (20% of total dataset) and 1/8 (10% of total dataset) of the training sets are used in these subsampling experiments. We use these sub-datasets into the training process of the DNN algorithm and examine the performance of each model on the testing sets. The AUC values are used as a measure of performance.

E. Deep neural network classifier structure

In this work, our aim is to compare TGL, MGL, SDA, and ADA approaches in deriving young stroke prediction model using EHRs. We use a multilayered feed-forward neural network as our classifier. The use of DNN model can automatically learn effective complex feature relationships computed from the EHRs at multiple levels of abstraction [4]. The architecture of our DNN model is composed of three fully connected hidden layers [1]. The number of neurons per hidden layer is equal to the dimension of input data (equals to 200), and hyperbolic tangent is used as the activation function. During the training process, the parameters of the model are randomly initialized. In order to speed up the training process, we apply a simple normalization approach by scaling the feature values to a range between 0 and 1. The DNN is implemented using the Keras (2015, GitHub) toolbox.

Figure2. Experiments for young age stroke prediction by using 10%, 20%, 40% and 80% datasets (AUC values).



III. RESULTS

In this study, the whole dataset includes 552,898 patients (up to 80% is used as training sets and 20% as testing sets in each cross validation process). Table 1 shows the numbers of patients and records in each group. There are 288,582 patients in the young age group and 264,316 patients in the old age group. In the follow-up 5-year period, 4,795 patients in the dataset had stroke events (215 in the young age group and 4,580 in the old age group). The young age group dataset contains a total of 3,092 stroke event records. As mentioned above, a total of 200 features are selected out of 7,932 generated features. The stroke prediction performance of proposed six experiments are compared.

Figure 1 shows the six experimental procedures for young stroke prediction using 10%, 20%, 40% and 80% of the entire datasets as training data. Overall, SDA and ADA-x methods gain higher AUC values than TGL and MGL. TGL and MGL methods have similar model performance. For a clear view of the predictive performance of different data amount, we also compare the effect of these methods with 10%, 20%, 40% and 80% datasets. In the experiment with 10% dataset, the SDA, ADA-L and ADA-M achieve best classification AUC values (0.784, 0.787 and 0.781) than other approaches (TGL: 0.741, MGL: 0.733, ADA-S: 0.746). In the experiment with 20% dataset, ADA-M and ADA-S achieve higher AUC values (0.801 and 0.802) than other approaches (TGL: 0.734, MGL: 0.741, SDA: 0.780, ADA-L: 0.781). In the experiment with 40% dataset, SDA and ADA-L perform better (AUC values: both 0.780) than other approach (TGL: 0.748, MGL: 0.749, ADA-M: 0.772, ADA-S: 0.772). In the experiment with 80% dataset, ADA-L, ADA-M and ADA-S achieve higher AUC values (0.797, 0.800 and 0.805) than other approaches (TGL: 0.775, MGL: 0.763, SDA: 0.786). These results highlight the importance of our proposed active data augmentation approaches when analyzing EHRs for rare diseases. In general, our method achieves higher AUC values than other traditional methods. The best performance happens while using 20% and

Authorized licensed use limited to: National Tsing Hua Univ.. Downloaded on October 08,2021 at 02:19:10 UTC from IEEE Xplore. Restrictions apply.

TABLE 2. THE TRAINING AND AUGMENTED DATA AMOUNTS IN EACH CROSS VALIDATION APPROACH WITH CORRESPONDING MODEL PERFORMANCE FOR YOUNG STROKE PREDICTION.

	No of young stroke records #	AUC of TGL	No of augmented data #	AUC of different methods	Maximal AUC improve ratio
10% data set	284	0.741	SDA: 10,276 ADA-L: 9,500 ADA-M: 4,636 ADA-S: 2,056	MGL: 0.733 SDA: 0.784 **ADA-L: 0.787*** ADA-M: 0.781 ADA-S: 0.746	6.2%
20% data set	474	0.734	SDA: 20,831 ADA-L: 10,998 ADA-M: 5,034 ADA-S: 2,522	MGL: 0.741 SDA: 0.780 ADA-L: 0.781 ADA-M: 0.801 **ADA-S: 0.802***	**9.3%**
40% data set	1,025	0.748	SDA: 41,869 ADA-L: 38,137 ADA-M: 16,043 ADA-S: 8,541	MGL: 0.749 SDA: 0.780 **ADA-L: 0.780*** ADA-M: 0.772 ADA-S: 0.772	4.3%
80% data set	2,444	0.775	SDA: 81,853 ADA-L: 6,380 ADA-M: 3,026 ADA-S: 1,482	MGL: 0.763 SDA: 0.786 ADA-L: 0.797 ADA-M: 0.800 **ADA-S: 0.805***	3.8%

Medium values of the training data per cross validation

* Highest AUC values within groups

80% datasets for training (the best AUC values are around 0.802-0.805).

Table 2 shows the training and augmented data amounts with their corresponding model performance for young stroke prediction. Overall, the performance of ADAs are better than those of TGL or MGL, which indicates that our approach can leverage the intrinsic related information of all age group data to improve the stroke prediction for young age population. The maximal AUC improvement ratio is found in the ADA-S method when training only with 20% of total data (from 0.734 to 0.802, a 9.3% performance improvement comparing to TGL, and from 0.741 to 0.802, an 8.2% performance improvement comparing to MGL). The ADA approach does not get a higher AUC improvement ratio when using more training data. This may reflect that our active selection method can quickly converge the model to achieve good performance with relatively small amount of data augmentation.

IV. DISCUSSION

In summary, we demonstrate that our proposed active data augmenter is effective in deriving young stroke prediction model. Our experimental results show an encouraging 9.3% improvement of AUC values can be achieved. This novel approach in developing automated system for the prediction of young stroke occurrence potentially offers advantages of improving predictive accuracy in other rare disease prediction while using EHRs. The improvement in the AUC values in our experiments may in part result from the balancing between amount of target event records and that of augmented data, as well as leveraging the internal relationship of stroke characteristics between age groups. Our active data augmenter approach can be a promising method to model and extract the implicit correlations among features from EHRs in rare disease prediction tasks.

V. CONCLUSIONS

In this evaluation of applying data augmentation with active-learning selection strategy in using EHRs for young stroke prediction, we demonstrate this method can achieve higher AUC values for prediction of the disease occurrence than non-data-augmented approaches (TGL or MGL approaches). Using EHRs data of other age group can help improve predictive power in young age group. Also, we would explore the next phase of algorithmic design to systematic develop an end-to-end system that incorporates both the prediction module and data augmentation procedure within one neural network architecture. To the best of our knowledge, this work presents one of the first methods in leveraging cross age group information to achieve improved recognition rate for the rare young stroke detection in a large scale population-based EHR database.

REFERENCES

[1] C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3110-3, 2017.

[2] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, et al., "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2530-3, 2015.

[3] C. Karmakar, B. Saha, M. Palaniswami, and S. Venkatesh, "Multi-task transfer learning for in-hospital-death prediction of ICU patients" 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3321-4, 2016.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-44, 2015.

[5] J. Guo, X. Wan, H. Lin, P. Li, G. Liu, and Y. He, "An active learning method based on mistake sampling for large scale imbalanced classification," 2017 International Conference on Service Systems and Service Management, pp. 1-6, 2017.

[6] K. M. Anderson, P. M. Odell, P. W. F. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," American heart journal, vol. 121, no.1, pp. 293-8, 1991.

[7] C. J. Hippisley, C. Coupland, and P. Brindle, "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study," British Medical Journal, vol. 357, pp. j2099, 2017.

[8] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Netw., vol. 21, no. 2-3, pp. 427-36, 2008.

[9] C. Ngufor, S. Upadhyaya, D. Murphree, D. Kor, and J. Pathak, "Multi-task learning with selective cross-task transfer for predicting bleeding and other important patient outcomes", 2015 IEEE International Conference on DSAA, pp. 1-8, 2015.

[10] J. Ding, X. Li, and V. N. Gudivada, "Augmentation and evaluation of training data for deep learning", 2017 IEEE International Conference on Big Data, pp. 2603-11, 2017.

[11] C. Y. Wu, Y. J. Chen, H. J. Ho, Y. C. Hsu, K. N. Kuo, M. S. Wu, et al., "Association between nucleoside analogues and risk of hepatitis B virus-related hepatocellular carcinoma recurrence following liver resection," JAMA, vol. 308, pp. 1906-14, 2012.

[12] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-8, 2008.

Authorized licensed use limited to: National Tsing Hua Univ.. Downloaded on October 08,2021 at 02:19:10 UTC from IEEE Xplore. Restrictions apply.