MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

© 2024 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI231066

Hierarchical Label Distribution Learning for Disease Prediction

Yi REN[a], Jing XIA[a], Ziyi YU[a], Zhenchuan ZHANG[a], Tianshu ZHOU[a], Yu TIAN[b] and Jingsong LI[a,b,1]

[a] *Research Center for Healthcare Data Science, Zhejiang Laboratory, Hangzhou, China*
[b] *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China*

Abstract. The prediction of disease can facilitate early intervention, comprehensive diagnosis and treatment, thereby benefiting healthcare and reducing medical costs. While single class and multi-class learning methods have been applied for disease prediction, they are inadequate in distinguishing between primary and secondary diagnoses, which is crucial for treatments. In this paper, label distribution is suggested to describe the diagnosis, which assigns the description degree to quantify the diagnosis. Additionally, a novel hierarchical label distribution learning (HLDL) model is proposed to make fine-grained predictions based on the hierarchical classification of diseases, taking into account the relationship among diseases. The experimental results on real-world datasets demonstrate that the HLDL model outperforms the baselines with statistical significance.

Keywords. Disease prediction, label distribution learning, hierarchical classification

1. Introduction

With the widespread of Electronic Health Records (EHR), the longitudinal experience of both patients and doctors can be easily recorded and learned by artificial intelligence. Intelligent clinical decision support anticipates the information at the point of care that is specific to the patient and provider needs. Many current efforts in this area have focused on developing specialized predictive models, such as disease prediction, to enable early intervention, diagnosis and informed decision-making.

The most of previous work tackled disease prediction tasks with the single label or multi-label classification methods. For example, Doctor AI [1] observes medical and medication uses to predict future physician diagnoses and medication orders with a temporal model using recurrent neural networks (RNN) model. As the black-box predictions cannot readily be explained to clinician, RETAIN [2] improves the clinically interpretable with the attention mechanism. Recently, Med-BERT [3] adapts the BERT framework to the structured EHR domain, which shows the better prediction accuracy with the pretrained embedding model. While logical labels can indicate the risk of diseases, they cannot determine which ones require more attention or which ones should be prioritized for treatment.

[1] Corresponding Author: Jingsong LI, email: ljs@zju.edu.cn.

In this paper, label distribution [4] is suggested to describe the diagnosis. Label distribution covers the whole labels, representing the degree to which each label describes the instance, i.e. the risk of diseases. Since patients may suffer from multiple diseases and the pathological changes of one organ or system may influence related ones, the model is designed to predict multiple diseases rather than being restricted to specific ones. Additionally, we intend to explore the relation among the diseases through big EHR data to contribute to pathological research. However, prediction becomes more challenging when considering more diseases, due to the class imbalance and the data sparsity. To solve these problems, this paper proposes a novel Hierarchical Label Distribution Learning (HLDL) method that employs a hierarchical neural network to integrate the global and local prediction, as well as medical ontologies. In summary, the contributions of this paper include: 1) Applying label distribution to better describe the risk of illness; 2) Using medical ontologies to enhance the learning process; 3) Introducing the HLDL method to predict diseases.

The rest of the paper is organized as follows. Section 2 introduces the HLDL method. Section 3 reports the experimental results and discusses. Section 4 concludes the paper and recommends future work.

2. Methods

This section introduces the Hierarchical Label Distribution Learning (HLDL) method, which includes a hierarchical neural network that combines global and local predictions. The model also takes into account medical ontologies and the relationship among diseases to enhance the accuracy, interpretability and robustness.

2.1. Hierarchical Label Enhancement

The complete collection of diagnosis codes is denoted as . It is assumed that a medical ontology, such as International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), typically represents the hierarchy of medical concepts in the form of a parent-child relationship, which can be illustrated as a tree-like structures as depicted in figure 1(a). Let $\mathcal{Y}^1 = \{y_1, y_2, \cdots, y_{m_1}\}$ represents the root labels, such as 'certain infectious and parasitic diseases', 'neoplasms' and so on. $\mathcal{Y}^2 = \{y_{m_1+1}, y_{m_1+2}, \cdots, y_{m_1+m_2}\}$ consists of the whole children of the root labels, such as 'intestinal infectious diseases' under the 'certain infectious and parasitic diseases', 'malignant neoplasms' under 'Neoplasms' etc. Therefore, the entire label set can be represented as $= \{\mathcal{Y}^1, \mathcal{Y}^2, \cdots, \mathcal{Y}^K\} = \{y_1, \cdots, y_{m_1}, y_{m_1+1}, \cdots, y_C\}$, where K is number of layers, and C is the total number of labels, $C = \sum_{k=1}^{K} m_t$.

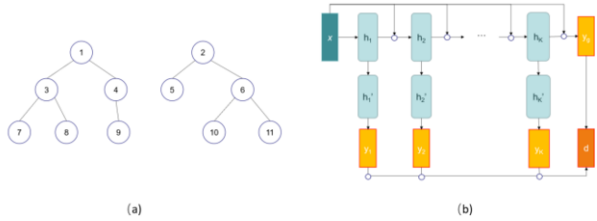

Figure 1. (a) shows a simple example of label set constructed to the tree structure according to some medical concept. (b) is the HLDL architecture.

Label distribution assigns a real number $d_x^{y_c} \in [0,1]$ to each label, representing the risk of disease y_c for patient x, i.e. $\boldsymbol{d}_x^{\boldsymbol{y}} = \{d_x^{y_1}, d_x^{y_2}, \cdots, d_x^{y_C}\} \in \mathbb{R}^C$. In this paper, we quantify the diagnosis with the priority hierarchically:

$$d_x^{y_i(k)} = \frac{r_i^{(k)}}{\sum_{j=1}^{m_k} r_j^{(k)}} \quad , \tag{1}$$

where $r_i^{(t)}$ denotes the ranking of y_i in the *k*-th layer. For example, given a diagnosis $\{y_7, y_8, y_{10}\}$ with priority, then $r_7 = 3$, $r_8 = 2$ and $r_{10} = 1$, according to equation (2), $d_x^{y_7} = 0.5$, $d_x^{y_8} = 0.3$, $d_x^{y_{10}} = 0.2$. Following the medical ontology shown in Figure 1(a), the codes in the second layer are $\{y_3, y_6\}$, thus $d_x^{y_3} = 0.67$ and $d_x^{y_6} = 0.33$. Similarly, the ancestors $d_x^{y_1} = 0.67$ and $d_x^{y_2} = 0.33$. For the irrelevant labels, $d_x^{y_4} = d_x^{y_5} = d_x^{y_9} = d_x^{y_{11}} = 0$. The label distribution $\boldsymbol{d}_x^{\boldsymbol{y}} = \{d_x^{y_1}, d_x^{y_2}, \cdots, d_x^{y_{11}}\} = \{0.67,0.33,0.67,0,0,0.33,0.5,0.3,0,0.2,0\}$.

2.2. Hierarchical Label Distribution Learning

Let $\mathcal{X} = \mathbb{R}^q$ denotes the feature space, and $\mathcal{D} = \mathbb{R}^C$ denotes the label distribution space. Given the training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{d}_i) | i = 1,2, \cdots N\}$, the main purpose of *hierarchical label distribution learning* (HLDL) is to build the model $f: \mathcal{X} \to \mathcal{D}$.

Inspired by the Hierarchical multi-label classification networks (HMCN) [5], HLDL follows the HMCN-F architecture and extends it to label distribution learning, as shown in Figure 1(b). The HLDL architecture consists of two parts: the global prediction network and the local prediction network.

In the global prediction network, the global features are learned with hidden layers: $\boldsymbol{h}_1 = \sigma(\boldsymbol{w}_1 \boldsymbol{x} + \boldsymbol{b}_1)$, where $\boldsymbol{h}_1$ denotes the first hidden layer, learning the global features for the first layer of the hierarchy, σ is a non-linear activation function (e.g., ReLU, tanh), $\boldsymbol{w}_1 \in \mathbb{R}^{s_1 \times q}$ is the weight matrix, and $\boldsymbol{b}_1 \in \mathbb{R}^{s_1}$ is the bias, s_1 is the dimension of $\boldsymbol{h}_1$. Inspired with the skip-connection [6], the second hidden layer is designed as: $\boldsymbol{h}_2 = \sigma(\boldsymbol{w}_2[\boldsymbol{h}_1 \oplus \boldsymbol{x}] + \boldsymbol{b}_2)$, where $[\boldsymbol{h}_1 \oplus \boldsymbol{x}]$ denotes the skip-connection. Accordingly, $\boldsymbol{h}_K = \sigma(\boldsymbol{w}_K[\boldsymbol{h}_{K-1} \oplus \boldsymbol{x}] + \boldsymbol{b}_K)$. Then the global prediction is generated as: $\boldsymbol{y_g} = sigmoid(\boldsymbol{w}_g \boldsymbol{h}_k + \boldsymbol{b}_g)$, where the label distribution $\boldsymbol{y_g} \in \mathbb{R}^C$ represents the overall risk of diseases.

The local neural networks are constructed hierarchically based on each global hidden layer: $\boldsymbol{h}'_k = \sigma(\boldsymbol{w}'_k \boldsymbol{h}_k + \boldsymbol{b}'_k)$, where $\boldsymbol{h}'_k$ learns the local features of the *k*-th layer, and the local prediction for level *k* is $\boldsymbol{y}_k = softmax(\boldsymbol{w}_{lk} \boldsymbol{h}'_k + \boldsymbol{b}_{lk})$.

The final prediction ensembles the local prediction and global prediction: $\widehat{\boldsymbol{d}} = \rho[\boldsymbol{y}_1; \boldsymbol{y}_2; \cdots; \boldsymbol{y}_K] + (1-\rho)\boldsymbol{y_g}$, where $\rho \in [0,1]$ is the parameter that regulates the trade-off regarding local and global predictions.

HLDL minimizes the sum of the local and global loss functions: $L = L_g + \tau L_p$, where $\tau, \epsilon \in \mathbb{R}$ is the trade-off parameter that weights the loss functions. The global loss function is composed of two parts: the widely used Kullback–Leibler (KL) divergence [7] to measure the distance between the predicted and the enhanced label distributions, and the least squares (LS) loss function to preserve the label distribution inherited from the initial logical labels $\boldsymbol{u}_i$:

$$L_g = L_{kl} + L_{ls} = \sum_{i=1}^{n} \boldsymbol{d}_i ln \frac{\boldsymbol{d}_i}{\boldsymbol{g}_i} + \sum_{i=1}^{n} \|\boldsymbol{g}_i - \boldsymbol{u}_i\|^2. \tag{10}$$

In addition to the Kullback–Leibler divergence, we also include the hierarchical violation for local loss function:

$$L_p = L_{kl} + L_{hv} = \sum_{i=1}^{n} \boldsymbol{d}_i ln \frac{\boldsymbol{d}_i}{\boldsymbol{p}_i} + \sum_{i=1}^{n} \boldsymbol{A}(\boldsymbol{1} - \boldsymbol{p}_i^{\mathrm{I}}) \boldsymbol{p}_i, \tag{11}$$

where $\boldsymbol{A} \in \{0,1\}^{c \times c}$ is the relationship matrix, i.e., $a_i^j = 1$ means y_j is the ancestor of y_i, otherwise, $a_i^j = 0$.

3. Results

3.1. Datasets

To evaluate the performance of HLDL, we use two publicly available real-world EHR datasets MIMIC-III [8] and MIMIC-IV [9]. In the MIMIC-III dataset, we collected medical records of 58929 hospital admissions of 46,517 intensive care unit (ICU) patients. Among them, 7,499 patients had more than two hospital admissions, of which 4,499 were used for training, 1500 were chosen for validation, and 1500 were chosen for testing. The remaining medical records of 39,018 patients were used for pretraining. The MIMIC-III dataset contains 1,233 3-digit ICD-9 codes, which can be grouped into 19 classes and 158 subclasses. For the MIMIC-IV dataset, we collected medical records of 453,905 hospital admissions of 190,173 patients. Among the 83,660 patients who has more than two hospital admissions, 16,732 were used for testing, 16,732 were used for validation, and 50,186 were used for training. The remaining 106,513 medical records were used for pretraining. The MIMIC-IV dataset contains two versions of the coding system: ICD-9 and ICD-10. Despite code mappings, we treated them as different classifications because ICD-10 codes are more detailed. The MIMIC-IV dataset contains 3,502 3-digit ICD codes, which can be grouped into 41 classes and 419 subclasses. Following Med-BERT, we embed the structured EHR data as the features of patients, which consists of ICD codes, procedure codes and drug codes in medical history.

3.2. Experimental Results

HLDL is compared with the state-of-the-art methods: Doctor AI [1], RETAIN [2,3], HMCN [5], and SA-BFGS [4]. As Doctor AI and RETAIN are two widely used methods for disease prediction. HMCN is a classical hierarchical multi-label classification method. And SA-BFGS is one of the LDL method which shown better performance. For the whole methods, the embedding size is 64, and the hidden units are selected from [256, 516, 1024, 2048], other settings are following the original papers, such as the activation functions etc. As for the HLDL, we set $\rho = 0.5$, $\tau = 2$. The activation function is ReLU, and the optimization is Adam. The iteration will stop when the accuracy on validation set does not improve after eight epochs. At this point, while the model has the best performance will be saved. Then the results of the models evaluated on testing set are recorded.

To evaluate the methods trained with multi-label and label distribution equally, Top-k recall [1] is used, which mimics the behavior of doctors conducting differential diagnosis. The experimental results are shown in Table1. All methods, except for SA-BFGS which is designed for LDL, are trained with both multi-label (recorded as the name) and label distribution (recorded with '+LDL'). As the Doctor AI, RETAIN, and SA-BFGS cannot deal with hierarchical structure, we flatten the labels. The results in Table 1 show that HLDL outperformed all other methods.

Table 1. The performance of compared methods on disease prediction task.

Methods	MIMIC-III			MIMIC-IV		
	top@10	top@30	top@50	top@10	top@20	top@30
Doctor AI	60.16	53.88	53.35	47.95	44.53	46.00
Doctor AI+LDL	63.21	56.56	55.72	49.18	46.06	47.26
RETAIN	61.25	53.62	53.02	48.89	44.54	45.97
RETAIN+LDL	63.45	57.35	56.44	48.64	45.23	46.67
HMCN	65.44	57.67	56.58	55.37	50.40	51.36
HMCN+LDL	68.77	61.54	60.31	58.02	53.83	54.86
SA-BFGS	58.89	52.74	52.72	40.90	36.81	36.16
HLDL	69.25	60.34	59.19	59.76	54.79	55.56
HLDL+LDL	**70.73**	**62.55**	**61.35**	**61.17**	**55.84**	**56.19**

4. Discussion

As the experimental results shown on Table 1, we can see our proposed method HLDL outperformed all other methods, followed by HMCN. This indicates that the hierarchical neural network is effective for multiple disease prediction, and our optimizations such as the skip-connection module and the loss function contribute to improving the prediction accuracy. The results confirm that the proposed method can make more accurate predictions with considering the relationship among disease and making full use of the hierarchical ontologies. Moreover, the methods trained with label distribution show better performance than those trained with multi-label datasets, which supports our suggestion that label distribution can better describe diagnosis.

5. Conclusions

The proposed HLDL method, which utilizes a hierarchical neural network ensemble to perform both global and local predictions, as well as incorporating medical ontologies and disease relations to improve model accuracy, interpretation, and robustness. The use of label distribution allows for more details of the risk of diseases for a given patient, leading to more accurate predictions. The experimental results on two real-world datasets demonstrate the superiority of HLDL over other methods and highlight the importance of label distribution in diagnosis prediction. Overall, HLDL has the potential to be a valuable tool for healthcare professionals in making more informed decisions regarding patient care.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62103382), the Zhejiang Provincial Natural Science Foundation of China (No. LQ21F030004), the China Postdoctoral Science Foundation (Grant No. 2021M692959), the Key Research Project of Zhejiang Lab (No. 2022ND0AC01) and the Key Program of Science and Technology Development Program of Hangzhou, China (Grant No. 2020ZDSJ0885).

References

[1] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. In: Finale DV, Jim F, David K, Byron W, Jenna W, editors. Proceedings of the 1st Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research: PMLR; 2016 Dec 10. p. 301-18.

[2] Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. Adv Neural Inf Process Syst. 2016; 29, doi: 10.48550/arXiv.1608.05745.

[3] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021 May; 4(1):86, doi: 10.1038/s41746-021-00455-y.

[4] Geng X. Label distribution learning. IEEE Trans Knowl Data Eng. 2016 Mar;28(7):1734-48, doi: 10.1109/TKDE.2016.2545658.

[5] Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks. International conference on machine learning; 2018 Jul 3; PMLR. p. 5075-84.

[6] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770-8.

[7] Cha S-H. Comprehensive survey on distance/similarity measures between probability density functions. City. 2007;1(2):1.

[8] Pollard TJ, Johnson AE. The MIMIC-III Clinical Database. 2016.

[9] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet. Available online at: https://physionet.org/content/mimiciv/10/. Accessed August 23, 2021.