Leveraging Hierarchical Category Knowledge for Data-Imbalanced Multi-Label Diagnostic Text Understanding

Shang-Chi Tsai   Ting-Yun Chang   Yun-Nung Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{r06946004,r06922168}@ntu.edu.tw y.v.chen@ieee.org

Abstract

Clinical notes are essential medical documents to record each patient's symptoms. Each record is typically annotated with medical diagnostic codes, which means diagnosis and treatment. This paper focuses on predicting diagnostic codes given the descriptive present illness in electronic health records by leveraging domain knowledge. We investigate various losses in a convolutional model to utilize hierarchical category knowledge of diagnostic codes in order to allow the model to share semantics across different labels under the same category. The proposed model not only considers the external domain knowledge but also addresses the issue about data imbalance. The MIMIC3 benchmark experiments show that the proposed methods can effectively utilize category knowledge and provide informative cues to improve the performance in terms of the top-ranked diagnostic codes which is better than the prior state-of-the-art. The investigation and discussion express the potential of integrating the domain knowledge in the current machine learning based models and guiding future research directions.

1 Introduction

Electronic health records (EHR) usually contain clinical notes, which are free-form text generated by clinicians during patient encounters, and a set of metadata diagnosis codes from the International Classification of Diseases (ICD), which represent the diagnoses and procedures in a standard way. ICD codes have a variety of usage, ranging from billing to predictive modeling of the patient state (Choi et al., 2016). Automatic diagnosis prediction has been studied since 1998 (de Lima et al., 1998). Mullenbach et al. (2018) pointed out the main challenges of this task: 1) the large label space, with over 15,000 codes in the ICD-9 taxonomy, and over 140,000 codes in the newer ICD-10 taxonomies (Organization et al., 2007), and 2) noisy text, including irrelevant information, misspellings and non-standard abbreviations, and a large medical vocabulary. Several recent work attempted at solving this task by neural models (Shi et al., 2017; Mullenbach et al., 2018).

However, most prior work considered the output labels independently, so that the codes with few samples are difficult to learn (Shi et al., 2017). Therefore, Mullenbach et al. (2018) proposed an attentional model to effectively utilize the textural forms of codes to facilitate learning. In addition to textual definitions of codes, the *category* domain knowledge may provide additional cues to allow the codes under same category to share parameters, so the codes with few samples can benefit from it. To effectively utilize the *category knowledge* from the ICD codes, this paper proposes several refined category losses and incorporate them into convolutional models and then evaluate the performance on both MIMIC-3 (Johnson et al., 2016) and our internal dataset. The experiments on MIMIC shows that the proposed knowledge integration model significantly improves the previous methods and achieves the state-of-the-art performance, and the improvement can also be observed in our internal dataset. The idea is similar to the prior work (Singh et al., 2018), which considered the keyword hierarchy for information extraction from medical documents, but our work focuses on leveraging domain knowledge for clinical code prediction. Our contributions are three-fold:

- This paper first leverages external domain knowledge for diagnostic text understanding.
- The paper investigates multiple ways for incorporating the domain knowledge in an end-to-end manner.
- The proposed mechanisms improve all prior

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 39–43
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/D19-62

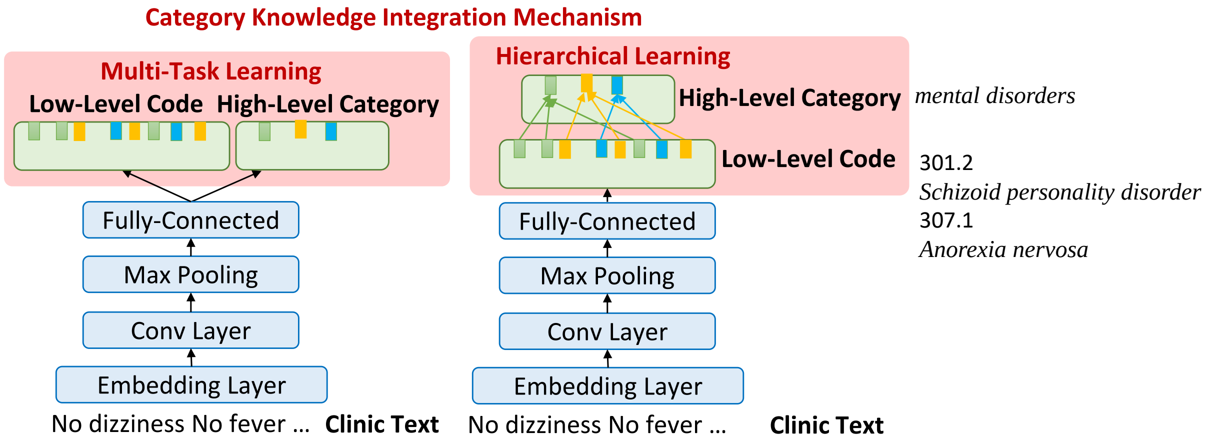

Figure 1: The architecture with the proposed category knowledge integration.

models and achieves the state-of-the-art performance on the benchmark MIMIC dataset.

2 Methodologies

Given each clinical record in EHR, the goal is to predict the corresponding diagnostic codes with the external hierarchical category information. This task is framed as a multi-label classification problem. The proposed mechanism is built on the top of various convolutional models to further combine with the category knowledge. Below we introduce the previously proposed convolutional models which are used for latter comparison in the experiment and detail the mechanism that leverages hierarchical knowledge.

2.1 Convolutional Models

There are various models for sequence-level classification, and this paper focuses on two types of convolutional models for investigation. The models are described as follows. Note that the proposed mechanism is flexible for diverse models.

TextCNN Let $x_i \in \mathbb{R}^k$ be the k-dimensional word embedding corresponding to the i-th word in the document, represented by the matrix $X = [x_1, x_2, ..., x_N]$, where N is the length of the document. TextCNN (Kim, 2014) applies both convolution and max-pooling operations in one dimension along the document length. For instance, a feature c_i is generated from a window of words $x_i, x_{i+1}, ..., x_{i+h}$, where h is the kernel size of the filters. The pooling operation is then applied over $c = [c_1, c_2, ..., c_{n-h+1}]$ to pick the maximum value $\hat{c} = \max(c)$ as the feature corresponding to this filter. We implement the model with kernel size = 3,4,5, considering different window sizes of words.

Convolutional Attention Model (CAML) Because the number of samples of each code is highly unbalanced, it is difficult to train each label with very few samples. To resolve this issue, the CAML model utilizes the descriptive definition of diagnosis codes, which additionally applies a per-label attention mechanism, where the additional benefit is that it selects the n-grams from the text that are most relevant to each predicted label (Mullenbach et al., 2018).

2.2 Knowledge Integration Mechanism

Considering the hierarchical property of ICD codes, we assume that using the higher level labels could learn more general concepts and thus improve the performance. For instance, the definitions of ICD-9 codes **301.2** and **307.1** are "*Schizoid personality disorder*" and "*Anorexia nervosa*" respectively. If we only use the labels given by the dataset, they are seen as two independent labels; however, in the ICD structure, both **301.2** and **307.1** belong to the same high-level category "*mental disorders*". The external knowledge shows that category knowledge provides additional cues to know code relatedness. Therefore, we propose four types of mechanisms that incorporate hierarchy category knowledge to improve the ICD prediction below.

Cluster Penalty Motivated by Nie et al. (2018), we compute two constraints to share the parameters of the ICD codes under the same categories. The between-cluster constraint, $\Omega_{between}$, indicates the total distance of parameters between mean of all ICD codes and the mean of each category.

$$\Omega_{between} = \sum_{k=1}^{K} \left\|\bar{\theta_k} - \bar{\theta}\right\|^2, \qquad (1)$$

where $\bar{\theta}$ is the mean vectors of all ICD codes, $\bar{\theta_k}$ is the mean vector of the k-th category. The within-cluster constraint, Ω_{within}, is the distance of parameters between the mean of each category and its low-level codes.

$$\Omega_{within} = \sum_{k=1}^{K} \sum_{i \in \mathcal{J}(k)} \left\|\theta_i - \bar{\theta_k}\right\|^2, \qquad (2)$$

where $\mathcal{J}(k)$ is a set of labels that belong to the k-th category. $\Omega_{between}$ and Ω_{within} are formulated as additional losses to enable the model to share parameters across codes with the same categories.

Multi-Task Learning Considering that the high-level category can be treated as another task, we apply a multi-task learning approach to leverage the external knowledge. This model focuses on predicting the low-level codes, y_{low}, as well as its high-level category, y_{high}, individually illustrated in Figure 1.

$$y_{high} = W_{high} \cdot h + b_{high} \qquad (3)$$

where $W_{high} \in \mathrm{I\!R}^{N_{high} \times d}$, N_{high} means the number of high-level categories, and d is the dimension of hidden vectors derived from CNN.

Hierarchical Learning We build a dictionary for mapping our low-level labels to the corresponding high-level categories illustrated in Figure 1. To estimate the weights for high-level categories, y_{high}, two mechanisms are proposed:

- Average meta-label: The probability of the k-th high-level category can be approximated by the *averaged* weights for low-level codes that belong to the k-th category.

$$y_{high} = \frac{1}{k} \sum y_{low}^{k} \qquad (4)$$

- At-least-one meta-label: Motivated by Nie et al. (2018), meta labels are created by examining whether any disease label for the k-th category has been marked as tagged, where the high-level probability is derived from the low-level probability of disease labels.

$$y_{high} = 1 - \prod_{k} (1 - y_{low}^{k}) \qquad (5)$$

	MIMIC-3		Internal
	Full	50	200
# training documents	47,424	8,067	17,762
mean length of texts	1,485	1,530	50.35
vocabulary size	51,917	51,917	25,654
OOV rate	0.137	0.137	0.373
# labels	8,922	50	200
mean number of labels	15.9	5.7	1.7

Table 1: Dataset comparison and statistics. From the full set of the internal data (1495 labels) to 200, only 6.0% of data points are discarded.

2.3 Training

The knowledge integration mechanisms are built on top of the multi-label convolutional models, which treat each ICD label as a binary classification. The predicted values for high-level categories come from the proposed mechanisms. Considering that learning low-level labels directly is difficult due to the highly imbalanced label distribution, we add a loss term indicating the high-level category in order to learn the general concepts in addition to the low-level labels, and train the model in an end-to-end fashion. Note that the high-level loss is set as $loss_{high} = \Omega_{between} + \Omega_{within}$ for cluster penalty and the binary log loss for other methods.

$$loss = loss_{low} + \lambda \cdot loss_{high}, \qquad (6)$$

where λ is the parameter to control the influence of the knowledge category and we choose $\lambda = 0.1$.

3 Experiments

In order to measure the effectiveness of the proposed methods, the following experiments are conducted.

3.1 Setup

We evaluate our model on two datasets, one is the benchmark MIMIC-3 data and another is the dataset collected by National Taiwan University Hospital (NTUH). MIMIC-3 (Johnson et al., 2016) is a benchmark dataset, where the text and structured records from a hospital ICU. We use the same setting as the prior work (Mullenbach et al., 2018), where 47,724 discharge summaries is for training, with 1,632 summaries and 3,372 summaries for validation and testing, respectively. We also obtain a subdataset from original MIMIC3-Full, called MIMIC3-50, which has the top 50 high frequency labels. NTUH dataset is collected

MIMIC3-50		**P@1**	**P@3**	**P@5**	**MAP**	**Macro-F**	**Micro-F**	**Macro-AUC**	**Micro-AUC**
CNN (Shi et al., 2017)		82.8	71.2	61.4	72.4	57.9	63.0	88.2	91.2
+ Cluster Penalty		83.5†	71.9†	62.4†	73.1†	58.3†	63.7†	88.5†	91.3†
+ Multi-Task		83.5†	71.3†	61.9†	72.5†	57.6	62.8	88.1	91.1
+ Hierarchical	*avg*	**84.5**†	**72.1**†	**62.4**†	**73.5**†	**58.6**†	**64.3**†	**88.9**†	**91.4**†
	at-least-one	83.4†	72.1†	62.4†	73.4†	58.5†	63.8†	88.4†	91.3†
MIMIC3-Full		**P@1**	**P@3**	**P@8**	**P@15**	**Macro-F**	**Micro-F**	**Macro-AUC**	**Micro-AUC**
CNN (Shi et al., 2017)		80.5	73.6	59.6	45.4	3.8	42.9	81.8	97.1
+ Cluster Penalty		80.9†	74.0†	59.5	45.2	3.3	40.5	82.1†	97.0
+ Multi-Task		**82.8**†	**75.8**†	**61.5**†	**46.6**†	3.6	**43.9**†	**83.3**†	**97.3**†
+ Hierarchical	*avg*	79.0	73.1	59.2	45.2	**4.3**†	42.7	83.0†	97.1
	at-least-one	82.1†	74.3†	59.7†	44.9	2.6	42.0	80.3	96.7
CAML (Mullenbach et al., 2018)		89.6	83.4	69.5	54.6	6.1	51.7	88.4	98.4
+ Cluster Penalty		88.4	82.4	68.8	54.0	5.4	51.2	87.5	98.3
+ Multi-Task		**89.7**†	83.4	69.7†	54.8	6.9†	52.3†	88.8†	98.5†
+ Hierarchical	*avg*	89.6	**83.5**†	**70.9**†	**56.1**†	**8.2**†	**53.9**†	**89.5**†	**98.6**†
	at-least-one	89.4	83.3	69.5	54.8†	6.2†	51.7	88.3	98.4

Table 2: The results on MIMIC-3 data (%). † indicates the improvement over the baseline.

Data-200	**Macro-F1**	**Micro-F1**
CNN	7.6	39.8
+ Multi-Task	**11.7**†	41.6†
+ Hierarchical (avg)	9.2†	**44.1**†
CAML	6.2	42.6
+ Multi-Task	14.5†	44.7†
+ Hierarchical (avg)	**18.4**†	**45.7**†

Table 3: The results on NTUH data.

from an internal hospital, where each record includes narrative notes describing a patients stay and associated diagnostic ICD-9 codes. There are total 1,495 ICD-9 codes in the data, and the distribution is highly imbalanced. Our data is noisy due to typos and different writing styles, where the OOV rate is 0.373 based on the large vocabulary obtained from PubMed and PMC. As shown in Table 1, our data, Internal-200, is more challenging due to much shorter text inputs and higher OOV rate compared with the benchmark MIMIC-3 dataset. We split the whole set of 25,375 records from Internal-200 into 17,762 as training, 2,537 as validation, and 5,076 as testing.

All models use the same setting as the prior work (Kim, 2014; Mullenbach et al., 2018) and use skipgram word embeddings trained on PubMed[1] and PMC[2] (Mikolov et al., 2013). We evaluate the model performance using metrics for the multi-label classification task, including precision at K, mean average precision (MAP), and micro-averaged, macro-averaged F1 and AUC.

3.2 Results

The baseline and the results of adding the proposed mechanisms are shown in Table 2. For MIMIC3-50, all proposed mechanisms achieve the improvement for almost all metrics, and the best one is from the hierarchical learning with average meta-label. The consistent improvement indicates that category knowledge provides informative cues for sharing parameters across low-level codes under the same categories. For MIMIC3-Full, our proposed mechanisms still outperform the baseline CNN model, and the best performance comes from the one with multi-task learning. The reason may be that multi-task learning has more flexible constraints compared with hierarchical learning, and it is more suitable for this more challenging scenario due to data imbalance. In addition, the proposed knowledge integration mechanisms using multi-task learning or hierarchical learning with average meta-label are able to improve the prior state-of-the-art model, CAML (Mullenbach et al., 2018), demonstrating the superior capability and the importance of domain knowledge.

To further investigate the model effectiveness, we perform the experiments on the NTUH dataset in Table 3. Due to shorter clinical notes and higher OOV rate, this dataset is more challenging and the results are lower than the ones in MIMIC-3. Nevertheless, the proposed methods still improve the performance by integrating category knowledge using multi-task learning or hierarchical learning with average meta-label. In sum, our proposed category knowledge integration mechanisms are

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] https://www.ncbi.nlm.nih.gov/pmc/

capable of improving the text understanding performance by combining the domain knowledge with neural models and achieve the state-of-the-art results.

3.3 Qualitative Analysis

From our prediction results, we find that our proposed mechanisms tend to predict more labels than the baseline models for both CNN and CAML. Specifically, our methods can assist models to consider more categories from shared information in the hierarchy. The additional codes often contain the right answers and sometimes are in the correct categories but not exactly matched. Moreover, our mechanisms have the capability of correcting the wrong codes to the correct ones which are under the same category. The appendix provides some examples for reference.

4 Conclusion

This paper proposes multiple mechanisms using the refined losses to leverage hierarchical category knowledge and share semantics of the labels under the same category, so the model can better understand the clinical texts even if the training samples are limited. The experiments demonstrate the effectiveness of the proposed knowledge integration mechanisms given the achieved state-of-the-art performance and show the great generalization capability for multiple datasets. In the future, we plan to analyze the performance of each label, investigating which label can benefit more from the proposed approaches.

Acknowledgements

We would like to thank reviewers for their insightful comments on the paper. This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant 108-2636-E002-003 and 108-2634-F-002-015.

References

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1101–1111.

Allen Nie, Ashley Zehnder, Rodney L Page, Arturo L Pineda, Manuel A Rivas, Carlos D Bustamante, and James Zou. 2018. Deeptag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain. *arXiv preprint arXiv:1806.10722*.

World Health Organization et al. 2007. International statistical classification of diseases and related health problems: tenth revision-version for 2007. *http://apps. who. int/classifications/apps/icd/icd10online/*.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Gaurav Singh, James Thomas, Iain Marshall, John Shawe-Taylor, and Byron C. Wallace. 2018. Structured multi-label biomedical text tagging via attentive neural tree decoding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2837–2842. Association for Computational Linguistics.