Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210178

A Deep Learning Framework for Automated ICD-10 Coding

Abdelahad CHRAIBI[a,1], David DELERUE[a], Julien TAILLARD[a], Ismat CHAIB DRAA[a], Régis BEUSCART[b] and Arnaud HANSSKE[c]
[a] *ALICANTE SARL, France*
[b] *ULR2694, Lille University, France*
[c] *KASHMIR-DataReuse Lab, Catholic Lille University (UCL), France*

Abstract. The International Statistical Classification of Diseases and Related Health Problems (ICD) is one of the widely used classification system for diagnoses and procedures to assign diagnosis codes to Electronic Health Record (EHR) associated with a patient's stay. The aim of this paper is to propose an automated coding system to assist physicians in the assignment of ICD codes to EHR. For this purpose, we created a pipeline of Natural Language Processing (NLP) and Deep Learning (DL) models able to extract the useful information from French medical texts and to perform classification. After the evaluation phase, our approach was able to predict 346 diagnosis codes from heterogeneous medical units with an accuracy average of 83%. Our results were finally validated by physicians of the Medical Information Department (MID) in charge of coding hospital stays.

Keywords. Medical informatics, Natural Language Processing, Deep Learning, Automated Coding

1. Introduction

To standardize and uniformize the protocol of reimbursement demands' treatment, the French insurance companies adopt the ICD-10, one of the most widely used classification system for classifying diagnoses and procedures. This system provides a codification of diseases and procedures, including a wide variety of signs, symptoms, and causes of injuries or diseases. ICD is used to translate diagnoses of diseases or other health problems into alphanumeric codes to facilitate data storage, research and analysis [1]. One EHR is attributed to each patient. At the end of an episode of care, all medical information of a patient result in textual documents such as discharge summaries, operative reports and progress notes, based on which diagnosis codes and procedure codes are assigned manually by human coders [2]. In this work, we considered the discharge summary (DS) which is a free-text document edited in the moment of patient departure. It describes the main health information about a patient during his hospitalization and provides final diagnosis, main interventions, treatments, etc. These free text notes have the advantage to be valuable sources of information and generally not subject to redaction's rules. Consequently, the use of abbreviations and synonyms, the grammatical errors, the spelling mistakes, the semantic ambiguities, the negation sentences, etc. entails issues and causes imprecision when manually matching ICD codes

[1] Corresponding Author, Abdelahad CHRAIBI; E-mail: abdelahad.chraibi@alicante.fr

to those descriptions. The automated coding methods become more attractive to assist the physicians when reporting diagnoses codes to become more productive, efficient, accurate, and consistent. It can be considered as a classification problem corresponding to a variety of NLP approaches that transform narrative text of DSs into structured data frames before applying Machine Learning (ML) or DL techniques to determine the appropriate assignment of codes. Since 1990, several researchers have been interested to creating an automated system for assigning ICD codes to clinical reports. Depending on the type of data, multiple methods have been applied ranging from simple regression to advanced DL approaches with the objective to maximize performance and improve the quality of cares [3]. In the last decade, Authors in [4] experimented probabilistic topic models on collected DSs within urology and hematology services by comparing models issued from both classical ML approaches (Decision Tree, Naïve Bayes, and SVM) and modern NLP approaches (supervised Latent Dirichlet Allocation (LDA) and labeled LDA). In [5] authors presented a multimodal machine learning model to cope with different type of data including unstructured text, semi-structured text and structured tabular data for which Text Convolutional Neural Network (CNN), Bidirectional LSTM and decision trees were respectively applied. [6] present a Hierarchical Attention bidirectional Gated Recurrent Unit for multi-label ICD-10 code assignment. Compared it to SVM-based one-vs-all model, continuous bag-of-words model and CNN model, the HA-GRU yield the most promising results when combined with careful data preparation steps. More recently, Almagro et al. [7] compared algorithms based on binary outputs, groups of subsets and extreme classification to automatically code Spanish electronic discharge summaries. [8] considered 6 principal diseases and proposed an unsupervised knowledge integration algorithm to analyze clinical narrative notes via semantic relevance assessment. [9] conceived a semi-automatic ICD-10 coding system based on regular expressions with promising results. The aim of this work is to assist hospitals in coding their stays by proposing an intelligent system based on NLP and DL approaches to automatically and accurately translates the free-text diagnosis descriptions into ICD codes. The rest of the paper is organized as follows: section 2 presents the dataset and methods. Section 3 and 4 give experimental results and discussion. Finally, section 5 presents conclusions and perspectives.

2. Methods

This study has been conducted using a database with more than 134K EHRs. The selected DSs correspond to 108K hospital admissions between 01-01-2016 and 31-03-2019 for which ICD-10 was used to assign diagnosis codes. All reports have been automatically de-identified using NLP technics including entity recognition techniques and regular expressions to constitute a large corpus of raw French text covering multiple medical specialties. This step has no influence on the accuracy of the final classification since even with the original corpus the results are the same. Following MID's recommendations, the DSs has been split in four groups according to the Length of Stay (LoS): group_0 (LoS = 0 night), group_1_2 (1 night <= LoS <= 2 nights), group_3_4 (3 nights <= LoS <= 4 nights) and group_5+ (LoS >= 5 nights). Some statistics related to the used dataset are presented in table 1, where for each group of stays the number of available DSs, the number of distinct used codes in the studied period and the number of initial selected codes for classification being coded at least 30 times over the period are presented in the second, third and fourth columns, respectively.

Table 1. Statistics related to the dataset

Group of reports	Number of DS	Count of distinct codes	Initial selected codes
Group_0	35308	1661	145
Group_1_2	20257	1866	109
Group_3_4	14489	1675	111
Group_5+	38475	2459	248

To produce the classification model, we adopt a pipeline involving some major steps: text preprocessing, feature engineering, data splitting, parameter tuning, and creation of classification model. A first training iteration is performed using a list of ICD codes being selected according to their occurrence frequency in the corpus of DSs (at least 30 occurrences) to have representative data for each target in the training and the test datasets. Other code selection iterations are applied at the end of each 'model validation' phase based on the F1-Score threshold to keep the top predictable codes. For a classification, almost all machine learning algorithms need to take input in numerical inputs and have not the capacity to interpret other types of data if not transformed. In order to transform the text into numerical values, we used a hybrid Bag of Words (BoW) considering frequency of ICD thesaurus in the text. The thesaurus is a list of medical concepts created from the official description of ICD-10 diagnoses by extracting unique terms. Then, exact and partial string matching (i.e., similarity computation) for thesaurus terms is applied to each DS and a BoW vector is created based on terms frequency. Feature selection methods are used to cope with popular data problems by reducing the data dimension with minimal loss of information. In this work, the variance threshold method is used to remove features with variation below a fixed cutoff without considering the relationship of features with the target variable. This cutoff value is manually fixed after several trials (i.e., training and evaluation cycles) and depends on the input data (e.g., in our case, it took a different value for each of the four groups). Artificial Neural networks (ANN) are popular classification methods been widely applied by many researchers to classify textual documents with different types of feature vectors. We adopted a multilayer ANN from DL4J library with backpropagation architecture, one of the most popular neural networks needing several parameters' configuration in the training phase [10,11]. In medical field, backpropagation is widely used to deal with different problem and data type (i.e., image, text, etc.) [12-14]. DL4J proposes Arbiter [15], a framework dedicated to hyperparameter optimization using grid search and random search methods. The use of Arbiter made the hyperparameters optimization an automatable task to guaranty optimal training performances. To evaluate the performance of our models on the test dataset, we used the standard evaluation metrics of accuracy, precision, recall and F-score.

3. Results

A set of experiments were carried out in to evaluate the effectiveness of the proposed approach. Our IA models were trained and evaluated on retrospective French data considering just the principal diagnosis code as label for the whole dataset to address the problem as a multi-class classification task. Table 2 summaries the obtained results for each group. The first indicator is the number of predictable codes which represents the final selection after all training iterations with a F1-Score threshold of 50%. The performances are then reported in terms of accuracy, precision, recall, and F-measure for each group. Average values are all superior to 80% indicating a well prediction effectiveness and a good capacity to adapt to different specialty and different LoS. Table

3 represents the number of predicted codes per group of LoS in four intervals of F1-Score. We can note that 62% of the identified codes have a F1-score above 80% which is very high and encourages further investigation.

Table 2. Evaluation of ICD-10 classification models

Type of Stay	Predictable Codes	Accuracy	Precision	Recall	F1- Score
Group_0	79	89,54%	90,04%	89,03%	89,30%
Group_1_2	74	87,76%	85,37%	83,76%	83,92%
Group_3_4	70	82,84%	83,67%	81,41%	81,70%
Group_5+	123	76,70%	78,99%	76,25%	76,23%
Average		**84,21%**	**84,52%**	**82,61%**	**82,79%**

Table 3. Evaluation of ICD-10 classification models

Type of Stay	F1-Score			
	F1 = 100%	80% <= F1 < 100%	50% <= F1 < 80%	F1 < 50%
Group_0	5	56	18	0
Group_1_2	4	48	22	0
Group_3_4	3	36	31	0
Group_5+	3	61	59	0

We can observe the degressive predictive performance while increasing the LoS which could be explained by the correlation between the LoS and the number of diagnoses. In fact, the number of codes corresponding to short stays is limited. On the contrary, for long stays, the number of possible diagnoses is higher, as well as the number of associated diagnoses.

4. Discussion

Our proposed automated coding models work properly by predicting a principal diagnosis ICD-10 code to each patient's stay according to its LoS. The returned prediction could be a diagnosis code or a null value in the case if the code is not recognized in the list of supported ICD-10 codes by the model, or if the DSs raw text is not enough explicit to distinguish the diseases or the reasons of hospitalization. In these situations, a human intervention is needed to code the stay. The starting step was to limit the study to ICD-10 codes with at least 30 occurrences to reduce the noise and bias. If we look a little more closely to group_0 in Table 1, we retained approximatively 9% of codes (i.e., 145 of 1661) covering more than 82% of cases (i.e., 29005 DSs). The fixed objective was to retain the maximum of ICD codes while ensuring high performances. Thus, an iterative learning approach is adopted to prune the list of retained ICD codes to focus on the well predictable one and ignore the rest. For group_0, this represents 5% of codes and only 34% of the volume of the data set. The prediction of these codes seems to be more interesting for coders as they are more time consuming to manually code.
It is difficult to compare our results with the literature since the problem definition and the finality are not always the same. For instance in terms of F1-Score, [7] obtained 46% for Spanish text classification, [4] reached 74% in hematology unit with 30 diagnosis, [3] attained 88% in radiology with 45 diagnosis, [8] obtained 76% with 6 diagnosis, [9] reached 46% with 500 diagnosis while in our work, we obtained 83% with 346 diagnosis.

5. Conclusion

The application of automated systems for coding diagnosis has recently been renewed using DL methods. In this work we constructed a model based on DL and NLP approaches to automate the ICD-10 coding process from French raw text in discharge summaries. Several ANN were trained corresponding to four categories of stay duration on a real-life dataset. From the tens of thousands of ICD-10 codes, we studied only 613 that are referenced at least 30 times and finally select 346 having the higher predictive effectiveness. The performance and the relevance of our models has been approved by physicians of the MID in a French hospital. The major limitation of our work is the imbalanced and mislabeled data which really depends on the quality and size of available databases. Further works will investigate these limitations and try the use of Convolution Neural Network, considering corpus features as image pixels, and compare it to our previous results. This opens another path of research concerning the use of GPU servers instead of CPU to try to reduce the computation time and eventually accelerate the convergence of our models.

References

[1] https://www.hcup-us.ahrq.gov/db/nation/nis/APR-DRGsV20MethodologyOverviewandBibliography.pdf
[2] https://www.caducee.net/DossierSpecialises/systeme-information-sante/pmsi.asp
[3] Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. In BMC bioinformatics. 2008 April; 9(3):S10
[4] Dermouche M, Velcin J, Flicoteaux R, Chevret S, Taright, N. Supervised topic models for diagnosis code assignment to discharge summaries. In International Conference on Intelligent Text Processing and Computational Linguistics 2016, April: pp. 485-497. Springer, Cham.
[5] Xu K, Lam M, Pang J, Gao X, Band C, Xie P, Xing E. Multimodal Machine Learning for Automated ICD Coding. arXiv preprint arXiv:1810.13348. 2018.
[6] Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad, N. Multi-label classification of patient notes: case study on ICD code assignment. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018, June.
[7] Almagro M, Martinez R, Fresno V, Montalvo S. ICD-10 Coding of Spanish Electronic Discharge Summaries: An Extreme Classification Problem. 2020. IEEE Access.
[8] Sonabend A, et al. Automated ICD coding via unsupervised knowledge integration (UNITE). International Journal of Medical Informatics. 2020;104135.
[9] Zhou L, Cheng C, Ou D, Huang H. Construction of a semi-automatic ICD-10 coding system. BMC Medical Informatics and Decision Making. 2020;20: 1-12.
[10] Hecht-Nielsen R. Theory of the backpropagation neural network. In Neural networks for perception (pp. 65-93). Academic Press. 1992.
[11] Kamath CN, Bukhari SS, Dengel A. Comparative study between traditional machine learning and deep learning approaches for text classification. In Proceedings of the ACM Symposium on Document Engineering 2018 (2018, August) (pp. 1-11).
[12] Al-Milli N. Backpropagation neural network for prediction of heart disease. Journal of theoretical and applied information Technology. 2013;56(1):131-135.
[13] Azmi MSBM, Cob ZC. Breast cancer prediction based on backpropagation algorithm. In 2010 IEEE Student Conference on Research and Development. (2010, December: pp. 164-168. IEEE.
[14] Gupta A, Shreevastava M. Medical diagnosis using back propagation algorithm. Int. J. Emerg. Technol. Adv. Eng. 2011;1(1): 55-58.
[15] https://github.com/deeplearning4j/Arbiter