doi:10.3233/978-1-61499-753-5-201

Automated Diagnosis Coding with Combined Text Representations

Stefan BERNDORFER[a,1] and Aron HENRIKSSON[b]
[a]*Faculty of Computer Science, University of Vienna, Austria*
[b]*Department of Computer and Systems Sciences, Stockholm University, Sweden*

Abstract. Automated diagnosis coding can be provided efficiently by learning predictive models from historical data; however, discriminating between thousands of codes while allowing a variable number of codes to be assigned is extremely difficult. Here, we explore various text representations and classification models for assigning ICD-9 codes to discharge summaries in MIMIC-III. It is shown that the relative effectiveness of the investigated representations depends on the frequency of the diagnosis code under consideration and that the best performance is obtained by combining models built using different representations.

Keywords. Electronic health records, diagnosis coding, predictive modeling

1. Introduction

The digitization of healthcare brought about by the adoption of electronic health record (EHR) systems has made vast amounts of data available for processing by computers [1]. Secondary use of EHR data enables the efficient and often effective provision of clinical decision support at the point of care by building predictive models that learn from large-scale observations of historical data to, e.g., automatically assign or suggest diagnosis codes. Automating the process of diagnosis code assignment can drastically reduce healthcare costs but is challenging due to the inherent difficulty of predicting one or more labels from a large set of classes [2]: ICD-9, for instance, includes around 14,000 unique codes organized in a hierarchical fashion. As a result, many studies have limited the task in some way, e.g. by focusing on a small subset of codes, as in the CMC challenge [3], or on a specific outcome such as mortality [4] or adverse drug events [5]. In reality, the number of distinct combinations of diagnosis codes in EHRs is extremely large and the distribution of codes highly skewed, both of which present challenges for supervised learning approaches [6]. To address these, Perotte et al. [7] proposed a classification strategy that exploits the hierarchy of ICD-9, demonstrating improved performance over a flat prediction model (F_1: 0.29 vs. 0.21), while a similar approach improved performance on adverse drug event detection [8]. In this study, we continue to explore these classification strategies, while focusing on yet another key issue: text representation. In a classification setting, a document is often represented as a bag of words, i.e. using simple (weighted) frequencies. Although such shallow representations often yield competitive performance, deep representations that account for the semantics of words have been proposed, improving performance on various

[1] Corresponding author: stefan.berndorfer@gmx.net

diagnosis coding tasks [9,10]. These representations derive vector representations of words (embeddings) based on their distribution in different contexts: the assumption is that words appearing in similar contexts (i.e. co-occurring with overlapping sets of words) have similar meanings. Here, we show that the relative effectiveness of these representations is related to the frequency of the considered class and that overall performance can be improved by combining shallow and deep text representations.

2. Methods & Materials

We investigate the use of various predictive models for ICD-9 coding of discharge summaries. Two text representations – shallow and deep – are provided to the learning algorithm and their effectiveness, w.r.t. predictive performance, is analyzed for diagnosis codes of varying frequency. Several strategies for combining predictive models that exploit different representations are then explored; the entire analysis is conducted using two classification strategies previously proposed in the literature.

Text Representations: Two popular text representations are used: (1) a shallow representation describing each document as a *bag-of-words* (BoW), i.e. the (weighted) frequency distribution of words in some vocabulary, here defined as the 10,000 words with the highest *Term Frequency - Inverse Document Frequency* (TF-IDF) scores in the training data; (2) a deep representation describing each document as a TF-IDF-weighted sum of semantic vectors that have been learned using the *continuous bag-of-words* (CBOW) model of *Word2Vec* (W2V) [11]. The CBOW model trains a single-layer neural network that learns to predict words based on their contexts, i.e. adjacent words within a symmetric window of a given size; the parameters learned in the hidden layer give us semantic vector representations of words.

Combination Strategies: Once predictive models have been trained using a given representation, they can be combined in an attempt to improve performance. A distinction exists between early fusion and late fusion. In the former, the combination takes place prior to learning, typically by combining feature sets. Here, a combination strategy named *Fusion* is investigated, in which features from the two representations are simply concatenated prior to learning a single predictive model. Various late fusion strategies are also explored. *Select One* chooses a representation and the corresponding model based on the observed best performance for the diagnosis code within a certain frequency interval. Two other strategies are based on simple set operations: *Union* takes the union of the predictions, while *Intersection* takes the intersection of the predictions made by the two models. Finally, *Probability Averaging* takes a weighted average of the class probabilities produced by the models; here, the weights are determined by the observed predictive performance scores for diagnosis codes within a certain frequency interval. A fitted sigmoid was used to obtain probability estimates from the trained Support Vector Machine (SVM) models [12].

Classification Models: Two classification models are used[2] [7]: the *flat SVM model* uses all available training examples, while the *hierarchical SVM model* exploits the ICD- 9 hierarchy[3]. In both settings, the multi-label problem is binarized with a one-versus-all model per diagnosis code. In the flat classification model, documents to which a given diagnosis code has been assigned serve as positive examples and all

[2] Based on PhysioNet project: https://physionet.org/works/ICD9CodingofDischargeSummaries

[3] http://bioportal.bioontology.org/ontologies/ICD9CM

others as negative examples. In the hierarchical model, codes are augmented by their ancestors as follows. Training is carried out from the root downwards: only codes from the parent's sub-tree are considered, where all instances rooted in the code itself serve as positive examples and the remaining ones as negative examples. The prediction follows the same hierarchical procedure: if a parent node has been predicted as negative, no child can be positive, while only leaf nodes serve as final predictions [13].

Experimental Setup: The experiments were conducted using data from the Medical Information Mart for Intensive Care III (MIMIC-III) [14], a publicly available database comprising de-identified health data for over 40,000 critical care patients. All discharge summaries and assigned ICD-9 diagnosis codes were extracted from the database. Codes occurring fewer than 50 times were filtered out, resulting in 59,531 non-empty discharge summaries with at least one assigned diagnosis code. The discharge summaries were tokenized, part-of-speech tagged and lemmatized, while common stopwords[4] were removed. The preprocessed corpus has a vocabulary size of around 125,000, with approximately 44 million instances. The average length of a discharge summary is 742 words (± 435.3). There are 1,301 distinct ICD-9 codes that occur a total of 634,375 times, resulting in an average of 10.66 (± 5.74) codes per summary. The code distribution is strongly skewed towards low-frequent codes: 83% of the codes occur in less then 1% of the discharge summaries. For classification, the LibLinear SVM [15] implementation was used and all representations were L2-normalized. The dataset was divided into a training set (80%), a development set (10%) and an evaluation set (10%). The following parameters were optimized using 5-fold cross-validation on the training set: the window size and dimensionality of the W2V models, as well as the c-value of the linear SVM. To limit the parameter optimization procedure, a sequential approach was taken whereby, first, the window size (5, 10, 25, 50, 100, 150, 200, 250) was optimized using a dimensionality of 200; then, the dimensionality was successively increased[5] (200, 400, 600, 800, 1000); finally, various c-values were explored (2^x, where $x \in \{0, 1, 2, 3, 4, 5\}$). The training set was also used for comparing the BoW and W2V representations: with the hypothesis that the effectiveness of a given representation may depend on the frequency of a diagnosis code, the predictive performance was analyzed in three subsets of the training set, corresponding to tertiles based on code frequency. Models were then trained on the entire training set and, based on the observations of the tertile analysis, the combination strategies were evaluated on the development set. Finally, the best single and combination models were trained on the tuning and development sets and compared on the evaluation set, where McNemar's test [17], with one degree of freedom, was used to verify the statistical significance of the results.

3. Results

The parameter tuning favored a large window size and dimensionality for the W2V spaces and various c-values for the SVM models (Figure 1). The tertile analysis shows that the predictive performance strongly decreases with a lower code frequency (Figure 2). The shallow BoW model performs better on high-frequent codes; however, for

[4] http://www.ranks.nl/stopwords [Accessed October 24, 2016]

[5] Increasing the dimensionality of semantic spaces can lead to improved performance [16].

medium- and low-frequent codes, the deep W2V representation outperforms BoW in both classification models.

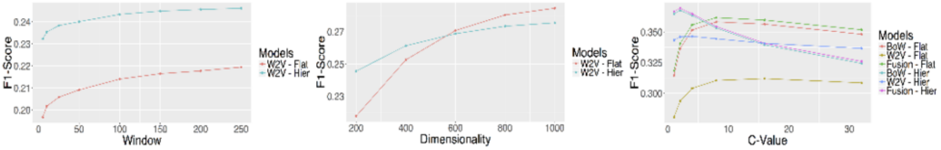

Figure 1. Parameter tuning: W2V window size, W2V dimensionality, SVM c-value.

A comparison of various combination strategies showed that the Union and Fusion models performed best in both classification models (Figure 2). The Union model leads to increased recall at the expense of precision, but outperformed the Fusion model in terms of F_1-score. When comparing the best single model with the best combined model on unseen data, the following was observed. In the flat model, BoW achieved 58.68% precision, 29.96% recall and 36.95% F_1-score. The Union model achieved 55.10% precision, 33.74% recall and 39.16% F_1-score. In the hierarchical model, the BoW representation achieved 43.96% precision, 35.98% recall and 37.97% F_1-score. The Union combination achieved 40.08% precision, 41.69% recall and 39.25% F_1-score. In both settings, the Union prediction model outperformed the BoW representation in terms of F_1-score, by 2.21 points in the flat and 1.28 points in the hierarchical setting. McNemar's test applied independently within each classification model showed that the differences in performance were significant ($p < 0.01$).

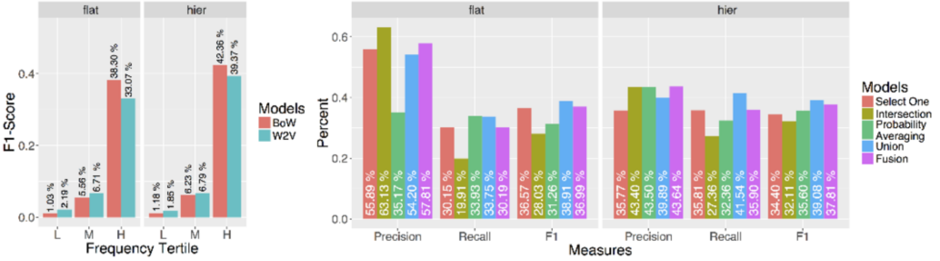

Figure 2. Results: Tertile Analysis and Combination Strategies.

4. Discussion

The tertile analysis revealed that the relative effectiveness of the two investigated text representations depends on the frequency of the diagnosis code under consideration. The deep representation outperformed the shallow counterpart for rare and medium-frequency codes, which can be explained by the lack of training examples to learn from: distributional semantics can then help the classifier to exploit similarities in word meaning between different surface tokens. The complementary nature of the representations was exploited by combining models trained using different representations. A number of combination strategies were evaluated, with the rather naive strategy of taking the union of the predictions outperforming the alternatives. This late fusion strategy hence outperformed the early fusion strategy; in another study on diagnosis code assignment where various late fusion strategies were compared to early fusion, this was not the case [18]. The combined model significantly outperformed the best single model and the results are substantially higher than those presented in [7], in part due to properly tuning the parameters but largely as a result of combining text representations. The performance gain of the Union model can be attributed to the fact that in large clinical datasets, an increase in recall of frequent codes mostly affects performance gains in terms of F_1-score [6].

References

[1] Carol Friedman, Thomas C. Rindflesch, and Milton Corn. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of Biomedical Informatics*, 46(5):765–773, October 2013.

[2] Mary H. Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, November 2010.

[3] John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In: Proc. of BioNLP 2007, p. 97–104. Association for Computational Linguistics, 2007.

[4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. arXiv preprint arXiv:1606.01865, 2016.

[5] Jing Zhao, Aron Henriksson, Lars Asker, and Henrik Boström. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Medical Informatics and Decision Making*, 15(Suppl 4):S1, 2015.

[6] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, 65(2):155–166, October 2015.

[7] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, March 2014.

[8] Jing Zhao, Aron Henriksson, and Henrik Boström. Cascading adverse drug event detection in electronic health records. In: International Conference on Data Science and Advanced Analytics (DSAA), 2015.

[9] Aron Henriksson, Martin Hassel, and Maria Kvist. Diagnosis code assignment support using random indexing of patient records – a qualitative feasibility study. In: Conference on Artificial Intelligence in Medicine, Springer, 2011, p. 348–352.

[10] Aron Henriksson, Jing Zhao, Henrik Boström, and Hercules Dalianis. Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection. In: International Conference on Data Science and Advanced Analytics (DSAA), 2015.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[12] John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers, pages 61–74. MIT Press, 1999.

[13] Yitao Zhang. A Hierarchical Approach to Encoding Medical Concepts for Clinical Notes. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, HLT-SRWS ’08, pages 67–72, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[14] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[15] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874, 2008.

[16] Aron Henriksson and Martin Hassel. Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In: Proceedings of Louhi Workshop on Health Document Text Mining and Information Analysis, 2013.

[17] Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. 10(7):1895–1923, October 1998.

[18] Aron Henriksson, Jing Zhao, Henrik Boström, and Hercules Dalianis. Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), p. 343–350, 2015.