Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200173

Machine Learning for Automatic Encoding of French Electronic Medical Records: Is More Data Better ?

Julien GOBEILL[a,b,1] , Patrick RUCH[a,b], and Rodolphe MEYER[c]
[a] *SIB Text Mining group, Swiss Institute of Bioinformatics, Geneva, Switzerland*
[b] *HES-SO / HEG, Information Sciences, Geneva, Switzerland*
[c] *Information Systems Department, University Hospitals of Geneva (HUG), Geneva, Switzerland*

Abstract. The encoding of Electronic Medical Records is a complex and time-consuming task. We report on a machine learning model for proposing diagnoses and procedures codes, from a large realistic dataset of 245 000 electronic medical records at the University Hospitals of Geneva. Our study particularly focuses on the impact of training data quantity on the model's performances. We show that the performances of the models do not increase while encoded instances from previous years are exploited for learning data. Furthermore, supervised models are shown to be highly perishable: we show a potential drop in performances of around -10% per year. Consequently, great and constant care must be exercised for designing and updating the content of such knowledge bases exploited by machine learning.

Keywords. Medical coding, machine learning, text mining.

1. Introduction

The coding, or encoding, of medical records with standard classifications serves multiple purposes for healthcare actors, including billing and reimbursement, quality control, epidemiological surveillance, or cohort identification for clinical trials [1,2]. This task is usually performed by professional coders, at the term of an episode of care (EOC). These coders have to master knowledge in the field of medicine, along with the handling of large code sets, coding rules and local guidelines [3]. For these reasons, manual coding is expensive and time-consuming. The amount of available electronic medical records (EMRs), potentially exploitable by computers, has steadily grown in recent years [4]. Hence, machine learning algorithms are today likely to produce accurate and effective tools for assisting human coders [5].

We report on the development and evaluation of a machine learning – or supervised – model for suggesting codes, from a large dataset of EMRs.

A citation attributed to Peter Norvig, Google's Chief Scientist, claims that "more data beats clever algorithms" [6]. Several reported works on text categorization tend to support this idea, and incite to direct efforts towards increasing the size of annotated

[1] Corresponding Author, SIB Swiss Institute of Bioinformatics, quartier Sorge - bâtiment Amphipôle, 1015 Lausanne, Switzerland; E-mail: julien.gobeill@sib.swiss.

learning collections [7,8]. Our study has particularly focused on the impact of exploiting growing amounts of learning data on the algorithm's performances. The figure below illustrates the global workflow of the automatic encoding process.

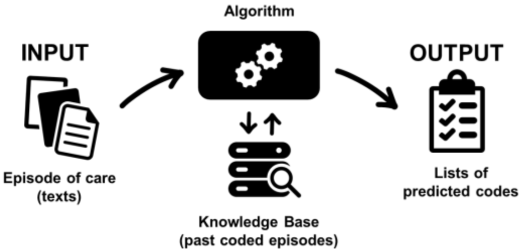

Figure 1. Global workflow of the machine learning encoding. The input is all textual documents from an EOC. The algorithm compares it with past coded episodes contained in the Knowledge Base, and retrieves the k most similar instances, based on statistical textual similarity. From the encoding of these similar episodes, the algorithm infers rankings of candidate CIM and CHOP codes, along with probability scores.

Our study was conducted in the University Hospitals of Geneva (HUG), with a real dataset ranging from 2011 to early 2016 (245 000 EOCs). In Switzerland, EOCs are encoded with two medical classifications derived from the International Classification of Diseases (ICD). Diagnoses are encoded with the CIM classification, issued from ICD-10, German modification; procedures are encoded with the CHOP classification, issued from the ICD-9-CM in 2008 but having evolved since. In Switzerland, both classifications are available in German, French and Italian, but all medical records in Geneva are written in French.

This encoding task is a multi-label automatic text categorization problem [9]. Most reported systems in the literature rely on the popular Support-Vector Machines (SVM). Yet, for large multi-label categorization, SVM has been reported to scale with difficulties [10], requiring the usage of data reduction or feature selection [1]. In our study, with our objectives, we chose to exploit a simple, yet powerful classification algorithm: the k-Nearest Neighbors (k-NN). From a Knowledge Base (KB) containing past encoded instances, this algorithm retrieves the instances that are the most similar to a new instance, and assigns to it the most common codes among these neighbors.

2. Data

At the HUG, encoding is performed by professional coders after an EOC. Coders have limited contacts with the healthcare personnel, but they review all produced documents. There are 4 318 different types of documents, including admission notes, exam or operative reports, progress notes, or discharge letters. Coders have to detect disorders that were addressed during the EOC, and to encode them into CIM codes. For example, for a patient admitted for an alcoholic cirrhosis of liver, the episode will be encoded with the CIM code K70.3 “Cirrhose alcoolique du foie”. The coders also have to detect procedures that were performed. In our example, if a liver biopsy was performed, the coder has to encode it with the CHOP code 50.13.10 “Biopsie du foie transveineuse”.
Thus, the whole dataset consists in 245 000 encoded EOCs, from 2011 to early 2016. It is worth noting that what is encoded is EOCs, not documents. In this perspective, all experiments were conducted at the EOC level. For each experiment, a test set of

randomly selected 250 EOCs was submitted to the tool. The algorithm is judged on its capacity to output the codes that were effectively assigned to these episodes by professional coders.

3. Methods

The investigated algorithm is a k-Nearest Neighbors (k-NN). This algorithm performs instance-based learning; it does not learn any discriminative function, but rather exploits a Knowledge Base (KB), containing previously labelled instances, in order to infer real-time decisions [10]. For such text categorization problems with numerous and multiple labels, k-NN has been reported to efficiently scale and to perform comparably with Support Vector Machines [11].

In this study, the KB contains previously encoded EOCs. A similar search engine is designed for retrieving nearest neighbors (i.e. EOCs), based on textual similarity. The Terrier information retrieval platform was used for this purpose [12]. The k parameter sets the number of neighbors exploited for making decisions. Once similar EOCs are retrieved, the second step of the k-NN is to infer encoding for the input episode. For this purpose, the algorithm simply computes what are predominant codes among the retrieved learning instances. The algorithm hence outputs ranked lists of disorders (CIM) and procedures (CHOP) codes, along with confidence scores.

The algorithm is judged on its capacity to output the codes that were effectively assigned to these episodes by professional coders. The evaluation relies on metrics issued from information retrieval [13]. As the tool potentially outputs hundreds of codes – along with decreasing scores – evaluated lists are limited to the top 10 predicted codes. In a semi-automatic workflow, ten is a convenient number of codes easily checkable by the coder. Hence, the recall at rank 10 curves presented in the Results section is the proportion of actual codes that were outputted by the tool in the top 10 ranks. Precision curves are not presented for clarity concerns, but show similar shapes.

4. Results

Preliminary studies (not reported) were conducted with a dataset limited to 2014 EOCs. For both disorders and procedures, the performances of the tool increased with the size of the KB. The most promising observation is that more data leads to better results: for 2014 EOCs, when the size of the KB doubles, the recall is improved by approximately 10%. Thus, with the complete dataset (2011 - 2016), higher performances were expected.

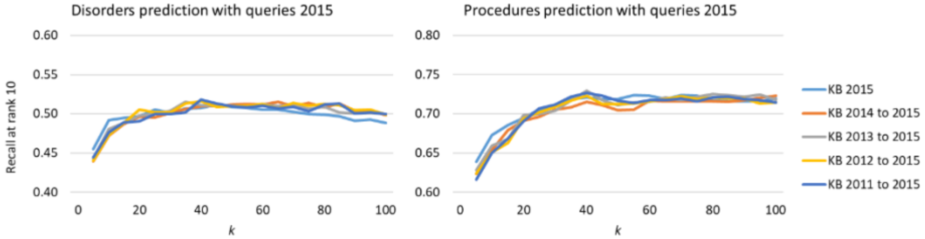

Figure 2. Performances (Recall at rank 10) of the encoding with 2015 queries, for disorders (CIM codes) and procedures (CHOP codes), depending on the k parameter. The KB is populated by previously encoded EOCs, from 52 000 for 2015 only, to 230 000 for the 2011 to 2015 range.

A first set of experiments was conducted for assessing the impact of training data quantity on the model's performances (Figure 2). The test set was a sample of EOCs from 2015, while we progressively increased the number of past EOCs for populating the KB. For both disorders and procedures, all curves are interwoven: the performances of the model do not increase while encoded EOCs from previous years are injected in the KB – yet the past EOCs do not damage the model.

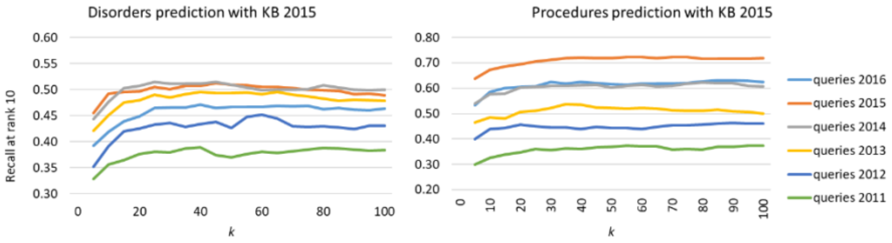

Figure 3. Performances (Recall at rank 10) of the encoding with 2015 KB, depending on the k parameter. Different test sets, containing EOCs sampled in a single year ranging from 2011 to 2016, were used.

A second set of experiments focused on the durability of the model (Figure 3). This time, the KB did not vary and only contained 50 000 encoded episodes from 2015. The model was evaluated with different test sets, each sampled in a single year ranging from 2011 to 2016. The different curves are this time remarkably distinct. With a KB containing previously encoded EOCs from 2015, the best results are reached for input EOCs from 2014 and 2015. Older episodes, from 2013 to 2011, obtain progressively weaker results. These results show that the encoding is considerably year-dependent, and suggest that a model built on a single year is highly perishable – with a potential drop in performances of around -10% per year.

Table 1. Final performances reached for KB and queries from 2015. F-measure is the harmonic mean of precision and recall. This study focused on optimizing recall values.

Coding	F1 at rank 10	P at rank 10	R at rank 10	R at rank 20
Diagnoses	39 %	32 %	51 %	62 %
Procedures	29 %	18 %	72 %	78 %

5. Discussion

The F-measure at rank 10 reached by the tool for 2015 data is 39.3% (see Table 1). This is competitive with F-measures (39.5% [14], 42% [5]) reported in other studies dealing with such large-scale data. Beyond state-of-the-art comparisons, the recall values – 62% for diagnoses and 78% for procedures at rank 20 – must be considered in the light of the inter-annotator agreement, which is a theoretical upper bound. In 2006, this agreement was reported about 79% in the University Hospitals of Geneva [15].

However, the results indicate that the performances of the machine learning model are considerably year-dependent. This could be explained by two main reasons. First, the classifications used for medical encoding in Switzerland evolve over time: every two years for diagnoses, every year for procedures. Codes can be created, merged, or migrated. Second, the coding guidelines proper to each hospital also evolve over time.

Indeed, on the top of diagnoses and procedures codes, rules defining reimbursement are applied by Swiss authorities based on DRGs (Diagnosis Related Groups), unknown from facilities. As two different accurate encodings of the same EOC can lead to differences in the reimbursement, each facility aims at continuously improving its proper coding guidelines, in order to better encode provided healthcare. The consequence is a typical case of concept drift in medical encoding.

6. Conclusion

Are more data better for automatic encoding of EMRs by machine learning? In University Hospitals of Geneva, the answer is mixed. For a unique year, injecting more data in the knowledge base leads to progressively better performances for the algorithm. Yet, injecting encoded episodes from past years does not improve – while does not harm – the performances. Another essential result is that KBs are considerably year-dependent, and a model built on previous years' EOCs appear to be perishable. In this study, a potential drop in performances of around -10% per year has been brought to light. Consequently, great and constant care must be exercised for designing and updating the content of such knowledge bases exploited by machine learning. As Peter Norvig has refined: "More date beats clever algorithms, but better data beats more data".

References

[1] R. Kavuluru, A. Rios, Y. Lu, An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records, AI in medicine, 65(2) (2015), 155-166.

[2] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, Measuring diagnoses: ICD code accuracy, Health services research, 40(5) (2005), 1620-1639.

[3] Y. Chen, H. Lu, L. Li, Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity, PloS one, 12(3) (2017), e0173410.

[4] EMC with Research & Analysis by IDC, The Digital Universe Driving Data Growth in Healthcare, available from: https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf (2014)

[5] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9 coding via deep transfer learning, Neurocomputing, 324 (2019), 43-50.

[6] A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data. Intelligent Systems, IEEE, 24(2) (2009), 8-12.

[7] M. Banko, E. Brill. Scaling to very very large corpora for natural language disambiguation. In : Proceedings of the 39th annual meeting on association for computational linguistics, (2001), 26-33.

[8] J. Gobeill, E. Pasche, D. Vishnyakova, P. Ruch, Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases, Database, (2013) bat041.

[9] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR), 34(1) (2002), 1-47.

[10] Z. Yao, W. Ruzzo, A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data, BMC bioinformatics, 7(1) (2006), 1-11.

[11] Y. Yang, X. Liu, A re-examination of text categorization methods. Proceedings of the 22nd annual international SIGIR conference on Research and development in information retrieval, (1999), 42-49.

[12] I. Ounis, G. Amati, V. Plachouras, et al, Terrier: A High Performance and Scalable Information Retrieval Platform, Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval, (2006).

[13] C. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Natural Language Engineering, 16(1) (2010), 100-103.

[14] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, Journal of the AMIA, 21(2) (2013), 231-237.

[15] P. Ruch, J. Gobeill, I. Tbahriti, A. Geissbühler, From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. AMIA Annual Symposium Proceedings, (2008), 636-640.