Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Inter-labeler and intra-labeler variability of condition severity classification models using active and passive learning methods

Nir Nissim[a,b,*], Yuval Shahar[a], Yuval Elovici[a,b], George Hripcsak[c,d], Robert Moskovitch[a,c]

[a] Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[b] Malware Lab, Cyber Security Research Center, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[c] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[d] Observational Health Data Sciences and Informatics, Columbia University, New York, NY, USA

ARTICLE INFO

Article history:
Received 1 March 2017
Accepted 3 March 2017

Keywords:
Active learning
Electronic health records
Phenotyping
Condition
Severity
Variance
Labeling

ABSTRACT

Background and objectives: Labeling instances by domain experts for classification is often time consuming and expensive. To reduce such labeling efforts, we had proposed the application of active learning (AL) methods, introduced our CAESAR-ALE framework for classifying the severity of clinical conditions, and shown its significant reduction of labeling efforts. The use of any of three AL methods (one well known [SVM-Margin], and two that we introduced [Exploitation and Combination_XA]) significantly reduced (by 48% to 64%) condition labeling efforts, compared to standard passive (random instance-selection) SVM learning. Furthermore, our new AL methods achieved maximal accuracy using 12% fewer labeled cases than the SVM-Margin AL method.

However, because labelers have varying levels of expertise, a major issue associated with learning methods, and AL methods in particular, is how to best to use the labeling provided by a committee of labelers. First, we wanted to know, based on the labelers' learning curves, whether using AL methods (versus standard passive learning methods) has an effect on the *Intra*-labeler variability (*within* the learning curve of each labeler) and *inter*-labeler variability (*among* the learning curves of different labelers). Then, we wanted to examine the effect of learning (either passively or actively) from the labels created by the majority consensus of a group of labelers.

Methods: We used our CAESAR-ALE framework for classifying the severity of clinical conditions, the three AL methods and the passive learning method, as mentioned above, to induce the classifications models. We used a dataset of 516 clinical conditions and their severity labeling, represented by features aggregated from the medical records of 1.9 million patients treated at Columbia University Medical Center. We analyzed the variance of the classification performance within (*intra*-labeler), and especially among (*inter*-labeler) the classification models that were induced by using the labels provided by seven labelers. We also compared the performance of the passive and active learning models when using the consensus label.

Results: The AL methods: produced, for the models induced from each labeler, smoother *Intra*-labeler learning curves during the training phase, compared to the models produced when using the passive learning method. The mean standard deviation of the learning curves of the three AL methods over all labelers (mean: 0.0379; range: [0.0182 to 0.0496]), was significantly lower ($p = 0.049$) than the *Intra*-labeler standard deviation when using the passive learning method (mean: 0.0484; range: [0.0275–0.0724).

Using the AL methods resulted in a lower mean *Inter*-labeler AUC standard deviation among the AUC values of the labelers' different models during the training phase, compared to the variance of the

Abbreviations: CAESAR, Classification Approach for Extracting Severity Automatically from Electronic Health Records; CAESAR-ALE, Classification Approach for Extracting Severity Automatically from Electronic Health Records– Active Learning Enhancement; EHR, Electronic Health Record; AL, Active Learning; SVM, Support Vector Machines; VS, Version Space; SNOMED-CT, Systemized Nomenclature of Medicine-Clinical Terms; ICD-9, International Classification of Diseases – Version 9; SVM-Margin, Support Vector Machines-Margin Method – An existing AL method oriented towards acquiring informative conditions that lie closest to the separating hyperplane (inside the margin); Exploitation, An AL method included in the CAESAR-ALE framework that is oriented towards acquisition of severe conditions; Combination_XA, An AL method included in the CAESAR-ALE framework that combines elements of the Exploitation method and the SVM-Margin method, so that it applies a hybrid acquisition strategy for enhanced improvement of the CAESER method.

* Corresponding authors at: Ben-Gurion University of the Negev, P.O.B 653, Beer-Sheva, 84105, Israel.
E-mail address: nirni@post.bgu.ac.il (N. Nissim).

http://dx.doi.org/10.1016/j.artmed.2017.03.003

induced models' AUC values when using passive learning. The *Inter*-labeler AUC standard deviation, using the passive learning method (0.039), was almost twice as high as the *Inter*-labeler standard deviation using our two new AL methods (0.02 and 0.019, respectively). The SVM-Margin AL method resulted in an *Inter*-labeler standard deviation (0.029) that was higher by almost 50% than that of our two AL methods The difference in the *inter*-labeler standard deviation between the passive learning method and the SVM-Margin learning method was significant ($p = 0.042$). The difference between the SVM-Margin and Exploitation method was insignificant ($p = 0.29$), as was the difference between the Combination_XA and Exploitation methods ($p = 0.67$).
Finally, using the consensus label led to a learning curve that had a higher mean intra-labeler variance, but resulted eventually in an AUC that was at least as high as the AUC achieved using the gold standard label and that was always higher than the expected mean AUC of a randomly selected labeler, regardless of the choice of learning method (including a passive learning method). Using a paired *t*-test, the difference between the *intra*-labeler AUC standard deviation when using the consensus label, versus that value when using the other two labeling strategies, was significant only when using the passive learning method ($p = 0.014$), but not when using any of the three AL methods.
Conclusions: The use of AL methods, (a) reduces *intra-labeler* variability in the performance of the induced models during the training phase, and thus reduces the risk of halting the process at a local minimum that is significantly different in performance from the rest of the learned models; and (b) reduces *Inter*-labeler performance variance, and thus reduces the dependence on the use of a particular labeler. In addition, the use of a consensus label, agreed upon by a rather uneven group of labelers, might be at least as good as using the gold standard labeler, who might not be available, and certainly better than randomly selecting one of the group's individual labelers. Finally, using the AL methods: when provided by the consensus label reduced the intra-labeler AUC variance during the learning phase, compared to using passive learning.

1. Introduction

Active learning (AL), a form of machine learning in which the learning method actively requires labels for specific instances in which knowing the label seems most beneficial to the learning process, has been at the focus of a substantial amount of research over the last decades. AL has been shown to be successful in decreasing the amount of labeling requirements, compared to a traditional passive learning method, in many domains including the cyber security [25–27,37–41,68–71] and biomedical domains [30–33,53–54]. While labeling and learning with an active learner is often much more efficient and achieves higher classification accuracy with a smaller labeled training set, the learning curve may vary greatly according to the labeler's expertise in the domain. The clinical domain is an excellent example of a domain in which there is a large number of potential experts with varying levels of expertise, depending on their training and experience. However, physicians, and particularly experts, are often very busy, and their time is expensive [43]. Thus, the focus of our current study is to examine the use of labelers with varying levels of clinical training and experience.

We have previously examined the effect of various learning methods on the specific task of determining the *severity level* of medical conditions. The severity level is an important aspect of each medical condition, which is expected to be useful for discriminating between sets of conditions or phenotypes. For the purposes of our research, we define severe conditions as those that are life threatening or permanently disabling. Such conditions would be considered as high priority in terms of the need to generate phenotype definitions for tasks such as pharmacovigilance [44,45,47]. Condition level severity classification can distinguish acne (mild condition) from myocardial infarction (severe condition). The bulk of the literature focuses on *patient level* severity, which generally requires individual condition metrics [8–11], although whole-body methods exist [11–13].

Severity level is also useful for prioritizing conditions that are important for specialized phenotyping algorithms. Although several consortiums and partnerships, including the Observational Medical Outcomes Partnership [1] and the Electronic Medical Records and Genomics Network [2,3], have developed methods for extracting conditions and their related characteristics from Electronic Health Records (EHRs), only a little more than 100 conditions/phenotypes have been successfully defined. Unfortunately, this represents just a small fraction of the approximately 401,200[1] conditions recorded in EHRs. Hurdles faced by experts when defining phenotype-extraction algorithms include overcoming definition discrepancies [4], data sparseness, data quality [5], bias [6], and healthcare process effects [7]. Condition severity can be one way of identifying conditions worthy of developing a specialized phenotype-extraction algorithm.

In our previous work, we developed an algorithm that we refer to as Classification Approach for Extracting Severity Automatically from Electronic Health Records (CAESAR) [13,47], which uses standard machine learning (also referred to as *passive learning*) to classify condition severity based on metrics extracted from EHRs [13] and requires medical experts to manually review and assign a severity status to each condition (i.e., severe or mild) independently from EHR metrics. We have recently developed and assessed an *Active Learning Enhancement* version of *CAESAR*, called *CAESAR-ALE*, which was initially published as a preliminary study [49], and was then extended into a more detailed paper [76]. Using three different AL methods, including two new AL methods that we developed, we demonstrated that the labeling burden on medical experts can be significantly reduced. All three AL methods decreased the labelers' efforts, compared to the passive learning methods applied by the original CAESER framework in which the classifier was trained on the entire set of conditions; depending on the AL strategy used in that study [13], the reduction ranged from 48% to 64%, which can result in significant savings, both in time and money.

Several labelers participated in our original study, and a separate learning curve was created for each labeler, depicting the classification model induced by using the labels provided by each labeler. The variance between the learning curves observed might be a result of the varying levels of clinical training and experience

[1] The number of SNOMED-CT codes as of September 9, 2014. Accessed via: http://bioportal.bioontology.org/ontologies/SNOMEDCT

Please cite this article in press as: Nissim N, et al. Inter-labeler and intra-labeler variability of condition severity classification models using active and passive learning methods. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.03.003