Unsupervised patient representations from clinical notes with interpretable classification decisions

Madhumita Sushil[1,2,3], Simon Šuster[2,4], Kim Luyckx[1,5], Walter Daelemans[2,4]
[1]Antwerp University Hospital, Belgium
[2]CLiPS Research Center, University of Antwerp, Belgium
[3]firstname.lastname@outlook.com
[4]firstname.lastname@uantwerpen.be
[5]firstname.lastname@uza.be

Abstract

We have two main contributions in this work: 1. We explore the usage of a stacked denoising autoencoder, and a paragraph vector model to learn task-independent dense patient representations directly from clinical notes. We evaluate these representations by using them as features in multiple supervised setups, and compare their performance with those of sparse representations. 2. To understand and interpret the representations, we explore the best encoded features within the patient representations obtained from the autoencoder model. Further, we calculate the significance of the input features of the trained classifiers when we use these pretrained representations as input.

1 Introduction

Representation learning techniques have been used extensively within and outside the clinical domain to learn the semantics of words, phrases, and documents (Baroni et al., 2014; Liu et al., 2016). We apply such representation learning techniques to create a patient semantic space by learning vector representations at the patient level. In a patient semantic space, "similar" patients should have similar vectors. Patient similarity metrics are widely used in several applications to assist clinical staff. Some examples are finding similar patients for rare diseases (Garcelon et al., 2017), identification of patient cohorts for disease subgroups (Li et al., 2015), providing personalized treatments (Zhang et al., 2014; Wang et al., 2015), and predictive modeling tasks such as patient prognosis (Gottlieb et al., 2013; Wang et al., 2012) and risk factor identification (Ng et al., 2015). When patient similarity is calculated as an ontology-guided distance between specific structured properties of patients such as diseases and treatments, it represents patient relationships corresponding to those properties. However, when the similarity measure is fuzzy, the different properties that influence the similarity value are unknown. We aim to capture patient similarity on multiple dimensions, such as complaints, diagnoses, procedures performed, etc., which would encapsulate a holistic view of the patients.

In this work, we create dense patient representations that are transferable across tasks from clinical notes in the freely available MIMIC-III database (Johnson et al., 2016). We focus on different techniques for creating patient representations using only textual data. We explore the usage of two neural representation learning architectures—a stacked denoising autoencoder (Vincent et al., 2010), and a paragraph vector architecture (Le and Mikolov, 2014)—for unsupervised learning. We evaluate the quality of the learned representations through multiple supervised tasks.

Dense representations can capture semantics, but at a loss of interpretability. We take a step towards bridging this gap by proposing different techniques to interpret the information encoded in the patient vectors obtained during the unsupervised learning phase, and to extract the features that most influence the classification output when they are used as input.

Workshop on Machine Learning for Health (NIPS 2017), Long Beach, CA, USA.

2 Learning Patient Representations

Stacked denoising autoencoder: Miotto et al. (2016) used a stacked denoising autoencoder (SDAE) (Vincent et al., 2010) to learn patient representations for disease prognosis using structured patient data combined with probabilistic topic models obtained from unstructured data. Suresh et al. (2016) used a sequence-to-sequence autoencoder to generate patient phenotypes using structured data. Given the success of these models, we explore the use of an SDAE for task-independent patient representation from unstructured data forgoing the use of intermediate techniques like topic modeling.

We sequentially train every layer of an SDAE as an independent denoising autoencoder to reconstruct the hidden layer output of the previous autoencoder from a corresponding corrupted version. We use the hidden layer value of the final autoencoder as the dense patient representations R. We use the sigmoid activation function for the encoders, and the linear activation function for the decoders. We train the network to minimize the mean squared reconstruction error using the RMSProp optimizer (Tieleman and Hinton, 2012). After a randomized hyperparameter search (Bergstra and Bengio, 2012), we obtain a 1-layer SDAE with 800 hidden units and 5% dropout noise.

Paragraph vector: Doc2vec, or 'Paragraph Vector' (Le and Mikolov, 2014), learns dense fixed-length representations of variable length texts such as paragraphs and documents. It supports two algorithms—a distributed bag-of-words (DBOW) algorithm, and a distributed memory (DM) algorithm. We use the DBOW algorithm for 5 iterations, with a window size of 3, a minimum frequency threshold of 10, and 5 negative samples per positive sample to train 300 dimensional patient vectors. We determined these settings also using the randomized hyperparameter search.

3 Feature extraction

When statistical models are deployed for clinical decision support, it is crucial to understand the features that influence the model output. A ranked list of the most influential features can assist such understanding, while facilitating error analysis, and exploratory analysis when unexpected features are ranked high. We propose two techniques to achieve model interpretability. First, to estimate how well the individual features are encoded in the patient vectors learned through the SDAE, we calculate the **squared reconstruction error** of the input features in the first layer of the pretrained autoencoder, averaged across all the training instances. Next, we extend the work by Engelbrecht and Cloete (1998) and use **sensitivity analysis** to calculate the significance of the original input words for different classification tasks for a selected set of instances, when the task-independent dense patient representations R are first generated using the SDAE, and R is then used as the input to the classifiers. This technique is transferable to the doc2vec representations and we plan to extend it in future. Given an input R to the classifier corresponding to the original inputs z to the SDAE model, the significance of the ith input feature, ϕ_{z_i} is defined as the maximum significance of the input feature i across all the K output units (o) of the classifier with respect to the N instances:

$$\phi_{z_i} = \max_{k=1\ldots K}\{S_{oz,ki}\}, \text{ where } S_{oz,ki} = \sqrt{\sum_{j=1}^{N}[S_{oz,ki}^{(j)}]^2 * N^{-1}}.$$

$S_{oz,ki}^{(j)}$ is the sensitivity of the kth output unit of the classifier w.r.t the ith input feature of the SDAE for an instance j:

$$S_{oz,ki}^{(j)} = \frac{\partial o_k^{(j)}}{\partial z_i^{(j)}} = \frac{\partial o_k^{(j)}}{\partial R_i^{(j)}} * \frac{\partial R_i^{(j)}}{\partial z_i^{(j)}}.$$

4 Dataset construction and preprocessing

We retrieve a set of adult patients ($\geq$18 years) with one hospital admission and at least one associated note (excluding discharge reports) from the MIMIC-III database (Johnson et al., 2016). We split it into a set of 24,650 patients for training, and 3,081 patients each for validation and testing. We represent patients with a concatenation of all their non-discharge notes. We tokenize the data using

Table 1: Classification results on different tasks using the BoW features, the SDAE and the doc2vec patient representations, and on concatenating the two dense representations (with κ score).

Approach	In_hosp	30_days	1_year	Pri_diag_cat	Pri_proc_cat	Gender
BoW	94.57	59.49	79.42	70.16	73.66	98.47
SDAE	91.94	79.65	79.80	65.00	67.46	87.75
doc2vec	91.95	76.80	81.34	68.07	65.83	97.70
(κ) SDAE + doc2vec	(58.65) 93.83	(00.00) 81.13	(15.81) 83.02	(64.38) 67.88	(58.91) 70.30	(72.00) 97.47

the Ucto tokenizer (Van Gompel et al., 2012) and lowercase it. To obtain patient representations using the SDAE, we replace the numbers, and certain time and measurement mentions with special tokens. We remove the punctuations, and the terms with frequency < 5. We use a bag-of-words (BoW) with their TF-IDF scores as features, to obtain a vocabulary size of 71,001. We also conducted the experiments with a bag-of-medical-concepts feature set, but they performed consistently worse. To train the doc2vec models, we remove the numbers, and the tokens matching time and measurement patterns (determined from the initial validation set results), and get a vocabulary size of 48,950.

5 Evaluation

5.1 Task description

We use the dense patient representations R as the input features to train feedforward neural network classifiers for the following independent tasks: binary prediction of patient mortality during the hospital stay (13.14%), within 30 days of discharge (3.85%), or within 1 year of discharge (12.19%); prediction of the 20 generic diagnostic categories, and the 18 generic procedural categories as encoded in the ICD-9-CM database (World Health Organization, 2004), corresponding to the most relevant diagnostic and procedural codes for a patient (the majority classes are 40.2% and 38.9% respectively); and gender prediction—male (56.87%) or female (43.13%). We evaluate the models using the area under the ROC curve for patient death for the mortality tasks, and the weighted F-score for the others, to correct for class imbalance. We minimize the categorical cross-entropy error using the RMSProp optimizer, and determine the hyperparameters using randomized search.

5.2 Results and Discussion

In Table 1, we compare the classification performance on using the dense patient representations obtained from the SDAE and the doc2vec models as the input features for all the tasks, compared to the BoW sparse features. We analyze the agreement between the SDAE and the doc2vec model outputs by calculating Cohen's κ score (Cohen, 1960) between them on the validation set. We find that the agreement scores are not high, which may indicate that the models learn complimentary information. We then concatenate the two dense representations to analyze model complementarity.

Our main finding is that all the dense representation techniques significantly outperform[1] the baseline for 30 days mortality prediction. However, although we see a large numerical improvement over the baseline on using the dense representations for 1 year mortality prediction, the differences are not statistically significant. We believe that the poor performance of the BoW model for 30 days mortality prediction may be due to the low number of positive instances, and that generalization assists feature identification in such cases. Grnarova et al. (2016) have previously shown significant improvements for these tasks on using a 2-level convolutional neural network as compared to the doc2vec vectors used in linear support vector machines. However, our results are not directly comparable because we use different data subsets, and non-linear neural classifiers with the doc2vec representations. The sparse inputs perform better than the SDAE representations for all the other tasks, and better than the doc2vec representations for in-hospital mortality and primary procedural category prediction. One probable reason is that the best predictors for the other tasks are the direct lexical mentions in the notes, which makes the BoW model a very strong baseline. Examples of such features obtained using the χ^2 feature analysis are 'autopsy', 'expired', 'funeral', and 'unresponsive' for in-hospital mortality prediction, and 'himself', 'herself', 'ovarian', and 'testicular' for gender prediction.

[1] All the statistical significance scores were calculated using the two-tailed pairwise approximate randomization test (Noreen, 1989) with a significance level of 0.05 before the Bonferroni correction for 36 hypotheses.

Table 2: The most significant features for the classifiers for one test instance each when the SDAE representations are used as the input. The true classes are 'patient death' for the mortality tasks (a common instance for 30 days and 1 year mortality prediction), and 'diseases of the digestive system', 'operations on the digestive system', and 'male' respectively for a common patient for the other tasks.

In_hosp	30_days	1_year	Pri_diag_cat	Pri_proc_cat	Gender
vasopressin	leaflet	magnevist	numeric_val	numeric_val	woman
pressors	structurally	signal	previous	no	female
focused	pacemaker	decisions	rhythm	of	she
dnr	sda	periventricular	no	enzymes	man
dopamine	periventricular	embolus	flexure	extubated	he
acidosis	excursion	underestimated	dementia	rhythm	male
levophed	non-coronary	calcified	brbpr	and	her
pressor	dosages	screws	of	the	his
cvvhd	microvascular	rib	sinus	vent	wife
cvvh	left-sided	shadowing	for	uncal	uterus
emergency	chronic	gadolinium	to	mso	him
pneumatosis	extubation	mri	tracing	to	urinal

The concatenation of the vectors learned by both models is not statistically different from the sparse representations under the given significance level for any task except 30 days mortality prediction, where the concatenation is better. This ensemble model significantly outperforms both individual models for primary procedural category prediction. For primary diagnostic category and gender prediction, the ensemble model is significantly better than the SDAE model, but not the doc2vec model. In these cases, there is no significant difference between the doc2vec and the BoW models. Hence, we observe that the concatenation helps in some cases and we recommend combining the two dense representations for unknown tasks. The doc2vec model uses a local context window in a log-linear classifier, whereas the SDAE model uses only the global context information and non-linear encoding layers. This may be one of the factors governing the differences between the two techniques.

Furthermore, we rank the features according to their mean squared reconstruction error when we pretrain the patient representations using the SDAE. We observe that infrequent terms such as spelling errors are reconstructed very well, as opposed to the frequent features in the dataset. To check for a correlation between this error and the feature frequency, we calculate the Spearman's rank correlation coefficient (Kokoska and Zwillinger, 2000) between the two parameters, and obtain a value of 0.8738. We believe that this behavior may be due to the high entropy of the frequent terms. Jo et al. (2017) also obtain misspellings and rare words as the top features when they use recurrent neural networks for patient mortality prediction in the MIMIC-III dataset.

In Table 2, we list the most significant features for the model outputs for one instance each in the test set. We find that the classifiers give high importance to sensible frequent features for most of the tasks, although the SDAE reconstructs the low frequent terms better during the pretraining phase. There is a minimal overlap between the sets of important features for different tasks. This shows that R is task-independent, and that the classifiers can identify task-specific important information when they are trained for a specific task.

6 Conclusions

We have shown that the dense patient representations significantly improve the classification performance for 30 days mortality prediction, a task where we are confronted with a very low proportion of positive instances. For the other tasks, this advantage is not visible. Moreover, we have shown that a combination of the stacked denoising autoencoder and the doc2vec representations improves over the individual models for some tasks, without any harm for the others tasks. Furthermore, during feature analysis, we have found that the frequent terms are not encoded well during the pretraining phase of the stacked denoising autoencoder. However, when we use these pretrained vectors as the input, sensible frequent features are selected as the most significant features for the classification tasks.

Acknowledgments

This research was carried out within the Accumulate SBO project (`www.accumulate.be`), funded by the government agency Flanders Innovation & Entrepreneurship (VLAIO), Belgium [grant number 150056].

References

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 238–247.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Engelbrecht, A. and Cloete, I. (1998). Feature extraction from feedforward neural networks using sensitivity analysis. In *Proceedings of the International Conference on Systems, Signals, Control, Computers*, pages 221–225.

Garcelon, N., Neuraz, A., Benoit, V., Salomon, R., Kracker, S., Suarez, F., Bahi-Buisson, N., Hadj-Rabia, S., Fischer, A., Munnich, A., et al. (2017). Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. warehouse and the needle in the needle stack. *Journal of Biomedical Informatics*, 73:51–61.

Gottlieb, A., Stein, G. Y., Ruppin, E., Altman, R. B., and Sharan, R. (2013). A method for inferring medical diagnoses from patient similarities. *BMC medicine*, 11(1):194.

Grnarova, P., Schmidt, F., Hyland, S. L., and Eickhoff, C. (2016). Neural document embeddings for intensive care patient mortality prediction. *Workshop on Machine Learning for Health, NIPS, arXiv preprint arXiv:1612.00467*.

Jo, Y., Lee, L., and Palaskar, S. (2017). Combining LSTM and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842*.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3.

Kokoska, S. and Zwillinger, D. (2000). CRC standard probability and statistics tables and formulae (pp. section 14.7). *Boca Raton, Fla.: Chapman & Hall/CRC*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., and Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, 7(311):311ra174–311ra174.

Liu, F., Chen, J., Jagannatha, A., and Yu, H. (2016). Learning for biomedical information extraction: Methodological review of recent advances. *CoRR*, abs/1606.07993.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094.

Ng, K., Sun, J., Hu, J., and Wang, F. (2015). Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*, 2015:132.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*. Wiley New York.

Suresh, H., Szolovits, P., and Ghassemi, M. (2016). The use of autoencoders for discovering patient phenotypes. *Workshop on Machine Learning for Health, NIPS, arXiv preprint arXiv:1703.07004*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Van Gompel, M., van der Sloot, K., and van den Bosch, A. (2012). Ucto: Unicode Tokeniser. Technical report, Tilburg Centre for Cognition and Communication, Tilburg University and Radboud Centre for Language Studies, Radboud University Nijmegen. http://ilk.uvt.nl/downloads/pub/papers/ilk.1205.pdf.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408.

Wang, F., Hu, J., and Sun, J. (2012). Medical prognosis based on patient similarity and expert feedback. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1799–1802. IEEE.

Wang, Y., Tian, Y., Tian, L.-L., Qian, Y.-M., and Li, J.-S. (2015). An electronic medical record system with treatment recommendations based on patient similarity. *Journal of medical systems*, 39(5):55.

World Health Organization (2004). *International statistical classification of diseases and related health problems*, volume 1. World Health Organization.

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2014). Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, 2014:132.