CSaRUS-CNN at AMIA-2017 Tasks 1, 2: Under sampled CNN for text classification

Arjun Magge, MS[1], Matthew Scotch, PhD[1], Graciela Gonzalez, PhD[2]
[1]Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ, USA;
[2]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Most practical text classification tasks in natural language processing involve training sets where the number of training instances belonging to each of the classes are not equal. The performance of the classifier in such a case can be affected by the sampling strategies used in training. In this work, we describe a cost sensitive and random undersampling variants of convolutional neural networks (CNNs) for classifying texts in imbalanced datasets and analyze its results. The classifier proposed in this paper achieves a maximum F1-score of 0.414 placing 2nd on the ADR dataset and achieves a maximum F1-score of 0.652 placing 6th on the medication intake dataset.

Introduction

Text classification tasks in natural language processing (NLP) can contain datasets where the number of training instances belonging to each class are not equal. The annotation cost for gold standard labels (i.e. labels assigned by humans) are mostly associated with the number of instances labeled in the dataset. When classes are highly imbalanced, the dataset may not contain enough training examples belonging to the minority class. Having highly imbalanced classes with very few instances belonging to minority classes in such datasets can lead to a drop in classification performance. Many classifiers tend to assign the majority class to a given training instance.

Learning from imbalanced data is a big problem and the subject has been studied extensively.[1] Standard approaches to classifying imbalanced datasets in NLP at the data level involve oversampling and undersampling.[2] The most common method to tackling this problem at the algorithmic level is using cost-sensitive learning.[3] In this work, we chose to experiment with the undersampling and cost-sensitive learning methods in CNN architectures to compensate for class imbalance.

For Task 1, the classification dataset contains tweets mentioning a drug and the objective is to classify the tweet into two classes i.e. 1) No-ADR : the tweet contains no evidence to indicate adverse drug reaction (ADR), and 2) ADR : the tweet contains evidence to indicate ADR. Detecting ADRs from social media and health forum texts has been an intensive area of research for early detection of ADRs and possible interventions.[4–8]

For Task 2, the dataset contains tweets mentioning a drug and the objective is to classify the tweet into three classes i.e. 1) Intake : the tweet contains evidence for medication intake, 2) Possible-Intake : the tweet contains evidence to suspect medication intake, and 3) No-Intake : the tweet contains no evidence of medication intake. For additional information about the dataset and its annotations, see Klein et. al.[9]

Method

Input The datasets for the tasks contained tweets-ids and their respective categorical annotations. The first set of annotations provided for the task was used as training dataset and the second set was used development/validation set. The original texts were available for only about 40% of the annotations for Task-1 and 60% for Task-2. In Table 1, we show the details the datasets for both tasks and their respective class distributions.

Classifier For the CNN classifier used in this paper, we implemented our models based on the original CNN architecture as proposed by Kim et. al. for sentence classification.[10] We use this CNN architecture to construct cost sensitive and random undersampling variants to tackle the class imbalance problem. The random undersampling variant (Undersampling-CNN) is constructed by randomly sampling equal number of class-instances in each epoch. This means that there are far fewer training instances in each epoch. The cost sensitive variant

Table 1: Dataset details for the ADR dataset and medication intake dataset.

	Category	Annotated	Available	Class-1	Class-2	Class-3
Task-1	Training Set	10,822	4966	4407	559	-
	Development Set	4845	2178	2024	154	-
Task-2	Training Set	8000	5244	1006	1611	2627
	Development Set	2260	1159	221	374	564

Table 2: Performance comparison of classifiers used in this paper. For task-1, the results are for ADR class i.e. class 1. For task-2, the results are micro-averaged scores for classes 1 and 2. Undersampling-CNN and CostSensitive-CNN was not computed and evaluated for Task-2.

	Implementation	Validation			Evaluation		
		P	R	F1	P	R	F1
Task-1	CNN	0.350	0.490	0.409	0.396	0.431	0.412
	Undersampling-CNN	0.435	0.352	0.389	0.467	0.357	0.404
	CostSensitive-CNN	0.493	0.393	0.438	0.437	0.393	0.414
Task-2	CNN	0.692	0.625	0.656	0.696	0.601	0.645

For our experiments we use a fixed maximum sentence length of 50 words. All tweets with length less than 50 words are padded with zeros. Sentences with more than 50 words are truncated. As pre-processing steps we tokenize each tweet and normalize punctuations. For word embeddings, we use the word vectors generated using millions of tweets containing drug names and made available by Sarker et. al.[11] for mining health related data online.

Hyperparameters For experimentation we use filter sizes in the range of 1 to 5 words with the number of filters i.e. model hidden dimensions in the range 50-150. The best models were obtained for filter combinations of 2,3,4 and 75 filters. A softmax cross-entropy function is used to compute the cost for optimization. For optimization, we use the Adam Optimizer with a learning rate of 0.001.[12] We employ dropout keep probability of 0.5 during training to prevent overfitting.[13] We also apply L2 regularization rate of 0.001 for training across 50 epochs. The model with the best performance on the validation/development set is saved and used on the evaluation/test set. The Undersampling-CNN had fewer training samples per iteration compared to training on the entire set. Hence, the Undersampling-CNN had to be trained at half the learning rate i.e. 0.0005 and took around 40-50 epochs to arrive at the optimal model as compared to 10-15 for CNN and CostSensitive-CNN models. Although we could add feature embeddings for each word in the architecture, we do not add any task specific features.

Results

In Table 2, we show the results for both tasks. For Task-1, the CostSensitive-CNN model was found to achieve the best score. As described earlier, the Undersampling-CNN takes longer to train on all the randomized training samples in majority class. However, the results did not show its improvement over the CNN or the CostSensitive-CNN model.

Conclusion and Future Work

In this work we evaluate a CNN classifer for detecting ADR and medication intakes as part of two shared tasks at AMIA-2017. The classifiers presented in this work placed 2nd and 6th in tasks 1 and 2 respectively. As improvements to the proposed classifiers, we would like to experiment with variants of cost sensitive training architectures in CNN for tackling class imbalance problems as well as strategies to introduce controlled synthetic sentence variants for oversampling the minority class.

References

1. Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

2. Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-supervised learning for im-

balanced sentiment classification. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1826, 2011.

3. Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

4. Xiao Liu and Hsinchun Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *International Conference on Smart Health*, pages 134–150. Springer, 2013.

5. Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, 2014.

6. Abeed Sarker, Karen OConnor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*, 39(3):231–240, 2016.

7. Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, 2015.

8. Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, 54:202–212, 2015.

9. Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. Detecting personal medication intake in twitter: An annotated corpus and baseline classification system. *BioNLP 2017*, pages 136–142, 2017.

10. Yoon Kim. Convolutional neural networks for sentence classification. *EMNLP*, 2014.

11. Abeed Sarker and Graciela Gonzalez. A corpus for mining drug-related knowledge from twitter chatter: language models and their utilities. *Data in brief*, 10:122–131, 2017.

12. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

13. Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.