Data Augmentation Based on Distributed Expressions in Text Classification Tasks

Yu Sugawara
Faculty of Information Science and Technology, Hokkaido University
suga@ist.hokudai.ac.jp

Abstract

We propose a data augmentation method that combines Doc2vec and Label spreading in text classification tasks. The feature of our approach is the use of unlabeled samples, which are easier to obtain than labeled samples. We use them as an aid to the classification model to improve the accuracy of its prediction. We used this method to classify several text data sets including the natural language branch of the AIWolf contest. As a result of the experiments, we confirmed that the prediction accuracy is improved by applying our proposed method.

1 Introduction

Analyzing human intentions in texts is a task in high demand in natural language processing. On the other hand, to solve this task well, it is necessary to prepare an enormous amount of natural language corpora that the intentions of each text are labeled. In particular, if the context is unusual, like in-game conversations, the preprocessed training data that meets the demand is rarely available. Thus we have to manually label intentions one by one or pay for crowdsourcing.

To cope with this situation, we propose a method that can estimate the intention of texts with high accuracy from a large number of unlabeled samples and a relatively small amount of labeled ones.

1.1 Data augmentation via unlabeled samples

There are several existing methods for performing data augmentation based on unlabeled samples. In S-EM(Nigam et al., 2000), a naive Bayes model is first constructed using only labeled samples. The trained naive Bayes model gives unlabeled samples an estimated probability of their label. Then, a new naive Bayes model is constructed using all the samples, both originally labeled and newly labeled. As with the EM algorithm, this procedure is repeated until the parameters of the model converge.

Many of the related methods involve minor changes to S-EM, such as replacing the algorithm used in intermediate steps with a more accurate one(Li and Liu, 2003).

1.2 Word2vec and Doc2vec

Word2vec(Mikolov et al., 2013) is a method that expresses a word as a distributed representation with a high dimensional vector. The regularity of addition and subtraction is shown by vector representation of words such that vector('king') - vector('man') + vector ('woman') approximates vector ('queen'). Word2Vec uses a Bag-of-Words model, which uses the number of occurrences of words in a sentence, and a Skip-gram model, which uses the word occurrence probability from the sequence of words in a sentence.

Doc2vec(Le and Mikolov, 2014) is a method to perform the same operation as Word2vec on a document. It converts a document into a vector representation in high-dimensional space. As with Word2vec, documents that are close in this space can be interpreted as having a similar context.

1.3 Label spreading

Label spreading(Zhou et al., 2003) is a semi-supervised learning method. The goal of semi-supervised learning is to estimate the label of unlabeled samples based on a small number of labeled samples. In label spreading, the label information is propagated from the labeled sample to the unlabeled sample at a close distance. This newly labeled sample also has a influence on the surrounding sample. By repeating this propagation, the label information of labeled samples is spread for all samples.

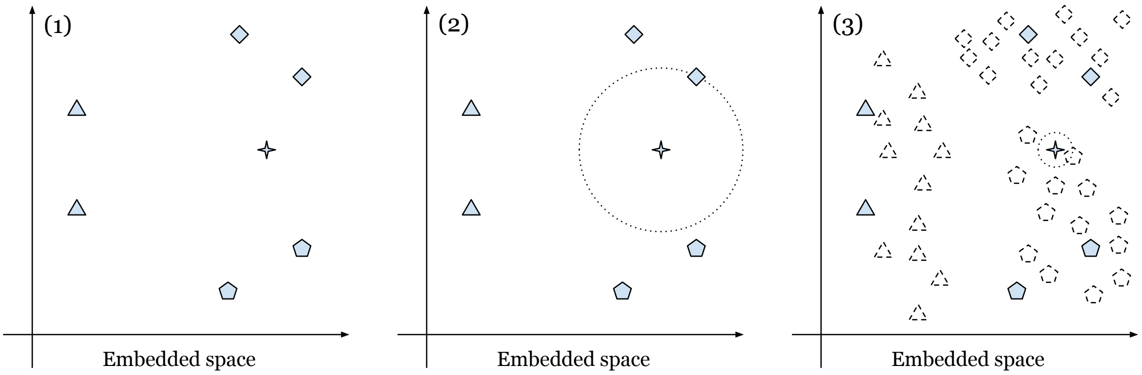

Figure 1: Concept of the proposed method. The figures enclosed by solid lines represent labeled samples embedded in the space. Those enclosed by dashed lines represent originally unlabeled samples. The difference in shape represents the label that the sample has.

2 Our proposed method

We propose a method to estimate the true label of the documents with high accuracy but from a relatively small amount of labeled data.

The model training process is as follows. First, we perform a word segmentation via morphological analysis on all the documents to obtain an ordered list of words. This operation is peculiar to the Japanese language, which is not normally written with a space between words. For that, it may not be necessary when applying this method to other languages such as English. Based on the result, the Doc2vec model is constructed using both labeled and unlabeled training samples. Thus each sample is made to correspond to the coordinate of the high dimensional space. After that, Label spreading is performed in this space. Labeled samples are used to label all the remaining unlabeled samples. The label information is propagated to surrounding samples in embedding space.

In the prediction process, we input the natural language document to the previously trained Doc2vec model to get the vector representation of the sample in the high dimensional space. The Nearest centroid algorithm(Tibshirani et al., 2002) is performed in this space, which estimates the label of the sample based on the neighboring samples. Finally, the true label of this sample is estimated.

We show this method schematically in Figure 1. (1) Our objective is to estimate the label of the sample embedded in the star position. (2) If we simply apply the nearest neighbor algorithm by using just labeled samples, the estimation is not reliable. (3) In our proposed method, the label of unlabeled samples is complemented by Label spreading at first. The Nearest centroid algorithm is applied based on both originally and newly labeled samples.

3 Experiments

3.1 Experimental setting

To verify the effectiveness of the proposed method, we conducted the following experiments. First, we prepared corpora that the intentions are labeled on. Then, we remove the label information from about 90% of the datasets. We trained the Doc2vec model with both labeled and unlabeled data, then use it to embed all samples to high dimensional space. After that, we performed Label spreading to recover label information. For comparison, we also prepared a model that simply executes the Nearest centroid using only the labeled data. Finally, we input the corpora not used for training and compared the prediction accuracy of the true label.

For Label spreading and Nearest centroid, we used the implementations of scikit-learn(Pedregosa et al., 2011).

3.2 Datasets

The following three corpora were used in this experiment.

Livedoor consists of documents published in an online news site. We labeled the topic category in which the news appears. There are nine categories such as ”sports”, ”life hacks” and so on. Our purpose is to estimate the topic category from a news article.

Label	Example (Japanese)
ASK_WHO_LIKE_WEREWOLF	>>Agent[xx] 君は誰が怪しいと考えているのかな？
ASK_WHY_DIVINE	>>Agent[xx] Agent[xx] はどうして私を占ったんだい？
COMINGOUT_VILLAGER	私は人間さ。
COMINGOUT_WEREWOLF	実は私が狼だったんだよ。
DIVINED_HUMAN	占い CO。Agent[xx] は人間だったよ。
DIVINED_WEREWOLF	占い師は私だよ。昨日の結果だが、Agent[xx] は人狼だと出た。
ESTIMATE_HUMAN	Agent[xx] は人間だと思う。
ESTIMATE_WEREWOLF	Agent[xx] が怪しいと思っているよ。
UNIMPORTANT	おはよう。CO はあるかな？
REQUEST_VOTE	Agent[xx] に投票して欲しい。

Table 1: Labels we defined in the AIWolfNLP.

Table 2: Outline of the datasets used in the experiment. Each column indicates the number of labels, the number of unlabeled samples and the number of labeled samples.

	# labels	# unlabeled	# labeled
livedoor	9	6638	663
wolfBBS	9	9343	1038
AIWolfNLP	10	1653	212

WolfBBS consists of utterances generated by humans on Werewolf BBS, an online BBS for playing the Werewolf game. Nine intentions are defined such as ”COMING OUT”, ”DIVINE RESULT”, and so on. Each utterance is annotated one of nine intentions.

AIWolfNLP consists of the utterances in the natural language branch of the 4th AIWolf Contest. We labeled the intention of each utterance generated in the TALK phase. We defined 10 intentions that seem to be useful in understanding the game situation such as ”DIVINED WEREWOLF”, ”REQUEST VOTE”, and so on. Examples of the correspondence between each text and assigned label are shown in Table 1. Our purpose is to estimate the intention of the utterance. In this dataset, just one agent’s utterances are labeled and others are unlabeled. This is a setting that assumes the case of actually participating in the natural language branch of the AIWolf contest. We have a complete set of utterances and intent pairs for the agents we created, but no information about other agents.

A summary of these datasets is presented in Table 2.

3.3 Experimental results

Table 3: The prediction accuracy on validation samples. The simple method discards unlabeled samples and runs Nearest Centroid with only labeled data. The proposed method first completes the labels of unlabeled samples and then runs Nearest Centroid with all the data.

	simple	proposed
livedoor	71.7%	80.3%
wolfBBS	49.2%	50.7%
AIWolfNLP	15.8%	57.4%

The experimental results for each dataset are shown in Table 3. In each dataset, the proposed method that exploits both labeled and unlabeled samples gained higher prediction accuracy than the method simply applying the Nearest centroid using just labeled samples.

4 Conclusion

We proposed an effective prediction method for document classification tasks when a large number of unlabeled samples and a few labeled samples are retained. Our experiments demonstrated that the proposed method gained significantly higher prediction accuracy than the model trained on only labeled samples. It is often the case that the text itself is available in large quantities, but only a few samples are labeled. This method will be quite useful in such situations.

As a prospect, we should conduct similar experiments on languages other than Japanese to confirm the usefulness of the method. The object of the experiment was limited to Japanese in this paper, but since this method has no language dependency, it can also be applied to any language.

References

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Pro-*

ceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1188–1196.

Xiaoli Li and Bing Liu. 2003. Learning to classify texts using positive and unlabeled data. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 587–594.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

R. Tibshirani, Trevor Hastie, B. Narasimhan, and Gilbert Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of National Academic Science (PNAS)*, 99:6567–6572.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 321–328.