하이퍼네트워크 모델을 이용한 텍스트 문장 분류

*작가멧, 김선, 장병탁
서울대학교 컴퓨터공학부
e-mail: *jakramate@bi.snu.ac.kr*, *skim@bi.snu.ac.kr*, *btzhang@bi.snu.ac.kr*

Text Sentence Classification Using Hypernetwork Models

*Jakramate Bootkrajang, Sun Kim, and Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University

Abstract

We propose a text sentence classification approach based on hypernetwork, which is a hypergraph model with weighted hyperedges. The hypernetwork memorizes word segments from sentences, and it is used to classify similar patterns. Our learning procedure is to adjust the weights of hyperedges towards minimizing prediction errors. For experiments, a PPI (Protein-Protein Interaction) filtering task was performed on a biomedical corpus, and the results show that the proposed method is promising for text sentence classification.

I. Introduction

Text classification have been receiving much attention due to the emergence of the digital storage of text documents. The overwhelming text resources, especially biomedical text, urge us to develop an automatic mechanism to filter out irrelevant text sentences. One major challenge in text classification at both document- and sentence-level is that the classifier must be able to capture the complex semantics of language to some extent [1].

Text sentences can be viewed as a collection of building blocks in which each block holds related words and contributes to the meaning of the whole sentence at various degrees. Similarly, the hypernetwork model [2] is a collection of hyperedges in which each hyperedge represents the relationship among features, i.e. words. In particular, the hypernetwork classifier works as a committee machine, where each hyperedge contributes to the final decision of the hypernetwork. Hence, these similar properties motivate the hypernetwork as a potential candidate for sentence classification tasks.

For experiments, we applied the hypernetwork model to a PPI (Protein-Protein Interaction) filtering problem. Hypernetwork classifiers learned word patterns underlying PPI sentences, and evaluated unseen examples based on the learned model. The experimental results show that our method can be effective for text sentence classification.

II. Hypernetwork Classifiers

2.1 Hypernetwork Models

A hypernetwork is a graphical generalized model based on hypergraph models [2]. A hypergraph is defined as $G=(V,E)$ where V is a set of vertices such that $V=\{v_1, v_2, ..., v_n\}$ and E is a set of hyperedges such that $E=\{e_1, e_2, ..., e_m\}$. A hyperedge is an edge in the hypergraph in which more than two vertices can be connected. Each hyperedge is defined as $e_i=\{v_{i1}, v_{i2}, ..., v_{ij}\}$, where j refers to a cardinality of the hypergraph.

A hypernetwork model is defined as $H=(X,E,W)$, where X represents set of all words in dataset, $E=\{e_1, e_2, ..., e_n\}$ is a set of hyperedges, where $e_i=\{x_{i1}, x_{i2}, ..., x_{im}, y_i\}$, i.e. a set with m words from set X and the class y_i. $W=\{w_i, w_2, ..., w_n\}$ is the weights of hyperedges.

One advantage of hypernetwork is that the higher-order correlation terms are explicitly represented by hyperedges, which is an attractive feature in machine learning methods. More details about hypernetwork models can be found in [2,3].

2.2 Learning Hypernetwork Classifiers

Hypernetwork classifiers repeats three simple steps, sampling, storing, and matching. Table 1 describes our hypernetwork learning algorithm for

two class problems.

(1) Begin with $H=(X,E,W)=\{\varnothing,\varnothing,\varnothing\}$.
- Generate hyperedge e_i from sentences by random hypergraph process [3].
- $E \leftarrow E \cup e_i$
- $X \leftarrow X \cup \{x_j | x_j \in e_i\}$
- $W \leftarrow W \cup \{w_i | w_i = w_{init}\}$
- Repeat until E is full.

(2) Let y and y^* be the correct and prediction classes, respectively. ans_0 and ans_1 are the weight sum of each class.
- Generate a hyperedge set T from training sentence by random hypergraph process.
- For each $t_j \in T$, compare to every $e_i \in E$. If t_j matches e_i by allowing one mismatch, then $ans_y \leftarrow ans_y + w_i$.
- $y^*=1$ if $ans_1 > ans_0$. $y^*=0$, otherwise.

(3) Let M_0 and M_1 be a set of matched edge of class 0 and class 1 in previous step.
- If $y^* \neq y$, update $w_i \leftarrow w_i + lRate$ for every hyperedges in M_y. $lRate$ is the learning rate.

(4) (2) and (3) are repeated until terminated.

Table 1. Hypernetwork learning algorithm.

Ⅲ. Experimental Results

The proposed method was applied to a PPI sentence corpus [6] for experiments. The dataset was preprocessed by deleting redundant components and stemming words. Also, insignificant words, i.e. stopwords were removed from sentences.

We have trained and evaluated 3-cardinality hypernetwork using 10-fold cross validation. The number of hyperedges was 2,000,000 and the learning rate was 1.3. Figure 1 illustrates the classification performance during learning procedure. After 100 iterations, our model achieves the average accuracy of 93% while precision and recall rates are about 94%. Therefore, it concludes that learning word patterns by hypernetwork is effective for sentence filtering tasks.

Ⅳ. Conclusion

We have demonstrated the usefulness of the hypernetwork model in text sentence classification. Here, word fragments in training data are randomly sampled to construct a hypernetwork. The hypernetwork is then learned by adjusting the weights of hyperedges. Applied to a PPI filtering task, the experimental results show that our hypernetwork approach performs well, and is enough to be an alternative classifier in sentence-based domain. One of the main advantages in hypernetwork is the human understandability for the learned model. Hence, our future goal is to analyze the behind structure of learned word patterns.

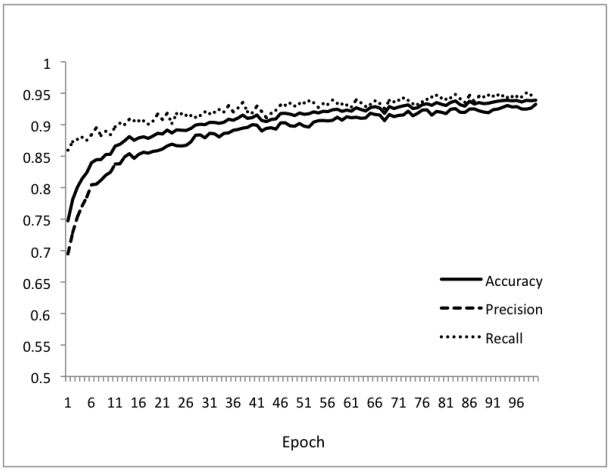

Figure 1. Classification performance during learning procedure.

Acknowledgement

This work was supported by KOSEF through the National Research Laboratory Program (No. M10400000349-06J0000-34910) and the Ministry of Education and Human Resources Development under the BK21-IT Program. The ICT at Seoul National University provided research facilities.

References

[1] D. D. Lewis, "Challenge in machine learning for text classification", *Proceeding of the Ninth Annual Conference on Computational Learning Theory*, pp. 1, 1996.

[2] B.-T. Zhang, "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory", *IEEE Computational Intelligence Magazine*, 3(3) pp. 49-63, 2008.

[3] B.-T. Zhang, "Random hypergraph models of learning and memory in biomolecular networks: shorter-term adaptability vs. longer-term persistency", *The First IEEE Symposium on Foundations of Computational Intelligence*, pp. 344-349, 2007.

[4] S. Kim, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, and B.-T. Zhang, "PIE: An online prediction system for protein-protein interactions from text", *Nucleic Acids Research*, 35, pp. W411-W415, 2008.