ADAPTING DISTANCE KERNEL TO DOMAIN ADAPTATION FOR SENTIMENTAL ANALYSIS

Saerom Park, Jaewook Lee & Woojin Lee
Department of Industrial Engineering
Seoul National University
Seoul, Republic of Korea
{psr6275,jaewook,wj926}@snu.ac.kr

ABSTRACT

Domain adaptation methods aims to improve the accuracy of the target predictive classifier while using the patterns from a related source domain that has large number of labeled data. To solve the domain adaptation problem, we propose new simple and intuitive method that can improve the learning of target data by calculating the distance of each instances in source and target domain. We added distance kernel based cross entropy term in loss function of logistic regression sentimental analysis classifier. We evaluated the proposed method by using cross domain sentiment analysis tasks of Amazon reviews in four domains. Our empirical results showed improvements in all 12 domain adaptation experiments.

1 INTRODUCTION

Recently, with the dissemination of smartphone uses and social network services, sentimental textual data such as reviews, recommendations, replies are proliferating. By using these overflowing data, sentiment analysis which determines the sentiment of the writer has been one of the most popular data application in businesses. There has been various applications in sentimental analysis in various domains such as movie review (Pang et al., 2002), or amazon product review (Blitzer et al., 2007).

Since there are large variety of data domains in textual data, collecting labeled data from all the different domains and building machine learning model for each of them are expensive and time consuming process. However, because traditional machine learning technologies rely on the assumption that the training data and the test data are drawn from same feature space and shares equal probabilistic distribution, it is difficult to build a single sentiment classifier that can be applied to all kinds of domains. For example, in kitchen domain, words such as ”defective”, ”reliable” are widely used, and in book domain ”unclear”, ”boring” can be used.

To build a robust classifier that could be applied to different domains, domain adaptation is receiving attention recently. Domain adaptation considers an environment that training data and test data are from different distributions (Glorot et al., 2011) and we have only labels in training data. It considers training data as source domain and test data as target domain. The purpose of domain adaptation is to build a predictive classifier for target domain by utilizing the knowledge in source domain.

There are two approaches to solve the domain adaptation problem. The first one is instance-based approach. This approach focus on revising the training of the classifier by adding various terms in loss function. There has been researches by using important reweighting to reweight the labeled instances from source domain (Sugiyama et al., 2007). The second approach is feature based domain adaptation. It aims to find common feature structure that can link two different domains for domain adaptation (Pan & Qiang, 2010). There has been research by Glorot et al. (2011), which used stacked marginalized auto-encoder to extract common feature between different domains.

In this research we propose new instance-based domain adaptation approach by using distance-kernel based loss function. By adding distance-kernel based cross entropy in loss function, we could train a model fits well in target domain by using the source input data, source label data, and target input data.

2 PROPOSED METHOD

The basic cross entropy loss function is shown as (1). It calculates cross entropy with binary classifier $f(\cdot)$ of source input data x_source_i with its corresponding labely_i. This cross entropy term is used for ordinary logistic regression.

$$-\frac{1}{N}(\sum_{i=1}^{N}[(y_i) \times f(x_{source_i}) - (1-y_i) \times log(1-f(x_{source_i}))] \tag{1}$$

Therefore, we make revised cross entropy term like in (2), where $k(\cdot,\cdot)$ is kernel function between inputs. We used radial basis function (rbf) kernel like in (3).

$$-\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}[y_i \times log(f(x_{target_j})) - (1-y_i) \times log(1-f(x_{target_j}))]k(x_{source_i}, x_{target_j}) \tag{2}$$

$$k(x_{source_i}, x_{target_j}) = e^{-\frac{||x_{source_i} - x_{target_j}||^2}{\sigma^2}} \tag{3}$$

By adding the term (2), we can approximately calculate the cross entropy error of target data with source labels by using the distance-kernel. Therefore, we proposed domain adaptation sentiment analysis method that uses the loss function (4) to train the classifier of target data in spite of the absence of target labels.

$$-\frac{1}{N}(\sum_{i=1}^{N}[(y_i) \times f(x_{source_i}) - (1-y_i) \times log(1-f(x_{source_i}))]$$
$$-\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}[y_i \times log(f(x_{target_j})) - (1-y_i) \times log(1-f(x_{target_j}))]k(x_{source_i}, x_{target_j}) \tag{4}$$

3 EXPERIMENT

3.1 DATA

In this paper we used amazon review data regarding four different domains: Books, DVDs, Electronics, and Kitchen. We regarded reviews with rating larger than 3 as positive reviews and rating smaller than 3 as negative ones. In our data there were large difference in number of reviews in each domain, and also the number of positive and negative reviews is imbalanced. Therefore, we used each of 3119 positive and negative reviews in every domain to balance the number of the data used in our experiment.

3.2 EXPERIMENTAL SETUP

We conducted 12 cross domain experiments with four domains. Each review test is preprocessed as bag-of-words and transformed in to binary vectors by using unigram model. We used 6786 sentiment words suggested by Baccianella et al. (2010) in our bag-of-words representation. Then we used linear dimension reduction method (SVD) with source and target domain together for extract the common features.
We compared three methodologies in our experiment. The baseline is using logistic regression model with cross entropy loss (1) that is trained on source data, and directly applied to target data. The goal baseline is using normal logistic model with cross entropy loss (1) by using target data as training. Finally our proposed model used logistic regression with loss function as (4). The result of our experiment is shown in Figure 1.

4 DISCUSSION

Compared to baseline, our proposed method showed better results in every 12 experiments. Moreover in some domain adaptation experiments it showed almost same performance as the goal of our

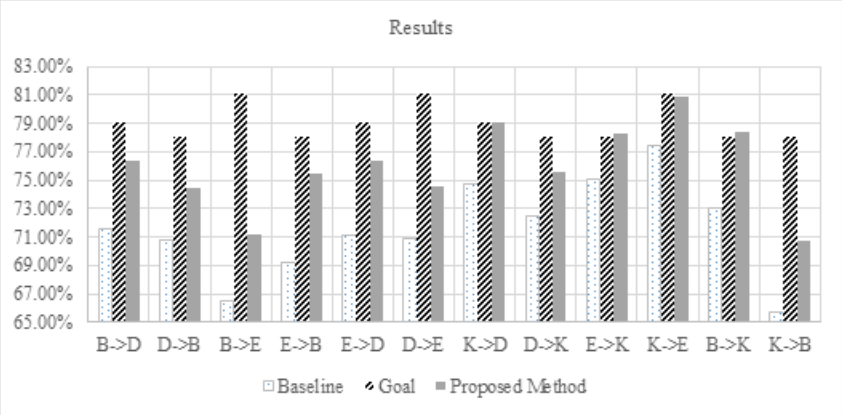

Figure 1: Result of our experiments.

research. We expect our paper will contribute to domain adaptation in sentimental analysis when labeled data in target data does not exist. The proposed methods is intuitive and simple approach that can apply in various domain adaptation problem.
In our further research we plan to consider combining our proposed loss function with previous methodologies in domain adaptation. Also conduct more experiments in various settings in order to find this method can be considered as robust model

REFERENCES

Stefano Baccianella, Esuli Andrea, and Sebastiani Fabrizio. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *LREC*, 10, 2010.

John Blitzer, Dredze Mark, and Pereira Fernando. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. 2007.

Xavier Glorot, Bordes Antoine, and Bengio Yoshua. Domain adaptation for large-scale sentiment classification: A deep learning approach. 2011.

Sinno Jialin Pan and Yang Qiang. A survey on transfer learning. volume 22, pp. 1345–1359. 2010.

Bo Pang, Lee Lillian, and Vaithyanathan Shivakumar. Thumbs up? sentiment classification using machine learning techniques. 2002.

Masashi Sugiyama, Krauledat Matthias, and Mller Klaus-Robert. Covariate shift adaptation by importance weighted cross validation. volume 8, pp. 985–1005. 2007.