Computers, Materials & Continua CMC, vol.60, no.3, pp.1207-1222, 2019

Sentiment Analysis Method Based on Kmeans and Online Transfer Learning

Shengting Wu[1], Yuling Liu[1, *], Jingwen Wang[2] and Qi Li[1]

Abstract: Sentiment analysis is a research hot spot in the field of natural language processing and content security. Traditional methods are often difficult to handle the problems of large difference in sample distribution and the data in the target domain is transmitted in a streaming fashion. This paper proposes a sentiment analysis method based on Kmeans and online transfer learning in the view of fact that most existing sentiment analysis methods are based on transfer learning and offline transfer learning. We first use the Kmeans clustering algorithm to process data from one or multiple source domains and select the data similar to target domain data to establish the classifier, so that the processed data does not negatively transfer the data in the target domain. And then create a new classifier based on the new target domain. The source domain classifier and target domain classifier are combined with certain weights by using the homogeneous online transfer learning method to achieve sentiment analysis. The experimental results show that this method has achieved better performance in terms of error rate and classification accuracy.

Keywords: Sentiment analysis, transfer learning, Kmeans, homogeneous online transfer learning.

1 Introduction

Sentiment analysis, also known as opinion mining or propensity analysis, uses the natural language processing (NLP) and computational techniques to automate the extraction opinions, feelings and subjectivity in the text [Hussein (2016)]. Sentiment analysis involves many real-life scenarios. For example, the government can understand the current public attitude towards policy and the overall direction of public opinion. And for Weibo, WeChat and other social media, Tmall, Amazon and other e-commerce platforms, it can recommend topics and products of interest to users based on sentiment analysis results. At present, sentiment analysis is also a research field for financial data analysis. According to the analysis results, it is recommended to users with reasonable financial products.

The existing sentiment analysis techniques are mainly divided into two categories,

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China.

[2] Department of Computer Science, University of Massachusetts Lowell, MA, USA.

[*] Corresponding Author: Yuling Liu. Email: yuling_liu@126.com.

lexicon-based sentiment analysis methods and machine learning-based sentiment analysis methods [Tripathy, Anand and Rath (2017)]. The lexicon-based sentiment analysis methods need to manually collect sentiment vocabulary and then construct a sentiment dictionary library [Zhang, Wei, Wang et al. (2018)], which mainly relies on open source sentimental dictionaries or extended sentimental dictionaries [Du, Tan, Cheng et al. (2010)]. At present, the domestic open source sentiment dictionary has the HowNet [Zhao, Bing and Liu (2010)], the Chinese emotional vocabulary ontology library of Dalian University of Technology [Chen, Lin and Yang (2009)] and the simplified Chinese emotional polarity dictionary of Taiwan University [Aiping, Peng and Liguo (2015)]. The foreign open source sentiment dictionary has the WordNet of the Princeton University [Lu, Castellanos, Dayal et al. (2011)]. Zubair et al. [Zubair, Aurangzeb, Shakeel et al. (2017)] proposed a lexicon-enhanced sentiment analysis framework using rule-based classification scheme. The machine learning-based sentiment analysis methods can automatically handle large-scale text data without manual intervention. Scholars have proposed a series of machine learning algorithms which have a high accuracy, such as support vector machine algorithm (SVM), decision tree, naive Bayes, random forest, adaboost and so on. The sentiment analysis method based on machine learning is divided into several types according to whether the training samples are labeled, including supervised learning [Shi and Li (2011)], semi-supervised learning [Xi-Shuang, Yi and Zhi-Guang (2014)] and unsupervised learning [Wang and Gupta (2015)]. Unsupervised learning is a popular machine learning technique, and its training samples are not labeled. Fernández-Gavilanes et al. [Fernández-Gavilanes, Álvarez-López, Juncal-Martínez et al. (2016)] proposed an approach based on an unsupervised learning to predict sentiment in online textual messages such as tweets and reviews. Since a large amount of data is unlabeled heterogeneous data, it also requires a lot of time consumption to use unsupervised learning algorithm. Xiang et al. [Xiang, Zhao, Li et al. (2018)] proposed a fast unsupervised heterogeneous data learning approach (TUMK-ELM). Most existing sentiment analysis methods are based on deep learning [Yin, Ye and Yao (2018)]. For example, some scholars conduct sentiment analysis on sentence type classification [Chen, Xu, He et al. (2017)], Chinese microblog [Sun, Li and Ren (2016)] and Chinese sentiment classification model [Xiao, Li, Wang et al. (2018)] based on convolutional neural network. Sentiment analysis based on convolutional neural network can also be applied to the field of financial data analysis [Sohangir, Wang, Pomeranets et al. (2018); Chen, Chen, Huang et al. (2016)].

Traditional lexicon-based and machine learning-based sentiment analysis methods require a large number of training samples to train a reliable classification model. However, it is difficult to obtain large amount of tagged data, and manually tagging of the data can take a lot of time and cost. Transfer learning can solve this problem, which is a hot topic in the field of data mining and machine learning. When there are few or no labeled data in target domain, transfer learning can improve learning performance of target domain by making full use of data in one or more source domains. Although there were a lot of transfer learning methods, most existing works focus on an offline learning fashion [Zhao and Hoi (2010)]. Transfer learning is impossible to effectively handle a lot of datasets and online data streams. Moreover, it cannot avoid the problems of under-adaption and negative transfer. The former refers to the failure to fully solve the

problem of cross-domain samples distribution mismatch. The latter refers to the fact that knowledge learned in the source domain will negatively affect the learning in the target domain, resulting in the performance degradation of transfer learning. An online learning method was proposed, which can respond immediately to solve these problems. Zhao et al. [Zhao and Hoi (2010)] proposed a new machine learning framework called online transfer learning. However, the existing algorithms directly train the source domain data to obtain a classifier. When there have mass data from multiple source domains during the transfer process, these large amounts of data may cause negative transfer due to noise or samples distribution differences. It can reduce the classification effect.

In this paper, we improve an online transfer learning algorithm and propose a sentiment analysis method based on Kmeans and online transfer learning. Firstly, the source domain data and target domain data are clustered by using Kmeans algorithm, then, select the source domain data which is similar to target domain data, so that the effect of dissimilar data on classification models is reduced. Secondly, by employing the homogeneous online transfer learning method proposed in literature [Zhao and Hoi (2010)], the task of online transfer learning is achieved. Finally, the proposed method is applied to sentiment analysis for reviews of Amazon's products of different domains.

The rest of this paper is structured as follows. We describe related work in Section 2, and our method is presented in Section 3. Then, in Section 4, we describe experimental results and analysis. Conclusions and future work is presented in Section 5.

2 Related work

Our work is based on two machine learning topics: online learning and transfer learning. The following is reviewed some important work in these two areas.

Online leaning, also known as random method, has been studied for many years [Hoi, Wang and Zhao (2014)]. Its advantage is that it is difficult to fall into local extreme points in the training process. In general, online learning uses a training sample to update the current model, which can reduce the spatial and time complexities of the learning algorithm, and improve real-time performance. At present, the perceptron learning algorithm [Rosenblatt (1958)] and the passive-aggressive algorithm [Crammer, Dekel, Keshet et al. (2006)] are two most well-known learning algorithms. They update the classifier by calculating the loss value of the predicted results and the actual results for current data, where the perceptron algorithm uses the value of (0, 1) as the loss value. The passive-aggressive algorithm uses the Hinge loss function to calculate the loss value. Some algorithms update the classifier by using the second order information to achieve better results [Yan, Wu, Tan et al. (2016)].

In recent ten years, transfer learning has been widely applied to data mining, computer vision and sentiment analysis. Most important work before 2009 can be found in the literature [Pan and Yang (2010)]. After that, many scholars have also proposed a series of transfer learning algorithms. The goal of transfer learning is to solve learning tasks in target domain by using data from one or more source domains. The multi-source domains transfer learning algorithms can be divided into two kinds: the method based on regularization and the method based on boosting. The former proposes a learning model with designing regular items [Xiang, Pan, Pan et al. (2011)], and in the latter method, the

data is transferred by adjusting different domains or the weight of samples [Eaton and Desjardins (2011)]. Our approach is similar to the method based on boosting, which can be applied to target domain by feature selection of the source domain data.

The traditional offline transfer learning has been researched many years, but there are a little literature considering online transfer learning. As the beginning work of online transfer learning, Zhao et al. [Zhao and Hoi (2010)] and [Yan, Wu, Tan et al. (2016)] focus on online transfer learning with only one source domain, and then systematically define the goals of online transfer learning. Online transfer learning can be divided into two models of homogeneous online transfer learning and heterogeneous online transfer learning according to differences and similarities of the feature space of data. In the homogeneous online transfer learning, the source domain data and target domain data have the same samples space and dimension and share a common feature space [Li, Song and Huang (2017)]. But the heterogeneous online transfer learning is the exact opposite in samples space and dimension. The representations of the two methods that are ultimately used to process the data are vectors. The first study of online transfer learning based on multi-source domains is literature [Ge, Gao and Zhang (2013)]. After that, Wu et al. [Wu, Zhou, Yan et al. (2017)] also proposed a method of transfer learning in multiple source domains, which adjust transferred weight of each source domain to obtain an ensemble classifier. Wu et al. [Wu, Wu, Zhou et al. (2017)] also introduced multiple sources which are homogeneous or heterogeneous. These online transfer learning methods in one or more source domains are easy to train and test all the data. But when the data scale is large, if the data from different fields is completely different from the target samples distribution, it will reduce the effect of the classifier. Therefore, in our proposed method, the data in the source domain is processed ahead of transfer learning. Then the selected source domain data are used for sentiment analysis based on online transfer learning.

3 Methodology

This section introduces the steps of the proposed method in detail, which include text preprocessing, text vector generation, feature selection algorithm based on Kmeans, homogeneous online transfer learning methods and sentiment analysis methods based on kmeans and online transfer learning. Firstly, the source domain data and a small number of target domain data are preprocessed for text segmentation and stop words deletion. Secondly, map the text to word vector by using Word2Vec model [Zhang, Wang, Yu et al. (2018); Giatsoglou, Vozalis, Diamantaras et al. (2017)]. Thirdly, calculate the similarity between the source domain word vector and target domain word vector based on Kmeans method, and then select the data in the source domain, which is similar to target domain sample. Finally, execute online transfer learning algorithm in the selected data and new target domain data, where the knowledge of source domain is transferred to the target domain and complete the task of sentiment analysis. The symbolic definitions used in this paper are listed in Tab. 1.

Table 1: Symbols and descriptions

Symbol	Description
Xs, Ys	source domain data and label: $X_s=R^m$, $Y_s=\{-1,+1\}$
Xt, Yt	target domain data and label: $X_t=R^m$, $Y_t=\{-1,+1\}$
h(x)	source domain classification function
f(x)	target domain classification function
α	support vector coefficients
$K_s(x1,x2)$	source domain classification function kernel
$K_t(x1,x2)$	target domain classification function kernel
Π(x)	normalized function
ω	weight
σ	kernel function parameters

3.1 Feature selection algorithm based on Kmeans

In recent years, many scholars proposed a number of sentiment analysis methods based on transfer learning. In the online transfer learning methods, the classifier should be established based on the source domain data or the auxiliary domain data. Then the classifier established at the beginning would combine with the target domain classifier in a certain way that realize the transfer of knowledge. However, in general case, the source domain data and target domain data have some differences, especially in the case of multiple source domains. If the source domain data is applied directly to the online transfer learning method, it can cause negative transfer because of the domains differences, which led to reduce the performance of the classifier.

In order to avoid the negative transfer, we need to cluster the data that is unlabeled. Yang et al. [Yang, Tan and Zhang (2018)] proposed a clustering method based on DBSCAN which is a density clustering algorithm. Different from the algorithm mentioned earlier, we propose a feature selection algorithm based on Kmeans, and apply it to sentiment analysis method. Kmeans is a kind of unsupervised clustering method and is presented in detail in literature [Hartigan and Wong (1979)]. Because the Kmeans algorithm is simple and easy to implement, it can accord with the requirement of online transfer learning. The Kmeans algorithm is introduced to cluster the source domain data and target domain data in this paper. Then we can select the data from multi-source domains data which is most similar to target domain. When the data is mapped to the word vectors, these word vectors also include syntactic and semantic information. It can measure the semantic similarity between two words by calculating the distance of word vector. After mapping the first round data from source domain and target domain to word vector with Word2Ve-

c model, the distance between the source domain word vector and the target domain word vector is computed by using the formula (1), where m and n are respectively the line number of the source domain word vector matrix and the target domain word vector matrix, and j is the word vector dimension. Then the average value of these distances is calculated based on the formula (2) and used as a measurement standard to select data, which is less than the average distance value. The selected data is more similar to the target domain data. The implementation process of the algorithm is described in detail in algorithm 1.

$$d_{mn} = \sum_{i=1}^{j} | x_{mi} - x_{ni} | \tag{1}$$

$$avg = \frac{\sum_{k=1}^{m} d_{kn}}{m} \tag{2}$$

Algorithm 1 Feature selection algorithm based on Kmeans

Input: X_s, a small number of X_t

Output: Source domain data X'_s after selection

Steps:

Step 1: Using the Word2vec model, the X_s and X_t are mapped from text to word vectors.

Step 2: For each row of source domain word vector, the distance between the source domain word vector and the target domain word vector is calculated by using the formula (1).

Step 3: Using the formula (2), the average distance value from all source domain word vectors to the target domain word vector is computed to use as a measurement.

Step 4: Select the data which the distance value in the word vector of source domain is less than the average value as X'_s.

3.2 Online transfer learning methods

Because the offline transfer learning cannot handle the way that the target domain data is received in a streaming fashion, an online and real-time way is needed to efficiently realize knowledge transfer. The main idea of online transfer learning is to combine the source domain classifier with the classifier established on the online data and then realize the transfer in the combination process. Thus, the problem is to confirm what weight are used while combining the two classifiers. Hoi et al. [Hoi, Wang and Zhao (2014)] proposed an online transfer learning method, which could update the weight and classification model with the dynamically adding of data. Thus, this paper uses the homogeneous online transfer learning method and combines it with the Kmeans-based feature selection algorithm for the task of sentiment analysis.

The classifier is established according to the source domain data, where the classification function *h(x)* is shown in formula (3). It is established based on support vector machines (SVM), where (x_s y_s) is the support vector of source domain data. Then for each target domain data, a classifier is built for the new data. The classifier established in target domain and the final combination fashion in source domain classifier and target domain are different in homogeneous OTL [Yan, Wu, Tan et al. (2017)] methods and heterogeneous OTL methods. We describe the homogeneous OTL method in detail as below.

$$h(x) = \sum \alpha_s y_s k_s(x_s, x) \tag{3}$$

The goal of homogeneous OTL is to combine *h(x)* with the classifier $f_t(x)$ in the target domain samples x_t from the *t*-th round. Firstly, a new classification function *f(x)* is constructed based on online data in the target domain. Secondly, the *h(x)* and *f(x)* are combined with a certain weight to realize the knowledge transfer from the source domain to the target domain. In order to combine two classifiers effectively in the learning of the *t*-th round, two weighting parameters ω_s and ω_t are introduced, which are source domain classifier weights and target domain classifier weights. Thirdly, the combination fashion in the *t*-th round is shown in formula (4). When the algorithm starts, $\omega_s=\omega_t=1/2$. In order to effectively conduct online transfer learning tasks, a passive-aggressive (PA) online learning method is used to update ω_s and ω_t, and the updated method is shown in formula (5) and (6). $s_t(g) = \exp\{-\eta\varepsilon * \Pi(g(x_t)), \Pi(y_t))\}$, $\varepsilon(z, y)$ is a loss function, $\varepsilon(z, y) = (z - y)^2$. The detailed steps of homogeneous OTL method are described in Algorithm 2.

$$u_t = sign(\omega_s \Pi(h(x_t)) + \omega_t \Pi(f(x_t)) - 1/2) \tag{4}$$

$$\omega_{s,t+1} = \frac{\omega_{s,t} * s_t(h)}{\omega_{s,t}(h) + \omega_{t,t} * s_t(f)} \tag{5}$$

$$\omega_{t,t+1} = \frac{\omega_{t,t} * s_t(f)}{\omega_{t,t} * s_t(h) + \omega_{t,t} * s_t(f)} \tag{6}$$

$$\tau_t = \min\{C, \frac{\varepsilon_t}{k_t(x_{t,t}, x_{t,t+1})}\} \tag{7}$$

$$f_{t+1} = f_t + \tau_t y_{t,t} k_t(x_{t,t}, \cdot) \tag{8}$$

Algorithm 2 Online transfer learning algorithm

Input: x_s, y_s, ω_s and ω_t
Output: $h(x), f_t(x)$
Steps:
Step1: Given X_t, Y_t, $h(x)$, and parameter C, and initialization: $\omega_{s,1}=\omega_{t,1}=1/2$, $f_1=0$;
Step2: For t=1, ... T, using the formula (4) to calculate the predictive label U_t, then using the formula (5) and the formula (6) to calculate the $\omega_{s,t+1}$ and $\omega_{t,t+1}$, set the loss value $\varepsilon_t = [1 - y_t f_t(x_{t,t})]$, if the $\varepsilon_t > 0$, then use the formula (7) and the formula (8) to update the target domain classifier.

3.3 Sentiment analysis methods based on Kmeans and online transfer learning

In this paper, an online transfer learning method is proposed to transfer the existing knowledge into the dynamically changing domain for sentiment analysis. However, the online transfer learning is more vulnerable to the negative transfer effect of the source domain data which reduce the classification performance of online data. Thus, we combine it with feature selection algorithm based on Kmeans and propose a sentiment analysis method based on Kmeans and online transfer learning. The detailed description of the method is given in Algorithm 3. Firstly, the source domain data X_S and a few target domain data X_T are first mapped text to word vectors by using the Word2vec model. Secondly, the Kmeans algorithm is used to compute the distance from each line of word vectors in the source domain to the target domain word vector. The average distance is measured as a criterion, and the source domain data with the distance value less than the average distance value is selected and set to X'_s. Finally, the source domain classifier is built according to X'_s and combined with the target domain classifier. The sentiment analysis task based on online transfer learning is implemented.

Algorithm 3 Sentiment analysis method based on Kmeans and online transfer learning

Input: X_s, Y_s, a small number of X_t, Y_t
Output: Sentiment label for target domain data
Steps:
Step 1: Using the Word2vec model, X_s and X_t are mapped from text to word vectors.
Step 2: Taking the target domain data word vector as the centroid, the Kmeans algorithm is used to select the suitable source domain data X'_s;
Step 3: The X'_s and new target domain data are iteratively executed online transfer learning algorithm, the source domain knowledge are transferred to the target domain, the target domain classification model M is established;
Step 4: Use M to classify online target domain data and obtain sentiment label.

4 Experimental results and analysis

The proposed method is implemented by using MATLAB software on Windows10 operation system. The datasets, parameter settings, and analysis of the experimental results are described in detail as below.

4.1 Experimental data settings

The sentiment analysis datasets use the review data from Amazon's different domains products. There are four kinds of books, DVDs, electronic products and kitchenware. The datasets are widely used in the study of transfer learning. All the data used to do the experiment is labeled, where the positive comment label is +1, and the negative comment label is -1. Each domain has 2000 pieces of comments data and the number of positive comment and negative comment are the same. After mapping the text to the word vector, data dimension of each field are 473857 dimensions, where the first dimension of the data is the sentiment tag and the other 473856-dimensional data are the word vectors. Thus the scale of the data in each domain is 20004*73857. The specific experimental data is given in Tab. 2. In order to evaluate the algorithm proposed in this paper, we do some experiments on the data. The data in one domains is used as the target domain data, and data in the other three domains are used as the source domain data. For example, when the book field is the target domain, the combination of DVD, electronic products, kitchenware is the source domains. Because it is online transfer learning, we also compare the accuracy of the algorithm when the target domain changes.

In order to compare the experimental results, we have implemented three other online learning methods. The first is the passive aggressive algorithm (passive-aggressive, PA) [Crammer, Dekel, Keshet et al. (2006)], where PA is set to do online learning tasks directly on target domain data without using any source domain data. The second is to add the source domain data on the basis of the first algorithm. All the source domain data are used to train the classifier, and then perform the online learning task on the target domain, that is, the source domain initializes passive-aggressive algorithm, short for PAIO. The third is the homogeneous online transfer learning algorithm (homogeneous online Transfer Learning, HomOTL) [Hoi, Wang and Zhao (2014)].

Table 2: Specific experimental data

Field	Total number of comments	Number of positive comments	Number of negative comments	Data dimension
book	2000	1000	1000	473856
DVD	2000	1000	1000	473856
Electronic Products	2000	1000	1000	473856
Kitchen	2000	1000	1000	473856

The parameters of the experiment are initially set as shown in Tab. 3. The kernel functions used by all algorithms are Gaussian kernel functions, that is $K(x, y) = \exp\{-\| x - y \|^2 / 2\delta^2\}$. The parameters of Gaussian kernel function in source domain is $\delta_1 = 4$. The parameters of Gaussian kernel function in target domain is $\delta_2 = 8$. In addition, the regularization parameters used in all algorithms are c=5. In order to obtain a more stable evaluation result, 20 tests are performed on the target domain, and the target domain data is randomly rearranged before each test.

Table 3: The initialization settings of experimental parameters

Parameter name	Parameter values
$\omega_{s,1}$	1/2
$\omega_{t,1}$	1/2
$f_1(x)$	0
C	5
δ_1	4
δ_2	8
Number of tests	20

4.2 Comparison and analysis of experimental results

The performance comparisons of each algorithm after running on different datasets are shown in Tab. 4. The target domain is respectively books, DVDs, electronic products, and kitchenware, and the efficiency of the algorithm is measured from the error rate and time-consuming by experiment. The experiment in this paper conducted 20 times on each target domain. Before each test, the data in the target domain is randomly arranged again as the input of the algorithm. Then the average value of the 20 experimental results is taken as the final error rate, and the standard deviation of the 20 experimental results are used as the error floating value. According to the experiment results, the proposed sentiment analysis method based on Kmeans and online transfer learning (abbreviated as KMEANS-HomOTL) in this paper is better than other algorithms. In terms of error rate, the KMEANS-HomOTL's total error rate is lower than the overall error rate of other algorithms. Compared with PA, PAIO and HomOTL, the error rate is lower on average 17.5%, 2.5% and 8.8%. The lowest error rate is the field of kitchenware, only about 19.52%. The PA online learning algorithm, which does not use the source domain data, has the highest error rate. The PAIO algorithm, with the assistance of data in the source domain and the transfer of knowledge, has an average error rate of 16% lower than that of the PA algorithm when other parameters and calculation methods are the same. The effect of the other two algorithms using the source domain data has also been significantly improved. It can be seen that the auxiliary function of the source domain

data can greatly enhance the performance of the algorithm.

Table 4: Error rate comparisons of algorithms

Algorithm name	Target domain(%)			
	Books	DVD	Electronic Products	Kitchen
PA	45.02±1.23	44.29±1.28	42.29±1.15	40.70±0.85
PAIO	30.36±0.56	29.28±0.59	25.31±0.79	23.09±0.56
HomOTL	36.90±1.11	36.03±1.11	32.00±1.15	30.10±0.56
Kmeans-HomOTL	29.24±0.10	28.81±0.11	22.36±0.11	19.52±0.07

By comparing the error rate of HomOTL and Kmeans-HomOTL, although these two methods use the same homogeneous online transfer learning algorithm in the same datasets, the error rate of the homogeneous online transfer learning with clustering algorithm is obviously lower. We can see when the data is large or the data from multi-source domains, the performance of the algorithm may be reduced because of the large samples differences. By processing the data from multiple source domains, we select the data which the samples distribution is as close as possible. Although the number of source domain data is reduced because there is no negative transfer effect, the performance of the algorithm is still good. About the time-consuming, it can be seen from Tab. 5, the execution time of the PA algorithm is shortest, because it has no source domain data. But for the other three algorithms, the size of the execution time varies little.

Table 5: Comparison of the time-consuming situation of the algorithm

Algorithm name	Target domain (s)			
	Books	DVD	Electronic Products	Kitchen
PA	0.30±0.01	0.30±0.01	0.36±0.18	0.29±0.01
PAIO	0.62±0.04	0.61±0.02	0.79±0.30	0.63±0.03
HomOTL	0.84±0.07	0.81±0.02	0.98±0.42	0.84±0.02
Kmeans-HomOTL	0.82±0.08	0.80±0.02	1.07±0.43	0.83±0.03

When the target domain changes and the number of samples changes (from 200 to 2,000 pieces), we also compare the classification performance and time-consuming of the four algorithms. The experimental results are shown in Figs. 1 and 2. The error rate of the four algorithms decreases with the increase of samples data in the target domain and the descending amplitude is very similar. The error rate of the HomOTL algorithm is the greatest. In the DVDs target domain, when the sample number increase from 200 to 2000,

the error rate of PAIO and Kmeans-HomOTL are getting closer, but they are quite different when the target domain sample is smaller. This shows that the algorithm proposed in this paper is better than other algorithms in the case of few samples in the target domain. It can be seen from Fig. 2, in terms of time-consuming, because there is no source domain data to be processed, the PA algorithm performs the shortest execution time. However, due to the use of Kmeans clustering algorithm to process source domain data, the overall execution time of the Kmeans-HomOTL algorithm will lengthen with the data size increases.

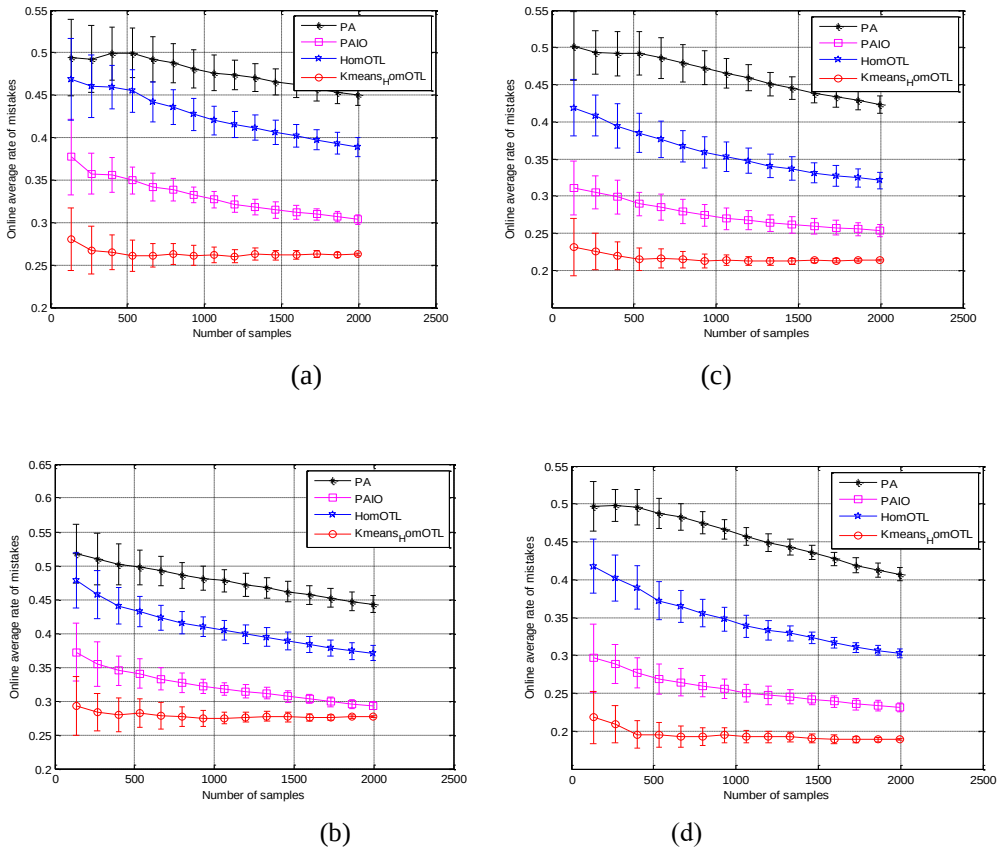

Figure 1: When the target domain is book (a), DVD (b), electronic products (c), Kitchen Utensils (d), the performance of four algorithms

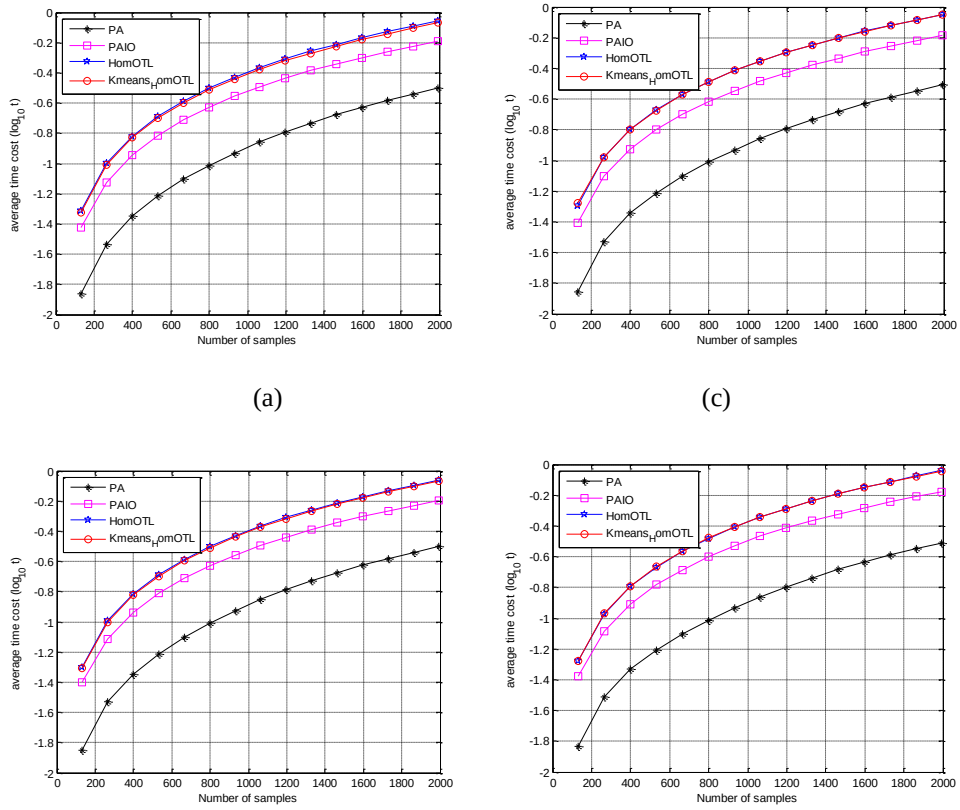

Figure 2: When the target domain is book (a), DVD (b), electronic products (c), Kitchen Utensils (d), the time-consuming of four algorithms

5 Conclusion

This paper proposes a sentiment analysis method based on Kmeans and an improved online transfer learning algorithm. Although the Kmeans-HomOTL algorithm reduces the error rate and improves the classification performance by selecting the source domain data to avoid negative transfer, the memory consumption is extremely high when dealing with a large-scale datasets. One of the main reasons is that the kernel function needs a large memory space to save the support vector, which leads to a lot of memory consumption. Therefore, we plan to improve the efficiency as well as the accuracy of our method in the future. In addition, when the data from different domains are heterogeneous, how to deal with the source domain data is also a direction worth studying.

Acknowledgments: This work was partially supported by National Natural Science Foundation of China (Nos. 61872134, 61502242), Natural Science Foundation of Hunan Province (No. 2018JJ2062), and 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province.

References

Aiping, L. I.; Peng, D. I.; Liguo, D. (2015): Document sentiment orientation analysis based on sentence weighted algorithm. *Journal of Chinese Computer Systems*, vol. 36, no. 10, pp. 2252-2256.

Chen, J. M.; Lin, H. F.; Yang, Z. H. (2009): Automatic acquisition of emotional vocabulary based on syntax. *CAAI Transactions on Intelligent Systems*, vol. 4, no. 2, pp. 100-106.

Chen, J. F.; Chen, W. L.; Huang, C. P.; Huang, S. H.; Chen, A. P. (2016): Financial time-series data analysis using deep convolutional neural networks. *Proceedings of 2016 7th International Conference on Cloud Computing and Big Data*, pp. 87-92.

Chen, T.; Xu, R, F.; He, Y, L.; Wang, X. (2017): Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, vol. 72, no. 1, pp. 221-230.

Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; Singer, Y. (2006): Online passive-aggressive algorithms. *Journal of Machine Learning Research*, vol. 7, no. 3, pp. 551-585.

Du, W.; Tan, S.; Cheng, X.; Yun, X. (2010): Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 111-120.

Eaton, E.; Desjardins, M. (2011): Selective transfer between learning tasks using task-based boosting. *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 337-342.

Fernández-Gavilanes, M.; Álvarez-López, T.; Juncal-Martínez, J.; Costa-Montenegro, E.; González-Castaño, F. J. (2016): Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, vol. 58, no. 1, pp. 57-75.

Ge, L.; Gao, J.; Zhang, A. (2013): OMS-TL: A framework of online multiple source transfer learning. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pp. 2423-2428.

Giatsoglou, M.; Vozalis, M. G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G. et al. (2017): Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, vol. 69, no. 1, pp. 214-224.

Hartigan, J. A.; Wong, M. A. (1979): Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108.

Hoi, S. C.; Wang, J.; Zhao, P. (2014): Libol: a library for online learning algorithms. *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 495-499.

Hussein, E. M. D. (2016): A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, vol. 30, no. 4, pp. 330-338.

Li, S.; Song, S.; Huang, G. (2017): Prediction reweighting for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1682-1695.

Lu, Y.; Castellanos, M.; Dayal, U.; Zhai, C. (2011): Automatic construction of a context-aware sentiment lexicon: an optimization approach. *Proceedings of the 20th International Conference on World Wide Web*, pp. 347-356.

Pan, S. J.; Yang, Q. (2010): A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359.

Rosenblatt, F. (1958): The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65, no. 6, pp. 386-408.

Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T. M. (2018): Big data: deep learning for financial sentiment analysis. *Journal of Big Data*, vol. 5, no. 1, pp. 3-11.

Sun, X.; Li, C.; Ren, F. (2016): Sentiment analysis for chinese microblog based on deep neural networks with convolutional extension features. *Neurocomputing*, vol. 210, no. 1, pp. 227-236.

Shi, H. X.; Li, X. J. (2011): A sentiment analysis model for hotel reviews based on supervised learning. *Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 950-954.

Tripathy, A.; Anand, A.; Rath, S. K. (2017): Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems*, vol. 53, no. 3, pp. 805-831.

Wang, X.; Gupta, A. (2015). Unsupervised learning of visual representations using videos. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794-2802.

Wu, Q.; Zhou, X.; Yan, Y.; Wu, H.; Min, H. (2017): Online transfer learning by leveraging multiple source domains. *Knowledge and Information Systems*, vol. 52, no. 3, pp. 687-707.

Wu, Q.; Wu, H.; Zhou, X.; Tan, M.; Xu, Y. et al. (2017): Online transfer learning with multiple homogeneous or heterogeneous sources. *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1494-1507.

Xiang, L.; Zhao, G.; Li, Q.; Hao, W.; Li, F. (2018): TUMK-ELM: a fast unsupervised heterogeneous data learning approach. *IEEE Access*, vol. 6, no. 1, pp. 35305-35315.

Xiao, Z.; Li, X.; Wang, L.; Yang, Q.; Du, J. et al. (2018): Using convolution control block for Chinese sentiment analysis. *Journal of Parallel and Distributed Computing*, vol. 116, no. 1, pp. 18-26.

Xiang, E. W.; Pan, S. J.; Pan, W.; Su, J.; Yang, Q. (2011): Source-selection-free transfer learning. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3, pp. 2355-2360.

Xi-Shuang, D.; Yi, G.; Zhi-Guang, L. (2014): Sentiment analysis of chinese micro-blog based on semi-supervised learning. *Journal of Shandong University*, vol. 49, no. 11, pp. 37-42.

Yan, Y.; Wu, Q.; Tan, M.; Min, H. (2016): Online heterogeneous transfer learning by weighted offline and online classifiers. *Proceedings of European Conference on Computer Vision*, pp. 467-474.

Yan, Y.; Wu, Q.; Tan, M.; Ng, M. K.; Tsang, I. W. (2017): Online heterogeneous transfer by hedge ensemble of offline and online decisions. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 99, no. 1, pp. 1-12.

Yang, K.; Tan, T.; Zhang, W. (2018): An evidence combination method based on SCAN clustering. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 269-281.

Yin, L.; Ye, X.; Yao, J. (2018): A sentiment analysis method based on BLSTM and CNN fusion. *Journal of Physics: Conference Series*, vol. 1087, no. 6, pp. 062058-062066.

Zhang, S.; Wei, Z.; Wang, Y.; Liao, T. (2018): Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems*, vol. 81, no. 1, pp. 395-403.

Zhao, Y. Y.; Bing, Q.; Liu, T. (2010): Sentiment analysis. *Journal of Software*, vol. 21, no. 8, pp, 1834-1848.

Zhao, P.; Hoi, S. C. (2010): OTL: a framework of online transfer learning. *Proceedings of the 27th International Conference on Machine Learning*, pp. 1231-1238.

Zhang, C.; Wang, X.; Yu, S.; Wang, Y. (2018): Research on keyword extraction of Word2vec model in Chinese corpus. *Proceedings of 2018 IEEE/ACIS 17th International Conference on Computer and Information Science*, pp. 339-343.

Zubair, A. M.; Aurangzeb, K.; Shakeel, A.; Maria, Q.; Ali, K. I. (2017): Lexion-enhanced sentiment analysis framework using rule-based classification scheme. *PloS One*, vol. 12, no. 2, pp. 171649-171671.