FSFP: Transfer Learning From Long Texts to the Short

Wei Fengmei[1,2,*], *Zhang Jianpei*[1], *Chu Yan*[1] *and Yang Jing*[1]

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin150001, China
[2] School of Technology, Harbin University, Harbin, China

Abstract: Transfer learning is a method that studies how to identify the useful knowledge and skills in the previous tasks, and uses them to the new tasks or domains. At present, the research on transfer learning mostly focuses on the field of long texts. However, the source data should be given for the transportation from long texts to the short ones, and the priori probability distribution of the data should be given at the same time. In order to solve the problems, the algorithm which is called FSFP (Free Source selection Free Priori probability distribution) is proposed. It can transfer knowledge from the long texts to the short ones. Latent semantic analysis is used to extract the key words as seed characteristic sets, which are semantically related to the long texts from the target domain. And then the graph structure of online information is built. With the help of the improved Laplacian Eigenmaps, the feature representations of high-dimensional data are mapped to a low-dimensional space. Lastly, the target data are classified in the constraint of minimizing the mutual information between the instance and the feature representation. The experimental results on large data sets show the effectiveness of the new algorithm.

Keywords: transfer learning, latent semantic analysis, Laplacian Eigenmaps, graph, mutual information

1 Introduction

Transfer learning, which the researchers have paid more and more attentions to recently, is a new mode of machine learning. It relaxes the traditional machine learnings requirement that the data from source field and the target domain must be under the same distribution. When the training data, especially the labeled training data in the target domain are too insufficient to produce an accurate classifier model, transfer learning studies how to transfer useful knowledge from the related auxiliary data, which can enrich the characteristics of the target domain, to help learning task on target domain[1,2,3].

With the rapid development of science and technology, the internet information becomes more diverse and short texts. Weibo, QQ news, online advertising and so on, play an increasingly important role in the network application. As short text data have few keywords and lack of context information, its high-dimensional sparse representation is hardly to express the contents of the texts completely and accurately. When the short texts as target data have only a small amount of labeled data, it is difficult for data analysis and classification. Compared with the short text, long texts carry more contexts and subject relevant keywords. And

long text appears in the network earlier; their classification technology is more mature. If we can learn knowledge and classification techniques from the long texts, to classify the short data will become easier.

The previous works on transfer learning form long text to long text may get well effect. But as the different data structures between long text and short text, if we apply transfer learning method of the long text domain to the field of short text directly, it hardly to get satisfactory results. Some algorithms achieved the goal that transfers knowledge from long texts to the short. However, in their work, a major assumption is that source data are provided by the problem designers. The problem how to get the right source data puts a big burden on the designer. Whats more, some solutions to transfer knowledge from long texts to the short demand to know the datas prior probability distribution, which is hardly to get.

In this paper, we propose a novel algorithm called FSFP. For simplicity, firstly we enter tags as keywords extracted from target data to a search engine, and select the first few pages of the web as the most relevant data. By using the algorithm of latent semantic analysisit can extract semantics keywords as seed feature sets from the source domain which is associated with the target domain

---

data. Then build undirected graph of the social media online. Tags make its nodes. We can get the Subgraph that contains all feature seed sets. By improving the algorithm of Laplacian Eigenmapsmap, each node is mapped to a low-dimensional space. We can get new feature representations of the short texts. Finally, minimizing the mutual information between the characteristics of the data and the target labels as a constraint, we can get an accurate classifier on the short text data.

In summary, three main contributions are as follows: 1) The proposed algorithm does not need to be given source data in advance and does not need to know the priori probability distribution of the data ; 2) We use online information as the auxiliary data which are data-rich, comprehensive, easy to get; 3) The algorithm has strong scalability. When the target domain data or tasks change, the algorithm can still be used.

2 Related Works

2.1 transfer learning in the domain of long text

In this section, we summarize our previous representative work on transfer learning. Dai et al. [9] proposed a Tradaboost algorithm, which improved the boosting technology in order to create an automatic weight adjustment mechanism. It can filter out most of the data similar to the target areas from the source field so that it can enrich the training data to improve the accuracy of the classifier; Mei et al. [5] proposed a WTLME algorithm. It bases on maximum entropy model, using instance weighted technology. The algorithm transfers model parameters studied from the original field to the target domain. It can reduce the time of re-collection and marking a large number of target data as well as the time of training model, achieving domain adaptation; Hong et al. [6] proposed a TrSVM algorithm. It requires weak similarity, which is defined by themselves, between two fields meeting certain constraints. Contact this constraint with the target classification, and embed the source data is into the support vector machine training. These transfer learning algorithms are based on instance, which can get a good result, but requires that the source data and the target data must be very similar. When the Source field and target area have a big difference, at the characteristic level, there are often some intersections which will become a bridge to connect two different domains, helping realize the knowledge transfer. Dai et al. [10] proposed a CoCC algorithm, in which the co-occurrence of words in the source field and the target domain were used as a bridge. The tag structures of the source field and the target domain were collaborative clustering at the same time. By minimizing the mutual information between words and samples, it can achieve the goal that transfer the tag structure of the source domain to the target domain; Xue et al. [8] proposed a TPLSA

algorithm. As the source and target domain are related, there may be some common topics, which can be used as a bridge to connect the two areas. From the source domain, transfer the useful information, similar to the target subjects to the target field and help the target classification task; Long et al. [4] proposed a GTL algorithm, which think that the document is a collection of hidden topics and has its inherent geometry between keywords and the document. By extracting the potential common themes between source domain and target domain, optimizing maximum likelihood function, and maintaining the geometric structure of the documents as the same time, it can make the process of transfer learning smoother; The above algorithm is mainly used in the same language of the text fields. When the languages of the two fields were different, Ling et al. [7] found that although the data has a different text feature, but it may have a great relevance on semantic. Therefore they proposed an information bottleneck model. First of all, the Chinese web pages were translated into English. Then put these web pages and English web pages into the information bottleneck model for information coding. Finally filtered out the common information in the two areas and solved the problems of translation inaccuracy, label drift and so on. It made cross-language transfer learning possible.

All these algorithms above are transfer learning methods in long text domain. It may get good results when transfer knowledge from long text to long text. But it is not applicable for short text as there are too many differences such as data structures, forms et al. between the two domains.

2.2 transfer learning from the long text domain to the short

Most existing works focus on this kind of transfer learning that transfer knowledge from long text to the long. Although rarely, there also some people did the work that transfer methods or skills form the long text to the short. For example, Jin et al. [11] proposed a DLDA model. It extracted two sets of topics from the source and target domains and used a binary switch variable to control the forming process of the documents. As it allowed the two areas share the topic structures, transferring part related and useful data from the long texts to the target domain. However, in the algorithm, the source data were required to provide before, and the priori probability distribution of the data were also needed. Source data would affect the target task. It put a great burden on the designer to find the right data. How to filter the similar, relevant, and useful data, and how to accurately calculate the datas prior probability distribution became a big problem.

3 Free Source selection Free Priori probability distribution (FSFP)

3.1 Problem Definition

In the target domain we have few labeled data. $T^i = \{(x_i, y_i) | i = 1, 2, \ldots, m\}$, x_i is the instance of the target domain and y_i is feature representation of the class label. $C = \{C_h | h = 1, 2, \ldots, N\}$ is the labels of T^i . There are a large number of unlabeled data as the same time. $T^u = \{x_j | j = m+1, m+2, \ldots, m+n\}$, m and n are the numbers of samples, and $m \ll n$.

In this paper, we will solve the problem that how to make use of online information to get a precise target classifier when there is few labeled short texts and no source data and no Priori probability distribution.

3.2 Construct the new features representation of the target data

As the target data have a few words, which can only provide a small amount of labels, the first we have to do is to expand the target label sets, called feature seed sets. The algorithm in our paper does not have to prepare source data in advance. In order to get auxiliary data, associated with the target areas, similar and useful for the target data, we take full advantage of online information, input keywords extracted from the target field to a search engine, and extract the first few pages of web as source data that semantically related to the target domain.

The next step is to identify the useful tags for the target task from the source domain. Traditional transfer learning methods often use bag of words model to represent the source and the target data. And then calculate the data similarity. According to the similarity, filter the most useful features from the source data. Although this method is simple, it views each word just as an individual, ignoring the relationship between the texts, especially the semantic relationships hidden in context keywords. However, the method of Latent semantic analysis organizes a text into a space semantic structure[12]. It assumes that there is a link, namely a potential semantic structure between the text and words. Use the Latent semantic analysis Statistical analysis of a lot of text sets, and get the context meaning of the words. And then extract the feature seed sets from the source domain.

First constructed the typical word - text matrix: $M = [a_{ij}]_{m \times n}$, a_{ij} is the logarithmic of the frequency that the i -th word appeared in the j -th text. Since each word appears only in a small amount of text, M is a high-level sparse matrix. Then apply the technology of singular value decomposition (SVD) and map words and texts from a high-dimensional space to a low-dimensional latent semantic space. Finally, we can get a new matrix $\tilde{M}$.

$$\tilde{M} = U \tilde{\Sigma} V^T \approx U \Sigma V^T = M \tag{1}$$

In the formula (1) ,U, V are orthogonal matrix. ($UU^T = VV^T = I$),$\Sigma = \text{diag}(a_1, a_2, \cdots, a_k, \ldots, a_v)$ ($a_1, a_2, \cdots, a_k, \ldots, a_v$ are the singular values of M)is a Diagonal matrix.

In the matrix $\tilde{M}$, the weight in the i-th row and j -th column represents the relevance of the word in the i-th row with text in the j-th column. Setting a threshold value λ . When the weight is greater than λ, the keyword can be a feature seed.

The second step of our work is to build a bridge between the labels. Social media is a website and technology that allows people to write, share, evaluate, discuss and communicate with each other. It provides tools and platforms for people to share opinions, insights, experiences and perspectives among themselves. Social media can be considered as a tag cloud, in which the co-occurrence labels carry a wealth of information. In the paper, social media becomes our auxiliary means, which can help build the bridge between the source labels and the target labels. First, each label is viewed as a node and connects the co-occurrence labels. So all the labels in the social media would appear in the graph. Then extract the subgraph that contains all feature seed sets from it. This can create a bridge between the labels in the source domain and the labels in the target domain. Finally, by improving the algorithm of Laplacian Eigenmapsmap, all the nodes in the subgraph can be mapped into a low-dimensional space. It effectively alleviates the problems such as data overfitting, low efficiency and so on, which are caused by the high-dimensional data. The next will improve the algorithm of Laplacian Eigenmapsmap so that it can comply with the requirements of our work.

The basic idea of Laplacian Eigenmapsmap is to represent the features of manifold local structure with Laplacian. High-dimensional manifolds composed by the n points$x_1, x_2, \ldots, x_n$, were given mapping in the D-dimensional space:

$$f : R^D \to R^d (d \ll D) \tag{2}$$

All the points are embedded in -dimensional space. Then in the d-dimensional space, we can get n points $y_1, y_2, \ldots, y_n$.

The Laplacian Eigenmapsmap algorithm [13] assumes that if the points are near in the high-dimensional space, the distances between them should be short when embedded into a low-dimensional space. As in the algorithm, it does not consider the classes information of the samples when calculate the neighbor distance. No matter the point inside the calss or outside the class, it gives the points same weight if the distances are the same. It is clearly unreasonable for target domain containing both labeled data and unlabeled data. In the paper, we improve the Laplacian Eigenmapsmap algorithm, using different methods to calculate the weight of labeled data and unlabeled data. And make points distance inside the class is less than the distance whose points are outside the class.

First construct the relative neighborhood graph. Use the method of unsupervised learning to calculate the distance between the unlabeled data. In the paper, we use Euclidean distance calculation method.

We use the method of supervised learning to calculate the distance between the labeled data. As follows:

$$D(x_i, x_j) = \begin{cases} \sqrt{1 - \exp(-d^2(x_i, x_j)/\beta} & c_i = c_j \\ \sqrt{\exp(d^2(x_i, x_j))/\beta} & c_i \neq c_j \end{cases} \quad (3)$$

In the formula (3), c_i and c_j are classes of the samples x_i and x_j. $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j. Parameter β can prevent $D(x_i, x_j)$ too large when $d(x_i, x_j)$ become larger. It can effectively control the noises. If the distance between sample points x_i and x_j is smaller than the threshold ε, the two points are neighbor points.

Then calculate the weight matrix $\mathbf{W}$. If x_i and x_j are neighbor points, $W_{ij} = 1$, else $W_{ij} = 0$.

Finally calculate the Laplacian generalized eigenvectors and embed them into low-dimensional space. The problem is to solve that:

$$\begin{cases} \min \sum_{i,j} \|Y_i - Y_j\| w_{ij} \\ s.t. \quad Y^T DY = I \end{cases} \quad (4)$$

In the formula (4), $\mathbf{D}$ is a diagonal matrix, . The formula (4) can be transformed into:

$$\begin{cases} \arg \min \mathrm{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y}) \\ s.t. \quad \mathbf{Y}^T \mathbf{D} \mathbf{Y} = I \end{cases} \quad (5)$$

In the formula(5), $\mathbf{L} = \mathbf{D} - \mathbf{W}$

With the improved Laplacian Eigenmapsmap algorithm, we can get a matrix $\mathbf{Y}$, each node can map to a low-dimensional space y. So each data can get a new feature representation.

3.3 Generate the target classifier

After getting the new feature representations of the target data, we can train a target classification with them. To complete the target task, we use the concept of mutual information. Mutual information measures how much a random variable contains another one. It can reflect the relationship between two statistics. The higher the degree of association become, the greater the mutual information is. Mutual information is defined as follows:

$$I(P;Q) = \sum_{x \in P} \sum_{y \in Q} p(x,y) \frac{p(x,y)}{p(x)p(y)} \quad (6)$$

Due to the target domain are short texts, containing few characteristics and some may still be useless or disturbance characteristics, in this paper we propose FSFP algorithm which can transfer usful features from the long texts to enrich the target characteristics. The

smaller of the mutual information between the new features representation and the label, the greater the degree of correlation of the data with the class. The likelihood of the data belonging to the class is greater. So the objective function is:

$$\min I_j(y_i, c_j) \quad (7)$$

In the formula (7), y_i represents the new feature representation of data, c_j represents the j-th class.

Algorithm 1 FSFP

Input the target data $T = T^l \cup T^u$ (T^l are labeled data, and T^u are unlabeled data), the labels C (the number of the labels id N), neighbor value k, feature seed threshold λ, parameter β, feature threshold ε.

Initialize $k, \lambda, \beta, \varepsilon$;

For $c = 1, \ldots, N$

1. Input the target labels to a search engine.
2. Extract the first 10 pages of data as the data that the most associated with the target areas.
3. Get the feature seed sets according to $\tilde{\mathbf{M}}$, k and λ.

$$\tilde{\mathbf{M}} = \mathbf{U} \tilde{\Sigma} \mathbf{V}^T \approx \mathbf{U} \Sigma \mathbf{V}^T = \mathbf{M}$$

4. Build the undirected graph of the social media.
5. Extract a subgraph that contains all these seed feature sets.
6. Filter the features of the target representation according to $\mathrm{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y})$, β and ε.

$$\begin{cases} \arg \min \mathrm{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y}) \\ s.t. \quad \mathbf{Y}^T \mathbf{D} \mathbf{Y} = I \end{cases}$$

Output the classification results of the target data according to I.

$$I(P;Q) = \sum_{x \in P} \sum_{y \in Q} p(x,y) \frac{p(x,y)}{p(x)p(y)}$$

---

4 Experimental evaluation

4.1 Data set

In order to make the results more authentic, we captured the actual data from the network as the experimental data sets in this paper. We crawled 300 pages randomly from C2C (cousumer to consumer), which was a shopping site, and it contained 97535 online advertising. From the the Sina Web, we randomly crawled 500 pages, in which there were 24047 weibo. We crawled 35184 QQ news randomly and in Taobao, we selected 58542 product reviews comprehensively. We used this online information as the target training data, among which the data with labels were only 5%.

Since our algorithm did not need to provide the source data, all the auxiliary data came from Internet. First of all, we extracted all the labels in the target domain as keywords and input them into a search engine (here we used Google). Then we selected the most semantically related keywords with the source domain. Since these keywords were used as the feature seeds, it might not be good if there were too many of them, or it might produce too many useless items in the process of these seeds expansion. Not only could it affect the efficiency of the experiment, but also affect the accuracy. In the paper, we selected the first pages as the most semantically related with the source data. Then built the undirected graph of the social media (here we used Delicious website). Extracted the subgraph according to the feature seed sets. So we extracted the most relevant characteristics from the short texts again, keeping the inherent structure of the source data. This was the second expansion of the data characteristics of the target domain.

4.2 Analysis of experimental results

We tested some possible factors that might affect FSFP algorithm firstly, and then set appropriate values for them. Finally, we compared the novel algorithm with other algorithms.

1) Parameter sensitivity test

There are two important parameters, λ and ε. When the number of the relevant pages is definite, the parameter of λ can decide the size of the seed feature sets. When we knew the features seed sets, the parameter of ε determined how many features we could filter from the social media. In this paper, we tested the sensitivity of the two parameters using online advertising for the experimental data. First of all, fixing the parameter of ε, we tested the experimental performance when λ was set up for 0.3, 0.5, 0.7, 0.9. The result was shown in Fig.1.
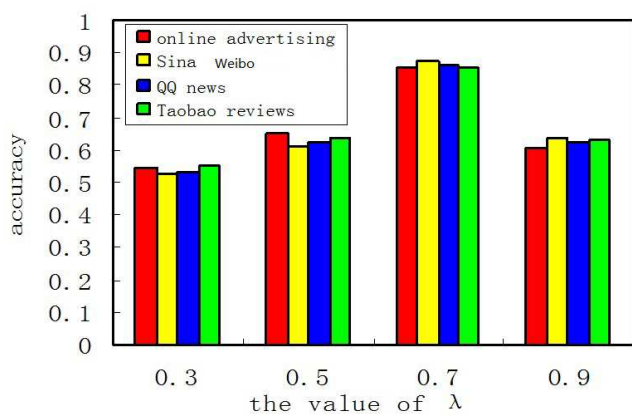

Fig.1 Sensitivity testing of λ

The value of λ and ε which was set too large or too small can affect the accuracy of the experiment. If the

value is set too small, that will make the features associated with the target domain filtered out.However, if they are set too big, some useless features would be mixed in, which may exert a noisy influence upon the final classification. From the experiment, we can see that when the value of λ is about 0.7, and ε is about 0.5, it can get a better result. That is shown in Fig.2. As λ decides the feature seed sets, it requires a higher relevance for the keywords; ε determines the final characteristics of the target data, to be set up a small value is beneficial for filtering possible useful features.

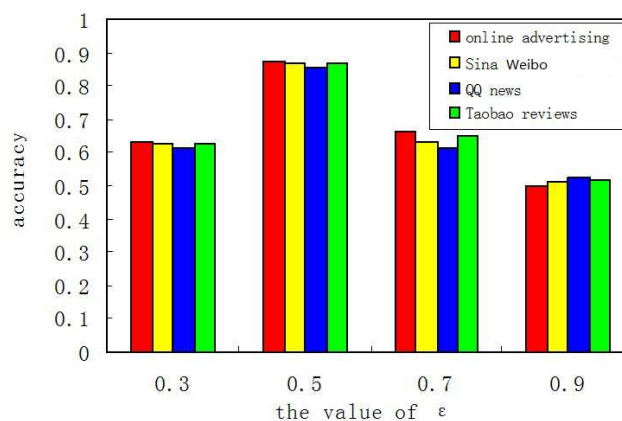Fig.2 Sensitivity testing of ε

2) The extraction of the relevant page in the search engine

As in this paper, there is no need to provide the source data, all the auxiliary data come from the online information. In the traditional transfer learning, the quantity of the source data must be much bigger than the target data. In our algorithm, there are two steps to extract auxiliary data. The size of feature seed sets affects the characteristics extraction of the social media tags. The effect of the size of feature seed sets or the number of related pages in the search engine are shown in Fig.3. The experimental results show that the fewer the relevant pages in the search engine are, the fewer feature seeds it contains. Further, the features selected in the social media are fewer which will reduce the final classification accuracy. However, as it is shown in Fig.3, it is not the best to select many relevant pages in the search engine. Filtering many useless feature seeds will introduce more useless feature items, which can affect the accuracy and efficiency of the algorithm. Through several tests, to extract about the first 10 semantically related webs, it can get a better experimental result.
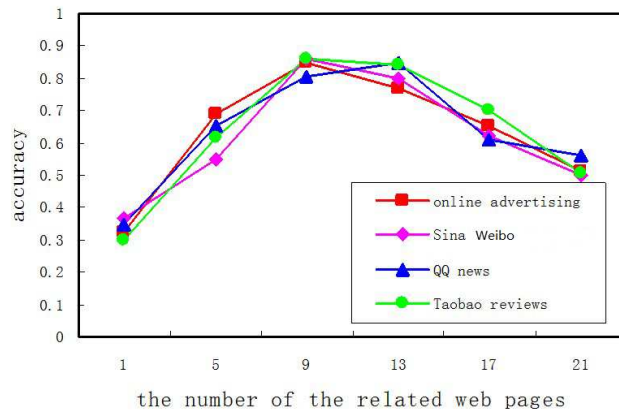
2038

W. Fengmei et. al. : FSFP: Transfer Learning From Long Texts to the Short...

Fig.3 The effect of the quantity of the relevant page extracted in the search engine

3) Compared Algorithms

In order to verify the effectiveness of the algorithm in the paper, we compared the FSFP with SVM which was a classification algorithm without transfer learning, Tr_SVM which was used to transfer knowledge from the long texts to the long, and DLDA that could transfer from the long texts to the short. In the algorithm of Tr_SVM, C and C_s were set 1.2. In the algorithm of DLDA $,\gamma_{tar}^{aux} = \gamma_{aux}^{tar} = \gamma_{small} = 0.2$ $\gamma_{tar}^{tar} = \gamma_{aux}^{aux} = \gamma_{big} = 0.5$. In the FSFP, k was set 7,λ was set 0.75, ε was set 0.5, β was set 2.

This experiment was completed in the environment of linux, using the C language. In order to make the experiment not affected by random factors and more authenticity, we did 10 times of the experiment and took the average of the experimental results. In the traditional transfer learning, when the target data was constant, the quantity of the data with labels can affect the algorithm. We compared the results of each algorithm by doing experiments on 4 data sets with different proportions of labeled data. Table.1 showed the accuracy changes of each algorithm when proportion of the labeled data increased from 5% to 15%. From the table, we could see the classification result of SVM without transfer learning is very poor. However, although Tr_SVM transferred knowledge from other domain, it ignored the differences between the long texts and the short, leading to not good result. The algorithms of DLDA and FSFP, took into account the different data structures of the two fields, the quality selecting useful features was higher than other algorithms. Their accuracy could be up to 70% -80%. In addition, compared with DLDA, FSFP did not to provide the source data and priori probability distribution of data before, so it reduced the influence caused by subjective factors. The feature seeds are the most semantically related to the target areas, and the expansion of the feature is the most relevant to the feature seeds. It improved the characteristics quality and increased the accuracy of the target classifier greatly.

The next experiment, doing on the data of online advertising, studied the accuracies of different algorithms, when the target labeled data is constant but the target data gradually increased. Suppose the quantity of the target data was M in the above experiments. In this experiment, we would test the result when target data change from $0.1 \times M$ to $1.5 \times M$. The experimental result was shown in Fig.4. From the figure, we can see that with the gradual increase of the target data, the accuracy of the algorithms increased. However, the classification accuracy of our algorithm proposed in the paper had always been better than the other algorithms. Especially when the target data were very few, the accuracy of FSFP is much higher than other algorithms.
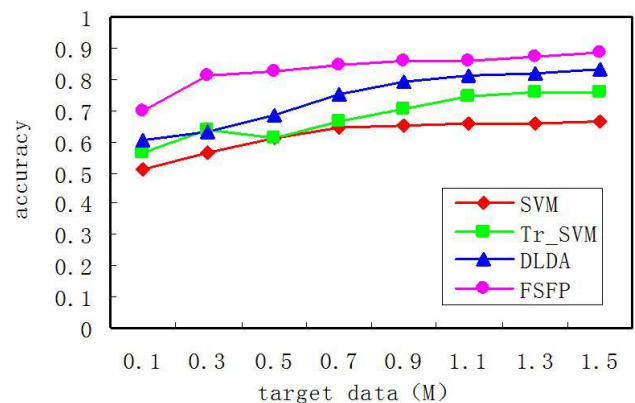

Fig.4 Performance comparison on different quantities of target data

5 Conclusions and future works

Transfer learning is a technique that finds useful knowledge and skills in the previous tasks and apply them to the new tasks or domains. In this paper, we solve the problem that transfer knowledge from the long texts to the short. Extract the most semantically related keywords as seed sets from the auxiliary long texts, which was selected online. Then the social media is used as a bridge connecting the source domain and the target domain. According to the feature seed sets, extract more related keywords from the long texts which can expand the features of the short texts again. As FSFP dose not have to prepare source data before and there is no need to know the priori probability distribution of data. So it reduced the effects caused by the subjective factors. And it can effectively improve the target classifier accuracy, taking into account both the semantics and the structure of the texts. The algorithm of FSFP is highly scalable. If the target domain is changed, the algorithm in the paper can also complete the target task. In the future, we will improve our algorithm to achieve more difficult transfer learning, for example transfer knowledge among images, texts, voice, video and so on.

Table.1 The experimental accuracy of the four algorithms(%)

Data Set	5				10				15			
	SVM	Tr_SVM	DLDA	FSFP	SVM	Tr_SVM	DLDA	FSFP	SVM	Tr_SVM	DLDA	FSFP
online advertising	50.2	61.4	70.3	81.2	55.3	66.2	75.7	86.2	60.1	70.3	80.7	88.3
Sina weibo	51.1	62.3	69.5	76.4	55.1	67.1	78.6	87.1	61.3	71.4	81.5	86.9
QQ news	49.8	65.1	72.4	78.1	52.4	60.7	80.1	88.6	64.5	70.1	81.9	90.1
Taobao Reviews	52.7	60.0	73.6	82.9	56.2	62.9	82.2	85.7	62.0	74.9	83.6	89.5

Acknowledgement

References

[1] S. J. Pan, Y. Qiang. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, **22**, 1345-1359 (2010).

[2] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno J. Pan, Gui-Rong Xue, Yong Yu and Qiang Yang. Heterogeneous Transfer Learning for Image Classification[C]. In Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, CA USA. August, 1304-1309 (2011).

[3] Sinno Jialin Pan and Qiang Yang. An Introduction to Transfer Learning[R].Hong Kong: Advanced Data Mining and Applications Conference, (2008).

[4] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen and Qiang Yang. Transfer Learning with Graph Co-Regularization[C]. In Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada, 22-26 (2012).

[5] Mei Canhua, Zhang Yuhong, Hu Xuegang, Li Peipei. Transfer learning algorithms based on maximum entropy model [J]. Computer research and development, **48**, 1722-1728 (2011).

[6] Hong Jiaming, Yin Jian, Huang Yun, Yiu Yubao, Wang Jiahai. TrSVM: A transfer learning method based on the similarity of domains [J]. Computer research and development, **48**, 1823-1830 (2011).

[7] Xiao Ling, Wenyuan Dai, et al. Can Chinese Web Pages be Classified with English Data Source[C]. In Proceedings of the 17th international conference on World Wide Web 2008, Beijing, China, 969-978 (2008).

[8] Guirong Xue, Wenyuan Dai, et al. Topic-bridged PLSA for Cross-Domain Text Classification[C]. In Proceedings of the 31st Annual International ACM SIGIR Conference. Singapore, 627-634 (2008).

[9] W. Dai, Q. Yang, G. Xue, et al. Boosting for transfer learning [C]. In Proceedings of The 24th Annual International Conference on Machine Learning, Corvallis, Oregon, USA, June 20-24, 193-200 (2007).

[10] Wenyuan Dai, Gui-Rong Xue, et al. Co-clustering based Classification for Out-of-domain Documents. In Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, Aug 12-15, 210-219 (2007).

[11] Ou Jin, Nathan Nan Liu, et al. Transferring Topical Knowledge from Auxiliary Long Text for Short Text Understanding[C].In Proceedings of the 20th ACM Conference on Information and Knowledge Management. Glasgow, UK, 775-784 (2011).

[12] Dumais S T. Using Latent Semantic Analysis to Improve Information Retrieval[C]. Proceedings of the ACM Conference on Human Factors in Computing Systems. Washington D. C., 281-285 (1988).

[13] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 1373-1396 (2003).

Wei Fengmei, Ph. D. student in Harbin Engineering University and Lecturer in the Dept. of computer, Harbin University, China. Her interests include data mining, machine learning and social network.

Zhang Jianpei, Ph. D., Professor in Harbin Engineering University. His interests include data mining and machine learning.

Chu Yan, Ph. D., Lecturer in Harbin Engineering University. Her interests include data management, mobile computing and machine learning.

Yang Jing, Ph. D., Professor in Harbin Engineering University. Her interests include data mining and Privacy.