Distance Based Active Learning for Domain Adaptation

Christian Pölitz

Fakultät für Informatik, LS VIII, Technische Universität Dortmund, 44221 Dortmund, Germany

Keywords: Transfer Learning, Active Learning.

Abstract: We investigate methods to apply Domain Adaptation coupled with Active Learning to reduce the number of labels needed to train a classifier. We assume to have a classification task on a given unlabelled set of documents and access to labels from different documents of other sets. The documents from the other sets come from different distributions. Our approach uses Domain Adaptation together with Active Learning to find a minimum number of labelled documents from the different sets to train a high quality classifier. We assume that documents from different sets that are close in a latent topic space can be used for a classification task on a given different set of documents.

1 INTRODUCTION

A large cost factor in using Machine Learning and Data Mining in Computer Linguistics and Natural Language Processing arises from the labelling of documents. For example, if we want to investigate the hypothesis that certain statements occur always in positive or negative context in a large set of documents. A usual approach would be to go through the documents and label parts of it as positive or negative and use them as examples for training a classifier. This can be quite expensive with respect to the documents and the task. The problem gets additional interesting in case we have labelled data sets from different domains. In our case a domain means that all documents follow a certain distribution p. In such situations we would like to use the the labelled data from all different domains even though they might have different distributions. We want to use the already labelled documents from domains (also called source domains) and adapt a trained classification model to a new so called target domain. This is a classical Domain Adaptation task as described by (BMP06) for instance. Further, we want to choose the documents from the different source domains that are best suited to train such classifier. The documents are actively sampled from all domains. Using these documents together with their labels the trained classifier should perform better on the new domain.

A possible application is to generate a classification model on some texts from Twitter, but we have only access to labels from documents from news articles. We might face words or whole sentences that are differently distributed in the domains. In such a case it is not clear how to transfer knowledge between the domains to use the news paper documents for training. Our idea to solve this problem is to use only news paper articles that are similar in a certain way to the Twitter tweets. These documents are transformed and together with their labels used to train a classifier that is applied to the tweets.

In this paper, we assume that we have access to labelled documents from some source domains. These domains share similarities in low dimensional subspaces with a given new target domain. For instance, tweets and news articles can talks about the same things but use different frequencies in wordings. A low dimensional latent space might cover the similarities between the domains based on co-occurring words. We want to find such a common latent subspace for the domains where semantically similar words are close. Latent Semantic Analysis (Dum04) for instance can be used to find a low dimensional latent subspace for a domain. Projecting the closest documents from the source domains to such a subspace, we train a classifier that is applied to the new target domain that is also projected into this subspace.

The paper is organized as the following: First, we explain how we can find low dimensional latent representations of documents in the different domains and what classifier we use in our training. Then, we describe how we use Domain Adaptation and Active Learning to train a classifier for a target domain using train data from different source domains. Based on distances from documents of the source domains to the common latent subspace, we identify most promising train data. Finally, we report results on our propose method on benchmark data sets.

Pölitz C..
Distance Based Active Learning for Domain Adaptation.
DOI: 10.5220/0005217302960303
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 296-303
ISBN: 978-989-758-076-5
Copyright © 2015 SCITEPRESS (Science and Technology Publications, Lda.)

2 RELATED WORK

We use methods from Active Learning, respectively Transfer Learning, and Domain Adaptation. Active Learning tries to manage the labelling process considering some intermediate results. A classifier that is trained on a small amount of labelled documents is used to estimate which further documents should be labelled to increase the quality of the classifier when trained also on these labelled documents. As candidates for further labelling we use the documents that are classified with least confidence. This strategy is called uncertainty sampling (LC94). Using a Support Vector Machine as classifier, the distance of a document to the margin is a proxy for the confidence in the prediction. In this paper we use this approach as proposed by Balcan et al. in (BBZ07) on many different source domains with label information. There are many different active sampling strategies in the literature. A general overview is given by (Set09).

Further, we assume that the documents are drawn from many different distributions respectively domains, but the labels have the same distribution given a document. In this case, instance weights can be used. In (JZ07), a classifier is trained on examples with labels and weights for each example. The weights are chosen such that the mass distribution of the examples from one domain adapts to the mass distributions of another domain. By this, they train a classifier using examples and labels from one domain that generalizes to another domain. A further approach is to model the commonalities of different domains as proposed by (BMP06) or (DM06) for instance.

Additionally, several approaches have been made to identify (latent) subspaces across different domains to identify common aspects in the domains. In (STG10) Si et al. propose to search for subspaces of two domains where the data distributions are similar in terms of Bregman divergence. In (SYvB^{+}11) Sugiyama et al. try to find low dimensional subspaces in which the domains differ. Only in this subspace a Domain Adaptation is necessary.

Recently, there have been efforts in combining Active Learning with Transfer Learning. Chan and Ng (CN07) couple Active Learning with Domain Adaptation for word sense disambiguation. They actively try to reduce the labelling cost for a target domain when they already have a trained model on a different source domain for word sense disambiguation there. In (SRD^{+}11), Saha et al. propose to train a classifier to distinguish a target form a source domain. In an Active Learning process they choose the most informative documents in the target domain using the classifier to decide if it belongs to the target or source domain. In case it belongs to the source domain a trained classifier on this domain can freely be used to label it. Further, Luo et al. use in (LJDC12) a similar approach to ours. They map the documents from a target and a source domain into a common latent factor space. In this space they train a classifier with actively chosen train documents. The difference to our approach is that the actively select samples from the target domains to label while we expect to have no access to any labels for the target domain.

Our approach is a combination of current Machine Learning methods to reduce the labelling cost. This means Transfer Learning, training of the classifier and the Active Learning strategies are coupled. In contrast to previous approaches, we assume that documents from each of the domains share the same support and have similar distributions on a low dimensional latent subspace.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

3 CLASSIFIER

To show the benefit of our proposed methods, we train a classifier on labelled documents that will be applied on new documents from a different distribution. We use a Support Vector Machine that has proven to be efficient in document classification, see (Joa02) for example. Given a set of documents with labels, we find a separating hyperplane in a Reproducing Kernel Hilbert space. In this paper we use the Bag-of-Words representations and embeddings into a latent subspace as training examples. In the Bag-of-Words approach, each document is mapped to a vector (a word vector) such that each component tells how many times a certain word occurs in the document. The embeddings are low dimensional vector representations of the documents that cover the most informative aspects.

During SVM training we minimize a regularized loss, formally $min_f \frac{1}{N}\sum_{i=1}^{N}[(1-y_i \cdot f(d_i))_+]+\lambda \cdot ||f||$ using the hinge loss $()_+$, y_i the labels and d_i the documents. Further, the classification results from a trained SVM are transformed into posterior probabilities $P(y|d_i)$ that can be used to estimate the confidence in the prediction of a document - see (Pla99) for instance. Later, we will use these confidence values for an Active Learning strategy.

4 DOMAIN ADAPTATION

For the applications as described in the introduction, we propose to combine Domain Adaptation techniques with Active Learning strategies. In Domain

Adaptation, we try to use documents from some different source domains to train a model that is used on a new target domain. In Active Learning, we try to find the documents that are potentially most helpful in training a classifier across the domains.

The main assumption in many Domain Adaptation papers is that the documents from the different domains may follow different distributions, but the conditional probability of a label given a document is the same over all domains. This is called the Covariate Shift assumption, see (SKM07) for more information. We further assume that this Covariate Shift assumption is only true on a low dimensional latent subspace.

We investigate two approaches for Domain Adaptation. First, we use Importance Sampling to adapt the source distribution to the target distribution. Assuming that the documents are differently distributed in different domains, we use an SVM with weighted examples as described below. The weights are estimated based on a regression model on the differences of the distributions of documents using Importance Sampling. Here, we assume that the Covariate Shift assumption is true on the whole document space.

Second and the main focus of this work, we assume that the Covariate Shift assumption is only true on a latent semantic subspace. We use large amounts of documents from different sources that already have been labelled to train a classifier that will be used on a new target domain with no label information.

While Importance Sampling weights the documents such that these weights reflect the adaptation to a different domain, we sample documents from the source domain proportional to the distance of the documents to the target domain.

This strategy makes sense under the assumption that documents, that are close and hence similar in some latent factor subspace, have the same label. We can define such a similarity as how much distance a document from a certain domain has to the target domain represented as low dimensional latent subspace. Later, we explore the benefit of performing the Domain Adaptation and the training of the classifier in a common latent factor space together with an Active Learning strategy.

In the next two subsection, we describe how Importance Sampling and Latent Subspace Methods can be used to optimally use the documents from source domains for the training of a classifier that is applied to documents from a new target domain with a different distribution.

4.1 Importance Sampling and Density Ratio Estimation

Under the Covariate Shift assumption on whole document space, Importance Sampling and Density Ratio Estimation can be used to adapt training data from a different source domain for training classifier on a target domain with new distribution.

If p_s and p_t are the document distributions from a source domain s and the target domain t with the same support, we can estimate the expected loss under the target domain using documents from source domain, using Importance Sampling. In Importance Sampling we sample from p_s, hence use documents from a source domain, but weight the examples by $r(d)$ for the documents d such that $r(d) \cdot p_s(d)$ has approximately the distribution p_t of the target domain. For further reading, we refer to (OZ00). $r(d)$ is called density ratio or weight function depending on the context. These weights are integrated into the risk minimization framework for the SVM using the hinge loss L. This means, we solve the following minimization problem: $min_f \frac{1}{N} \sum_{i=1}^{N} r(d_i) \cdot [(1 - y_i \cdot f(d_i))_+] + \lambda \cdot ||f||$ See (LLW02) for further details.

In practice the density ratio $r(d)$ is estimated as regression model as proposed by (YSK$^+$13) for instance. A major shortcoming on this approach is that the density ratio estimation can only be applied between two domains. But in our case we want to use data from different domains, that might have all different distributions. In theory, the Importance Sampling will produce samples from the target distribution using the source distribution. But this is only true for a single source distribution. In the next subsection, we propose to use latent subspaces in order to find common substructures among different domains to train the classifier across many source domains.

4.2 Latent Subspace Methods

Under the Covariate Shift assumption on a latent subspace, we expect that documents that are close to a latent subspace from a new target domain can adapt to the new target domain by projecting them onto this subspace. We can train a classifier on these projected documents. We expect that this classifier can be safely applied to documents from the target domain.

We use latent subspace methods to extract the most important parts of the documents of a target domain. Based on this subspace, we can estimate distances from the documents of the source domains to the target domain. Using the Covariate Shift assumption we expect that documents that are close to the latent subspace from the target domain are similar

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

enough to use them for the training of a classifier that is applied only in a different target domain.

There are different possibilities to model latent subspaces in the document domains. The overall goal is to make the different source domains more similar to the target domain by mapping the documents into a corresponding latent subspaces. This means, the subspace shall keep invariant parts of source and target domains. By this, a trained classifier on the source domains can also be applied the target domain.

Assuming we have extracted a subspace S from the documents of the target domain, we define the distance δ of a document d to this subspace as a proxy for the closeness and hence similarity to the target domain. The estimation of the distance depends on the subspace and how we represent the documents. We investigate two possible representations and corresponding subspaces that represent core aspects of the documents.

First, we use the Bag-of-Words approach to represent a document as word vector. Latent Semantic Analysis (LSA) (Dum04) is used to extract a latent subspace from the term document matrix D that is build by all word vectors. In LSA, we perform a singular value decomposition on the matrix D such that $D = L^\top \cdot E \cdot R$. L is a basis in the space spanned by the terms and R is a basis of the space spanned by the documents. E is a diagonal matrix containing the singular values of D. The projection onto the latent subspace in the space of the documents that corresponds to the largest k singular values is noted as R_k. By this any document d, represented as Bag-of-Words, can directly be projection onto this space by $R_k \cdot R_k^\top \cdot d$.

Second, we represent the documents as sequence of words, respectively tokens of word and documents ids. This means any document is a sequence of term ids. We use Latent Dirichlet Allocation (LDA) (BNJ03) to extract a latent subspace from the documents. LDA models the documents as random mixture model of a number of topics. The topic distribution follows a Dirichlet distributions. The parameter estimation is done via Gibbs sampling as proposed by Griffiths et al. in (GS04).

To map a document into the subspace extracted by LDA, we embed it into the simplex spanned by the posterior distributions of the latent factors given a document. This means we map d to $[p(t_1|d), \cdots, p(t_k|d)]$. Since we use a Gibbs sampler for LDA we can simulate the process of assigning a document to a topic and hence can estimate the posterior probability simply as: $p(t_i|d) = \frac{n_i(d)+\alpha}{n(d)+k\cdot\alpha}$. $n_i(d)$ is the number of times topic i is assigned to document d in the simulation, $n(d)$ the number of times document d is assigned to any topic, k is the number of topics and α is the meta parameter from the LDA procedure.

4.3 Distances

An important measure for our proposed approach is the distance of a given document to a latent subspace. For LSA, the length of the orthogonal projection onto the latent subspace is the distance from a document represented as word vector to the latent subspace extracted by the LSA. Formally, we note the distance as: $\delta(d_i, S_R) = \left\|R_k \cdot R_k^\top \cdot d_i\right\|^2$. The length can be calculated for each document regardless of the domain it comes from. This is possible since all document are modelled by the Bag-of-Words approach and are represented as elements of the same vector space.

The distance from a document to the latent space extracted by LDA is not as straight forward as for LSA. Since LDA extracts a latent space over probability distributions a natural distance is the KullbackLeibler divergence (KL51). Formally we note: $\delta(d, S_D) = \sum_t log(\frac{p(t|d)}{p(t|D)}) \cdot p(t|d)$. We extract the posterior distributions $p(t|d)$ and the topic distribution for the whole data set $p(t|D)$ via the Gibbs sampler.

Here, an additional problem arises when we want to calculate the distance of documents from different domains to a domain that is modelled by a the latent subspace extracted by LDA. Similar to the Euclidean case we need to map the documents into the subspace. We need the Gibbs sampler again to estimate the posterior distributions of the latent topics for the new documents. In a simulation, the Gibbs sampler is applied only to the new documents, while keeping the posterior distributions for all other documents fixed. This embeds a new document d_n from the source domains into the corresponding subspace of the target domain via $[p(t_1|d_n), \cdots, p(t_k|d_n)]$.

In the next section, we explain how we use the latent subspace model in an Active Learning scenario to reduce labelling cost for the case when we have an unlabelled target domain and many labelled source domains from different document distributions.

5 ACTIVE LEARNING ACROSS DIFFERENT DOMAINS

In this section we describe how we use Active Learning together with Domain Adaptation in order to reduce number of labelled documents needed from the different domains in a classification task. We generally assume that the distribution of the documents differ among the different domains. Formally this means $p_i(d) \neq p_j(d)$, for two different domains i and j and a document d. Further, we assume that the distributions of the labels for a given document are the same among the domains on a latent subspace S using a projection matrix P_S, hence $p_i(y|P_S \cdot d) = p_j(y|P_S \cdot d)$.

After an initial training of an SVM on the nearest documents from the source domains to the target domain, we apply the model to all documents from the source domains. For these documents we estimate a confidence value for the prediction based on the probabilistic outputs as described above. Based on this value we choose those documents that result in a low confidence in the prediction for retraining our SVM model.

Since now we have different domains, we expect that not all domains or all documents are similar useful for the retraining. Even within one domain we expect some documents to be more useful for the training than others. That is why we include the distance measure into our Active Learning strategy.

We use two criteria for choosing the documents from the different source domains for the training. First, as in confidence based Active Learning, we estimate a confidence value for the prediction of a trained model on the documents from the source domains projected onto the latent space of the target domain. Next, we integrate the distance of the documents from the source domains to the latent space to estimate their potential latent value for the training when we apply the SVM on the target domain.

The equation $\sigma(d) = (\lambda \cdot \gamma(f,d) + (1-\lambda) \cdot \delta(d,D_t))^{-1}$ defines the selection faction as the inverse of the weighted sum of the confidence value γ and the distance δ of the corresponding document to the target domain. The larger this value is, the more similar is the document to the target domain while on the other hand the SVM is uncertain in its prediction of this document. Among all documents from the domains, except the target domain, the closest ones that are predicted with least confidence are chosen for retraining.

6 EXPERIMENTS

In this section we perform extensive experiments on benchmark data sets to validate our proposed method. In the first experiment we test how good the latent subspace representations, respectively the embeddings into a latent subspace of documents from the different domains, can be used for training a classifier. Using the vector space model, each documents is represented by a large vector with each telling component the number of times a certain word appears in the document. LSA is used to find a latent subspace that covers the most important parts of the documents.

Table 1: Accuracy on separating documents about organizations from documents about people, respectively places. We use only data from the source subcategories for the training. The Baseline is an SVM trained on the whole vector space of the source domain. LSA means the SVM is trained on the projection of the source domain onto the subspace extracted from the target domain by LSA.

Data sets	Baseline	LSA
Org−People	80.3	82
Org−Places	61.7	70.8

We use the Reuters data set in the same configuration as done by (DXYY07). The documents are about organizations, people and place. The task is to distinguish the documents about organizations from the documents about people, respectively places. The training is done on a subset of subcategories and the testing on different subcategories.

Table 1 shows the results of an SVM classifier trained on a source domain of the subcategories that contains only documents about organizations. For comparison we use a simple baseline method. This method uses no latent representation, but simply the Bag-of-Words representation. The performance of the SVM in these subspaces is much better compared to the baseline. This shows already a potential benefit of a projection onto a latent subspace to make the domains more similar.

Next, we validate our Active Learning strategy. We train the SVM on initial 300 documents from the source domains that are closest to the latent subspace extracted from the target domain. Next, we applied the SVM to all remaining documents in the source domains and calculate the selection factor s from Equation **??**. The documents with highest selection factors are chosen to be used for the retraining of the SVM.

Figure 1 show the increase in accuracy of a trained SVM using our proposed Active Learning strategy. We see that already after 600 respectively 900 documents we get better results as the baseline that has been trained on the whole source data set.

In our second experiment, we investigate how good our proposed methods performs when we have more than one source domain. Here, we expect that some domains might be better suited for the training than other ones. Beside the Bag-of-Words representation together with LSA, we also tested the representation as sequence of words together with LDA. As distance measure we used for the Bag-of-Words representation again the Euclidean distance and for the representation as sequence of words the KL-divergence.

We used the Amazon review data set in the same configuration as Blitzer et al. in (BMP06). The review documents are about books (B), Dvds (D), elec-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

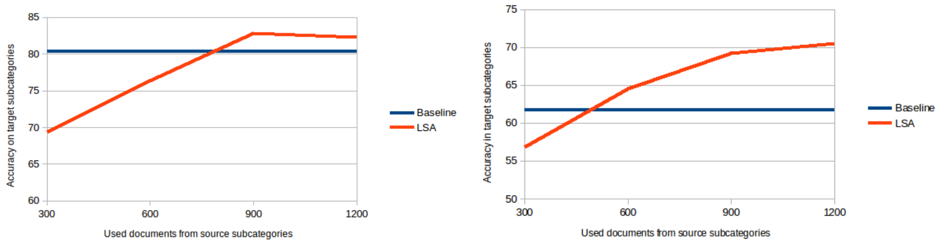
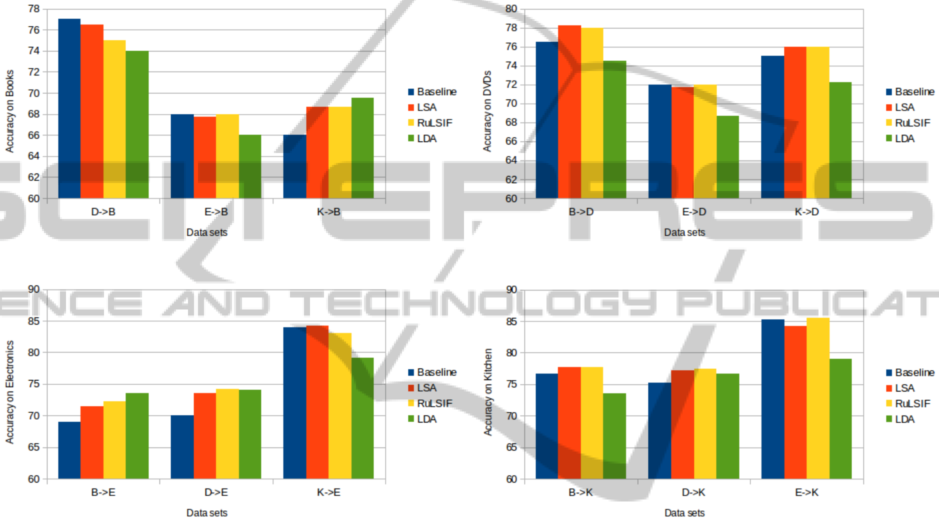


Figure 1: Accuracy on the task to separate documents about organizations from the documents about people (on the left) respectively places (on the right(using our Active Learning strategy. The subspaces are extracted via LSA. The Baseline is an SVM on the whole vector space and uses all the documents.



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 2: Results on the target domains. For each source domain a classifier is trained and applied to a target domain: source domain $- >$ target domain. The baseline is an SVM trained on the whole source domain. LSA and LDA are the latent subspace methods and RuLSIF is the domain adaptation method via importance weight.

tronics (E) and kitchens (K). One of these domains was always being used as target domain without considering labels. The other domains are used as source domains for the training. The sets of documents from each domain were split into a train set of 1600 and a test set of 400 documents. In the Active Learning scenario we used 1600 documents from all domains except from the target domain for possible training. As baseline we trained an SVM directly on the source domain in its original Bag-of-Words representation. Compared to this, we tested the Bag-of-Words representation together with LSA and the sequence of tokens representation together with LDA. Further we tested Importance Sampling for domain adaptation on the Euclidean subspaces. We applied a method called RuLSIF as introduced by (YSK$^+$13) to estimate the importance weights for the documents as discussed above.

We tested how good the different source domains can be used for training the SVM that is applied on a different target domain. Figure 2 shows the accuracy on the different domains when the SVM model is trained purely on one of the other domains.

The main result is that there is always one domain that is best suited for the target domain. Since we might have no information about the domains or even the possible best domain, the projection on the subspace can increase the accuracy even on the worst suited domains.

Next, we investigated our distance assumption by training the SVM model on the closest documents from the source domains to a corresponding target domain. We use the same number of train documents as before. Table 2 shows the accuracies on the target domains. For a given target domain, we chose the 1600 closest documents from the other domains for training and applied the trained SVM model on the test sample of the target domain. The accuracies are between the best and the second best results of the subspace method on only one domain. This is what we expected. With no domain information this is the best we can get. This means, that our closeness measure is

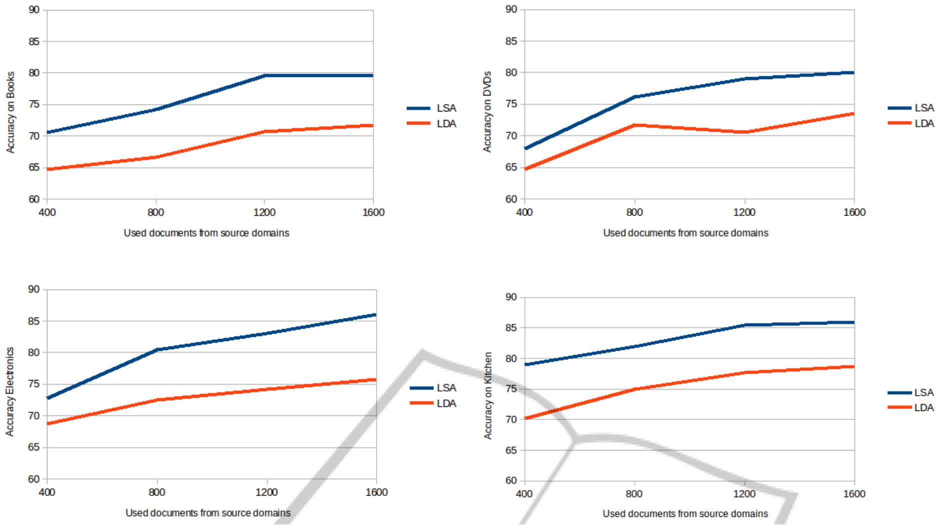


Figure 3: Results on the target data sets with our proposed Active Learning strategy.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 2: Accuracy on the different target domains. For training we used the documents from the source domains that are closest to given target domain.

	Books	DVDs	Electronics	Kitchen
LSA	75	75.7	80	83.5
LDA	69.25	70	75.7	80.2

a good indicator for which documents to be used for the training.

Finally we tested our proposed Active Learning strategy among all domains for a given target domain. Figure 3 shows the accuracies when we apply our proposed Active Learning strategy. Similar to the experiments before, LSA outperforms LDA. Further, we see that already after two third of the available documents are used for the training, we reach a higher level of accuracy compared to the experiments before.

7 CONCLUSION AND FUTURE WORK

In this paper, we explained an approach to perform Active Learning across different domains using Transfer Learning. We argued that the distance of documents is a good measure of how appropriated documents from different domains are for the training of a classifier for a certain target domain. We calculated the distance of documents to (different) domains as distance to a latent subspace of the corresponding target domain. Finally, we defined an Active Learning strategy that integrates this distance measure to choose potentially useful documents from many different domains for the training of an SVM that is applied to a target domain where no label information are available. The results on benchmark data sets show the potential of our proposed methods. Compared to previous approaches we are now able to easily use large amounts of documents from different domains for training of any other domain.

REFERENCES

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, pages 35–50, Berlin, Heidelberg, 2007. Springer-Verlag.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

Yee Seng Chan and Hwee Tou Ng. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Hal Daumé, III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126, May 2006.

Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 210–219, New York, NY, USA, 2007. ACM.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

Jing Jiang and Chengxiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the Association for Computational Linguistics*, ACL'07, pages 264–271, 2007.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, ICML'94, pages 148–156. Morgan Kaufmann, 1994.

Chunyong Luo, Yangsheng Ji, Xinyu Dai, and Jiajun Chen. Active learning with transfer learning. In *Proceedings of ACL 2012 Student Research Workshop*, pages 13–18, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Yi Lin, Yoonkyung Lee, and Grace Wahba. Support vector machines for classification in nonstandard situations. *Journal Machine Learning*, 46(1-3):191–202, March 2002.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):pp. 135–143, 2000.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

Burr Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, December 2007.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L. DuVall. Active supervised domain adaptation. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part III*, ECML PKDD'11, pages 97–112, Berlin, Heidelberg, 2011. Springer-Verlag.

Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.

Masashi Sugiyama, Makoto Yamada, Paul von Bnau, Taiji Suzuki, Takafumi Kanamori, and Motoaki Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183 – 198, 2011.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS