International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 14871-14876

http://www.ripublication.com

Cross Domain World Wide Knowledge For Sql Database By Using Transfer Learning

[1]G.Mohana Prabha and [2]Dr.S.Chitra

[1]Assistant Professor/CSE, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India
Email: *sekarprabha@gmail.com*
Ph: 91-9865175910
[2]Principal, Er.Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India
Email: *schitra3@gmail.com*

Abstract

The Classification learning technique has a major drawback in classifying both labeled and unlabeled data. In case of supervised learning technique, when there is a lack of labeled data instant the classification performance is reduced. It is difficult to train a system using both labeled and unlabeled data instance in supervised learning. To solve this problem, semi supervised learning to train both labeled and unlabeled data have been proposed. The semi supervised learning (SSL) is a machine learning technique to train the artificial intelligent system in such a way that it performs classification with both labeled and unlabeled data.. On the other side, it is very difficult to train a model when the data distribution is different, it also reduces the classification performance. To solve this problem domain adaptation techniques in transfer learning are introduced by capturing the shared knowledge from some related domains where labeled data are available, and use the knowledge to improve the performance of data mining tasks. A cross domain platform using semi-supervised learning (SSL) bridges web client and SQL server thereby providing effective information retrieval. Whatever the database operation done in web client is reflected over the SQL database server.

1. Introduction

Text classification, which aims to assign a document to one or more categories based on its content, is a fundamental task for Web and document data mining applications, ranging from information retrieval, spam detection, to online advertisement and Web search.

Traditional supervised learning approaches for text classification require sufficient. labeled instances in a problem domain in order to train a high quality model. However, it is not always easy or feasible to obtain new labeled data in a domain of interest (hereafter, referred to as the target domain). The lack of labeled data problem can seriously hurt classification performance in many real world applications. To solve this problem, transfer learning techniques, in particular domain adaptation techniques in transfer learning is introduced by capturing the shared knowledge from some related domains where labeled data are available, and use the knowledge to improve the performance of data mining tasks in a target domain. In transfer learning terminologies, one or more auxiliary domains are identified as the source of knowledge transfer, and the domain of interest is known as the target domain. However, transfer learning may not work well when the difference between the source and target domains is large. In particular, when the distribution gap between the source and target domains is large, transfer learning can hardly be used to benefit learning in the target domain. For example, when we use some financial documents as the source domain and information technology documents as the target domain, the differences are so large that the performance in the target domain may decrease. Another problem is when the source and target domains have a large divergence in feature space; for example, the source data might be written for one audience and the target data for another. In these situations, traditional transfer learning might not work well.

2. Semi Supervised Learning

Domain adaptation could also be viewed as transductive transfer learning if the source domain and the target domain had no information gap. In this case, the problem can be reduced to a semi supervised learning problem. However, when there is information gap, how to exploit semisupervised learning is not clear. In this section, we first review some semi supervised learning research works. Semi supervised learning addresses the problem when the labeled data are too few to build a good classifier and makes use of a large amount of unlabeled data, together with a small amount of labeled data to enhance the classifiers. Many semisupervised learning algorithms have been developed in the past 10 years. Some notable models developed include co-training, transductive SVM, and some other graph-based regularization methods like mincut

In traditional supervised learning problems, we are given a set of labeled instances from a specific domain as a training set. A learning machine is trained on the training set and will be applied on newly incoming instances from the same domain to obtain their labels. The condition that training and test sets are drawn from the same domain guarantees the consistence and generalization ability of the learning machine. However, in practice we may not be able to ensure that the training and test sets are from the same domain.

3. Domain Adaptation

Domain adaptation has attracted more and more attention in the recent years. In general, previous domain adaptation approaches can be classified into two categories are as previously defined one is instance based approach another one is feature based approach. Instance-based methods try to seek some reweighting strategies on the source data, such that the source distribution can match the target distribution. Feature-based methods try to discover a shared feature space on which the distributions of different domains are pulled closer. Both types are trying to discover the relation between source and target domains within the scope of two domains. For example, instance-based transfer learning models assume that there is a subset of instances sharing similar distributions in different domains, and then they emphasize the impact of these data in the models since they are more "similar." For the feature-based domain adaptation models, they assume that different domains may share some features, for instance, a subset of explicit features or implicit features.

Data Mining with Online Knowledge Repository is a major component of our approach is to use online knowledge repositories as auxiliary information sources to help bridge the gap between the source domain and the target domain. Therefore, we review some latest approaches of data mining with online knowledge repositories. In recent years, understanding and using online knowledge a repository to aid real world data mining tasks has become a hot research topic. There are more and more works trying to use the Wikipedia for feature enrichment. Unlabeled data in co-training help to reduce the size of the version space. Transductive builds the connection between class distribution and decision boundary by putting the boundary in low density regions.

The goal is to find a labeling of the unlabeled data such that a linear boundary has the maximum margin on both the original labeled data and the unlabeled data. It can be viewed as an SVM with additional regularization term on unlabeled data. Graph-based semi supervised methods define a graph where the nodes are labeled and unlabeled examples and edges reflect the similarities between examples.

Blum and Chawla view semi supervised learning as a graph min cut problem, where positive labels act as sources and negative labels act as sinks and the mincut problem aims to find a minimum set of edges whose removal would block all flow from sources to sinks. Nodes connected to the sources would be labeled as positive and those connected to the sinks would be labeled as negative. In this work, we will exploit the ability of semi supervised learning to aid the problem of domain adaptation.

4. Feature Space

In traditional supervised learning problems, we are given a set of labeled instances from a specific domain as a training set. A learning machine is trained on the training set and will be applied on newly incoming instances from the same domain to obtain their labels. The condition that training and test sets are drawn from the same domain guarantees the consistence and generalization ability of the learning machine . However, in practice we may not be able to ensure that the training and test sets are

from the same domain. We want to use it to classify reviews from other domains such as books or music . Another example is that we may have trained a classifier to classify news into topical categories but we want to use it also on blogs. In these cases, we do not want to re-label the data in the new domains but hope to borrow the knowledge from the old domains.

When the differences between the source and target domains are large, the model trained on the source domains cannot generalize well for the target domain data A natural approach to follow is to consider transductive (or semisupervised) learning, since unlabeled data from the target domain is available. However, some previous works have found that after introducing some unlabeled data in the target domain, transductive learning is still not sufficient in improving the performance. The reason may be that transductive or semisupervised learning generally assumes that the decision boundary lies in the low-density region of the feature space.2 When the distributions of source and target domains are different, there may exist a low density region between different domains which is a gap that disconnects the same-class data in different domains. We refer to this gap as the information gap in domain adaptation. For example, Fig. 1 illustrates an example where the feature space is the Wikipedia concept space. From Fig. 1, we could see that there exist some information gaps between the source domain and target domains. To solve the problem of domain adaptation under large information gaps, an intuitive idea is to find the shared part of different knowledge between the domains, and ignore the differences. One instantiation of this idea is to make use of the abundant and potentially useful information sources that are around, and use them to connect the information separated by the gap. Such an intuition motivates us to think of a different way for solving the domain adaption problem, i.e., through finding an information bridge.

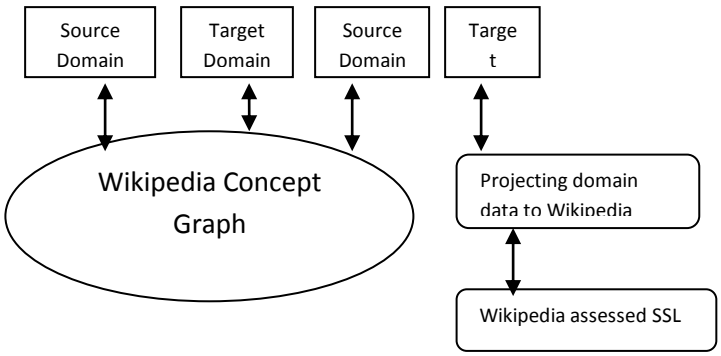

Fig. 1. Measuring similarities between documents based on the knowledge base.

5. Margin as Information Gap

An intuitive way to understand the concept of information gap is to consider reparability of the source and target domains. Consider the simplest case when we want to transfer knowledge from a single source domain to a target domain. Intuitively, the difficulty in separating these domains shows how large the information gap is between them. If the two domains can be easily separated, then there exists a

large information gap between them, which may prevent our adapting the original learned model from the source to the target domain. On the contrary, if the two domains cannot be separated from each other easily, then the information gap is small, in which case we can treat the two domains as essentially data that are sampled from a single underlying distribution. In other words, the original "domain adaptation problem" is transformed into a classification problem under the supervised setting or a semi supervised (transductive) setting.

A similar idea is used where a classifier is trained to distinguish the source and target domains and the classification error is used as an empirical estimation for domain distance.

6. Convergence and Stability

The paper first demonstrates that our algorithm can reduce the information gap between domains during the process of including the unlabeled data from the related domains. We randomly sample three tasks for each of the three data sets and display the performance together with their corresponding margin sizes. For each iteration, we include the top 100 unlabeled data that are closest to the decision boundary for TSVM. The x-axis is the iteration count. We find that our algorithm is able to reduce the information gap and converge quickly..A very interesting observation is that our optimal margin threshold is relatively stable over different data sets.

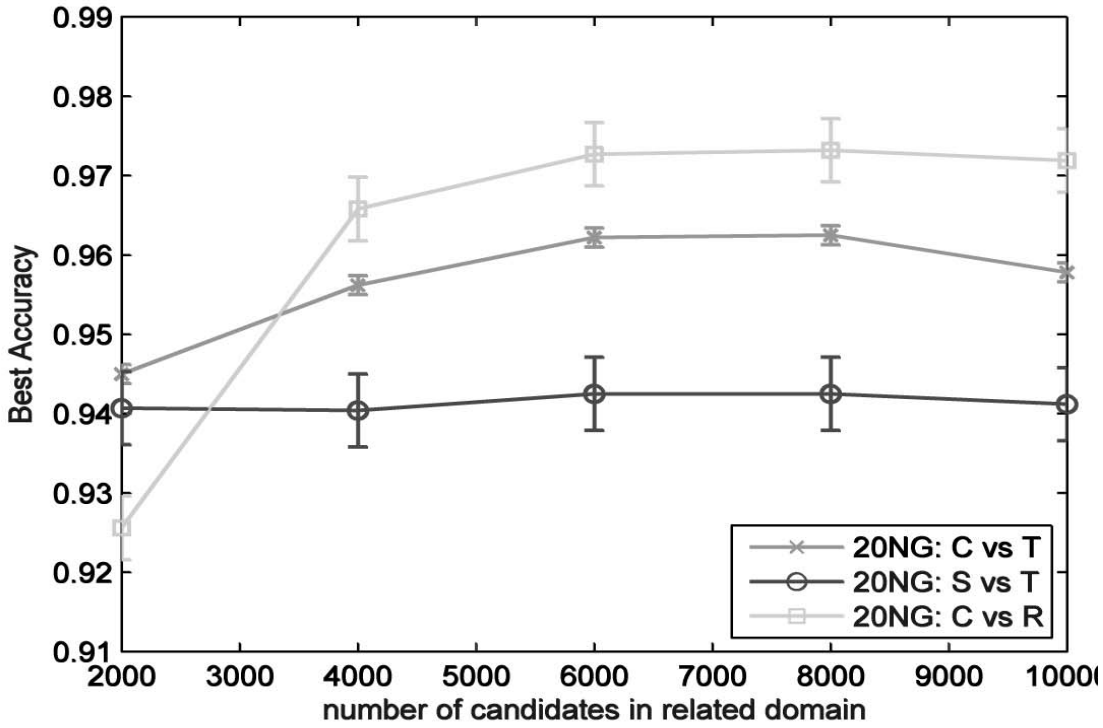

Fig 2 The relation between the number of auxiliary data and the best performance

7. Conclusion and Future Works

In this paper, we model the learning problem as a semi supervised learning problem aided by a method for filling in the information gap between the source and target domains with the help of an auxiliary knowledge base (such as the Wikipedia). By conducting experiments on different difficult domain adaptation tasks, we show that our algorithm can significantly outperform several existing domain adaptation approaches in situations when the source and target domains are far from each other. In each case, an auxiliary domain can be used to fill in the information gap efficiently.

8. References

[1] S.I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks," Proc. 24th Ann. Int'l Conf. Machine Learning (ICML '07), pp. 489-496, June 2007.

[2] J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," Proc. 45th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '07), June 2007.

[3] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu, "Topic-Bridged PLSA for Cross-Domain Text Classification," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 627-634, July 2008.

[4] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying, "A Spectral Regularization Framework for Multi-Task Structure Learning," Proc. 21st Ann. Conf. Neural Information Processing Systems (NIPS '07), Dec. 2007.

[5] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, "Self-Taught Learning: Transfer Learning from Unlabeled Data," Proc. 24th Ann. Int'l Conf. Machine Learning (ICML '07), pp. 759-766, June 2007.

[6] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning Bounds for Domain Adaptation," Proc. 21st Ann. Conf. Neural Information Processing Systems (NIPS '07), Dec. 2007.

[7] S.J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Trans. Knowledge and Data Eng., preprint, 12 Oct. 2009, doi: 10.1109/TKDE.2009.191.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of Representations for Domain Adaptation," Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 137-144, Dec.2006.

[9] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-Clustering Based Classification for Out-of-Domain Documents," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 210-219, Aug. 2007.

[10] H.D. III, "Frustratingly Easy Domain Adaptation," Proc. 45th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '07), June 2007.