Multi-Arm Active Transfer Learning for Telugu Sentiment Analysis

Subba Reddy Oota[1], Vijaysaradhi Indurthi[1], Mounika Marreddy[2]
Sandeep Sricharan Mukku[1], and Radhika Mamidi[1]

[1] International Institute of Information Technology, Hyderabad,
[2] Quadratyx, Hyderabad.
ootа.subba@students.iiit.ac.in,vijaya.saradhi@research.iiit.ac.in
mounika0559@gmail.com,sandeep.mukku@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract. Transfer learning algorithms can be used when sufficient amount of training data is available in the source domain and limited training data is available in the target domain. The transfer of knowledge from one domain to another requires similarity between two domains. In many resource-poor languages, it is rare to find labeled training data in both the source and target domains. Active learning algorithms, which query more labels from an oracle, can be used effectively in training the source domain when an oracle is available in the source domain but not available in the target domain. Active learning strategies are subjective as they are designed by humans. It can be time consuming to design a strategy and it can vary from one human to other. To tackle all these problems, we design a learning algorithm that connects transfer learning and active learning with the well-known multi-armed bandit problem by querying the most valuable information from the source domain.
The advantage of our method is that we get the best active query selection using active learning with multi arm and distribution matching between two domains in conjunction with transfer learning. The effectiveness of the proposed method is validated by running experiments on three Telugu language domain-specific datasets for sentiment analysis.

Keywords: Active Learning, Transfer Learning, Multi-Arm Bandit

1 Introduction

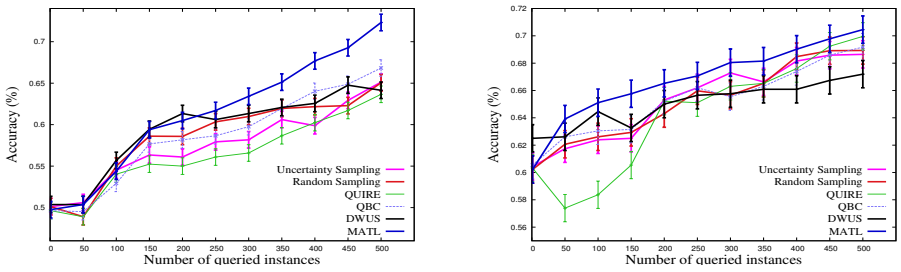
People comment on online reviews and blog posts in social media about trending activities in their regional languages. There are many tools, resources and corpora available to analyze these activities for English language. However, not many tools and resources are available to analyze these activities in resource poor languages like Telugu. With the dearth of sufficient annotated sentiment data in the Telugu language, we need to increase the existing available labeled datasets in different domains. However, annotating abundant unlabeled data manually is very time-consuming, cost-ineffective, and resource-intensive.

To address the above problems, we propose a Multi-Arm Active Transfer Learning (MATL) algorithm, which involves transfer learning [1] and a combination of query selection strategies in active learning [3]. One of the prerequisites

for transfer learning is that the source and target domains should be closely related. We use Maximum Mean Discrepancy (MMD) [2] as a measure to find the closeness between two distributions of the source and target domains. In this paper, we experiment with sentiment analysis of Telugu language domain specific datasets: Movies, Political and Sports[1]. By considering each domain as the source or target domain, we have a total of 6 domain pairs: M-P, M-S, P-M, P-S, S-M, S-P. Figure 1 shows two domain pair results. We evaluate the accuracy with three different classification techniques viz., support vector machines (SVM), extreme gradient boosting (XGBoost), gradient boosted trees (GBT), and meta learning of all these approaches and record the accuracy.

2 Approach & Results

In Multi-Arm active transfer learning approach, it takes both source domain: S = {unlabeled data instances (S_U), labeled data instances (S_L)}, and target domain: T = {unlabeled data instances (T_U), labeled data instances (T_L), test data instances (T_T) (used for measuring classification accuracy at each iteration)}, iterations (n) as an input. A decision making model is built along with this approach to predict the posterior probability for each instance of S_U. After calculating the sampling query distribution $\phi(S(n))$, based on multi-arm bandit approach a best sample instance $x_{i_n} \in S$ is selected for querying. If $x_{i_n} \in S_U$, then this selected sample instance (x_{i_n}) is labeled with an oracle/labeler as y_{i_n} and added to S_L. Now the classifier (C_n) is trained on the total set {updated S_L,T_L}. Using MMD [2], the distance between two distributions is calculated. This process is repeated until reached query budget. The classification model C_n is tested on target test data T_T to measure the accuracy. The reward ($r_n(a_k(n))$) and observation($o_n(a_k(n))$) is updated by comparing the label y_{i_n} given by the oracle/labeler with the classifier ($C_n(x_{i_n})$).



(a) P-S **(b) S-P**

Fig. 1. Performance comparison on Sentiment Analysis

References

1. Gong, B.: Discriminatively learning domain-invariant features for unsupervised domain adaptation. (2013)
2. Gretton, A., Smola, A.J.: A kernel method for the two-sample-problem (2007)
3. Settles, B.: Active learning literature survey. Tech. rep. (2010)

[1] `https://github.com/subbareddy248/Datasets/tree/master`