Statistical Machine Learning Project2 Report

Hongyu Chen
1062897

1. Introduction

Data mining and machine learning technologies have already achieved significant success in many fields including classification, regression and clustering. Nevertheless, many machine learning methods work well only under a common assumption: the training and test data are drawn from the same feature space and the same distribution, which leads to an issue that when the distribution varies, most statistical models need to be rebuilt from scratch using newly collected training data. In many real-world applications, it is expensive or impossible to re-collect the needed training data and rebuild the models. Propelled by the idea to reduce the need and effort to re-collect the training data and train the model anew, the concept transfer learning is proposed to handle the issue above.

In this project, we have three file: 'MALE.csv', 'FEMALE.csv', 'MIXED.csv'. In each file, each row represents a student with some related features and a label of their mark. We need to choose two file as our source domain and the remaining one as the target domain, then use the samples from source domain to train our model and finally predicate the mark (the label) of the samples from target domain.

2. Related Work

A lot of work has been done in transfer learning field[1]. Generally, we can categorize transfer learning under three sub-settings, *inductive transfer learning*[2,3], *transductive transfer learning*[4,5,6,7] and *unsupervised transfer learning*[8,9], based on different situations between the source and target domains and tasks.

In the *inductive transfer learning* setting, the target task is different from the source task, no matter when the source and target domains are the same or not. In this case, some labeled data in the target domain are required to induce an objective predictive model for use in the target domain.

In the *transductive transfer learning* setting, the source and target tasks are the same, while the source and target domains are different. In this situation, no labeled data in the target domain are available while a lot of labeled data in the source domain are available.

Finally, in the *unsupervised transfer learning* setting, similar with *inductive transfer learning* setting, the target task is different from but related to the source task. However, the *unsupervised transfer learning* focus on solving unsupervised learning tasks in the target domain, such as clustering, dimensionality reduction and density estimation. In this case, there are no labeled data available in both source and target domains in training.

Apart from the traditional categorization, some new neural network architecture such as CycleGAN[10] can also be used to do transfer learning task.

3. Task 1

In task1, I implemented 6 baseline methods evaluated by Daum´e III and Marcu[11] and a new feature-augmented method proposed by Hal Daum´e III[12] (we'll call this method as FEDA in later part).

3.1 Baseline

Some details about the 6 baselines are demonstrated as below:

- The SRCONLY baseline ignores the target data and trains a single model, only on the source data.
- The TGTONLY baseline trains a single model only on the target data.
- The ALL baseline simply trains a standard learning algorithm on the union of the two datasets. However, a potential problem with the ALL baseline is that if N << M (N are samples from target domain, M are samples from source domain), then M may "wash out" any affect N might have.
- The WEIGHTED baseline solves issue above to some degree. For instance, if M = 10 * N, we may weight each example from the source domain (M) by 0.1.
- The PRED baseline is based on the idea of using the output of the source classifier as a feature in the target classifier. Specifically, we first train a SRCONLY model. Then we run the SRCONLY model on the target data (train, dev and test). We use the predictions made by the SRCONLY model as additional features and train a second model on the target data, augmented with this new feature.
- In the LININT baseline, we linearly interpolate the predictions of the SRCONLY and the TGTONLY models. The interpolation parameter is adjusted based on target development data.

3.2 FEDA

In a nutshell, this method can be regarded as the feature augmented method. Suppose Ds and Dt are samples from source domain and target domain respectively, their feature spaces are $X \in R^F$ (F > 0), then the features of Ds and Dt can be augmented as:

$$\varphi(Ds) =< X, X, 0 >$$

$$\varphi(Dt) =< X, 0, X >$$

The first F space belongs to common feature space, the second F space belongs to source domain feature space and the third F space belongs to target domain space.

In general, for K domains (one target domains and K-1 source domains), the augmented feature space will consist of K+1 copies of the original feature space. This process is shown in Fig1.

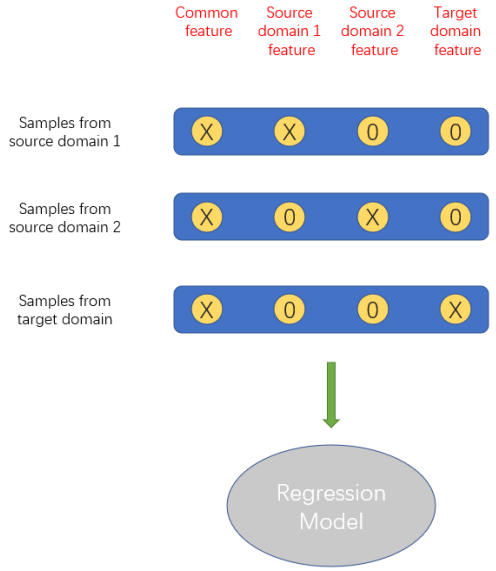

Fig1. Data processing of FEDA, feature space is augmented with the number of total domains, and different sample will be mapped to different final feature space according to their original domain.

3.3 Model choice

For the training process, I choose two different model to train: Full-Connected Neural Network and BayesianRidge Regression Model. The prediction result for each target domain with different algorithms will be demonstrated in Section 5.

Additionally, some hyper-parameters in neural network such as dropout rate can affect the result in obvious way. After several

tuning process, the best settings in my neural network architecture should be: the number of FC layers (3~4), dropout rate (0~0.3), training epoch (80~120), training batch (50).

4. Task 2

Considering the imbalanced training samples (many source-domain samples and few target-domain samples) and inspired by GAN (Generative Adversarial Network)[13] and VAE (Variational Auto-Encoder)[14], I think if we can use the ample source-domain samples to train a regression model and learn a mapping from target domain to source domain, we can naturally predicate the mark of a sample from target domain. The idea is shown in Fig2 and I'll call it FSMM (feature space mapping method) as my method in later part.

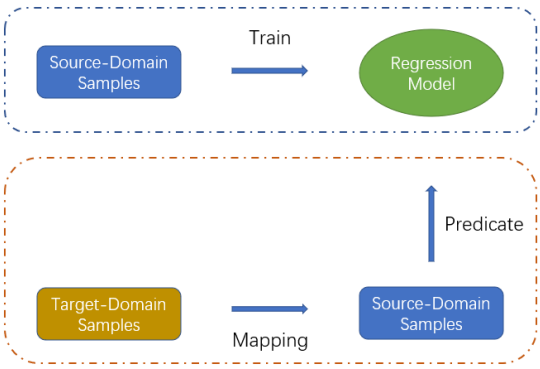

Fig2. Above part is using source-domain samples to train a regression model. Below part is using another model to learn the mapping from target-domain feature space to source-domain feature space.

5. Result and Analysis

Experiment results of the FEDA in Task1 and the results of FSMM in Task2 are shown in Table1. Because of space limit, the results of 6 baselines are hidden, but you are free to get them by running 'baseline.py'.

Some pre-definition details in Table1:

- Model: FEDA and FSMM.
- Learning Algorithm: Full-connected neural network (NN) and BayesianRidge Regression (BR).
- Target: Including 3 target domains – male, female, mixed.
- Hit-x: Rate of hit-x, represents the accuracy of the mark prediction, if ||predicted mark – true mark|| < x, this prediction will be labelled as Hit-x (on 100 target samples).

Model	Learning Algorithm	Target	Hit-0	Hit-2	Hit-5
FEDA	NN	Male	8%	16%	20%
	NN	Female	5%	15%	20%
	NN	Mix	6%	11%	16%
	BR	Male	6%	19%	22%
	BR	Female	4%	19%	16%
	BR	Mix	8%	20%	22%
FSMM	/	Male	3%	15%	21%
	/	Female	6%	12%	24%
	/	Mix	6%	11%	20%

Table1. Experiment results of FEDA and FSMM models on 'MALE/FEMALE/MIXED.csv'.

6. Conclusion

In conclusion, this project demonstrated a new field to me. With the help of this project, I have implemented several transfer learning models according to their original paper and have a deeper understanding about transfer learning, which is beneficial to my future research study.

Reference

[1]. Sinno Jialin Pan and Qiang Yang, “A survey on transfer learning” in *IEEE Transactions on Knowledge and Data Engineering*, Oct 2010, pp. 1345-1359.

[2]. W. Dai, Q. Yang, G. Xue, and Y. Yu, “Boosting for transfer learning,” in *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, Oregon, USA, June 2007, pp. 193–200.

[3]. R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: Transfer learning from unlabeled data,” in *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, Oregon, USA, June 2007, pp. 759–766.

[4] H. Daum´eIII and D. Marcu, “Domain adaptation for statistical classifiers,” *Journal of Artificial Intelligence Research*, vol. 26, pp. 101–126, 2006.

[5] B. Zadrozny, “Learning and evaluating classifiers under sample selection bias,” in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Alberta, Canada, July 2004.

[6] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of Statistical Planning and Inference*, vol. 90, pp. 227–244, 2000.

[7]. A. Arnold, R. Nallapati, and W. W. Cohen, “A comparative study of methods for transductive transfer learning,” in *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 77–82.

[8] W. Dai, Q. Yang, G. Xue, and Y. Yu, “Self-taught clustering,” in *Proceedings of the 25th International Conference of Machine Learning*. ACM, July 2008, pp. 200–207.

[9] Z. Wang, Y. Song, and C. Zhang, “Transferred dimensionality reduction,” in *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008*. Antwerp, Belgium: Springer, September 2008, pp. 550–565.

[10]. Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *proceedings of 2017 ICCV*, 2017.

[11]. Hal Daum´e III and Daniel Marcu. “Domain adaptation for statistical classifiers,” in *Journal of Artificial Intelligence Research*, 2006.

[12]. Hal Daum´e III, “Frustratingly Easy Domain Adaptation,” in *proceedings of ACL 2007*, 2007.

[13]. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, "Generative Adversarial Networks," in *proceedings of 2014 NIPS*, 2014

[14]. Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," 2013.