CE888 Assignment 1 - Project Proposal: Learning under covariate shift

Suraj Ghuwalewala

Abstract—The changing nature of several environments can lead to a change in the distribution of the population data which in turn can affect the performance of a machine learning model. This project focuses on the methods to detect and adapt to one such shift known as covariate shift. The original and the updated datasets will be used to train classifiers and their individual performances will be compared to gain a deeper understanding of the effect of covariate shift in the learning performance.

1 Introduction

This paper focuses on detection and correction of covariate shift, which is a phenomenon in which the population distribution of the training data is not same as the test data [1]. Such shift occurs in the data because of the constantly evolving and changing nature of the environment. Such shift can drastically affect the performance of the learning models trained on an old dataset. So, the detection and correction of this shift becomes crucial. Two such datasets have been chosen from Kaggle. These datasets will undergo detection of covariate shift, then appropriate methods will be used to correct the shift if detected. Finally, some basic classifiers will be trained on these datasets and their performance will be evaluated using several performance evaluation metrics.

The subsequent sections will talk in detail about the project. Section 2 will focus on the literature review and the background study of the topic, while section 3 will detail the plan of action and datasets. Subsequently, section 4 will throw some light on the planned experiments for the evaluation and section 5 will discuss the performance metrics that will be used to evaluate the classifiers. Finally, section 6 will conclude the whole project.

2 Background

According to [1], the term "dataset shift" was used in 2006, which is broadly defined as "case where the joint distribution of inputs and outputs differs between training and test stage, i.e., when,

$$P_{\text{train}}(y, x) \neq P_{\text{test}}(y, x)$$

. [1] also mentions that the concept of dataset shift has been referred in the literature with various terminologies like concept shift, changes of classification, changing environments, contrast mining and fracture points. Now, in real world applications, dataset shift is broadly classified into three types: (1) covariate shift, (2) prior probability shift and (3) concept shift [2]. This paper focuses on covariate shift, also known as 'population drift', which refers to a "case where the population distribution may change over time. (i.e) it only appears in X $\rightarrow$ Y problems and is defined as a case where;

$$P_{\text{train}}(y, x) = P_{\text{test}}(y, x)$$

BUT,

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$

[1]. Such shift in the distribution of the dataset can deter the performance of the machine learning models over time, and it is quintessential to detect and correct this imbalance. [2], [3] suggests exponentially weighted moving average (EWMA) mnodel for covariate shift detection in non stationary environments. Also, [4], [5] has used the same EWMA model to detect covariate shift in brain computer interface (BCI) systems.

Two most well knowns techniques for correction of the covariate shift are *Kullback-Leibler Importance Estimation Procedure (KLIEP)* [6] and *Kernel Mean Matching (KMM)* [7]. [8] has applied *Frank-Wolfe algorithm* to KLIEP and KMM and demonstrated it to be more effective for covariate correction. On the other hand, [9] has proposed a method for reweighing the observations while retaining low variance. [10] uses a novel cross validation technique called *importance weighted cross validation* to select the most optimal model for fitting onto the covariately shifted dataset for classification problems.

3 Methodology

The project revolves around the comparative analysis of the effect of covariate shift on predictions from learning model. The idea was to initially train a model on a dataset with covariate shift, then correct the shift and train a separate model, and finally compare the performance of the two models. The two datasets chosen from *Kaggle* [11] were *Santander Customer Transaction Prediction* [12] and *Google Analytics Customer Revenue Prediction* [13]. These datasets were directly downloaded into *Google Colab* notebook environment using the Kaggle API for initial analysis.

The first step of the analysis will be checking for null values and and imputing them. Then, the distribution of the data along several feature vectors will be plotted for both train and test sets to visually analyse the pattern in distribution. Also, the shift in the distribution of each feature

S. Ghuwalewala is a Post-Graduate student at the School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK.
E-mail: sg19394@essex.ac.uk.

was detected using the method described in [14]. One of the way to deal with covariately shifted features to drop them, but it can lead to information loss. Several other methods have been discussed in the background (Section 2). The python implementation of these methods will be studied and suitable techniques will be used to deal with covariate shift. Basic classifiers like k-nearest neighbour(k-nn), support vector machines (svm) and naive bayes (nb) will be used for the comparative analysis of these datasets.

4 Results

As explained in the Section 3, every feature in the data will be tested for presence of covariate shift. These features will be treated with the methods mention in section 2 upon further analysis of each method. Since both the datasets have a large number of feature vector, Principal Component Analysis (PCA) will be used to generate several sets of components covering different ratios of information, and each set can be test for performance in terms of metric score and time taken for execution. Three basic classifiers - kNN, SVM and NB will be used for evaluating the performance on each PCA decomposed dataset with and without covariate shift. The evaluation metrics has been described in the subsequent section.

5 Discussion

The evalutation of the above mentioned classifiers will done using the metrics like classification accuracy, area under curve (auc) as well as the kaggle competetion specific metrics like root mean square error (rmse). The predictions for both datasets will uploaded to the respective Kaggle competitions and the performance score on the private leaderboard will be used as the basis for analysing the individual performance.
This evaluation will assist in understanding the true effect of covariate shift in the performance of the classifiers. In today's constantly evolving environment, the patterns and trends in the data is bound to change. So understanding the effect of such change on the machine learning models can be used to design the models to adapt such changes robustly and perform with higher accuracy.

6 Conclusion

The aim of this project is to identify the effect of covariate shift on the performance of the classifiers. Several methods, as discussed in the background and methodology sections will be tried to detect and eventually correct or adapt covariate shift. There is a need to study and evaluate each of the above mentioned methods in detail for implementation. The final evaluation will done by uploading the required metrics by the respective Kaggle competitions. The plan to go ahead with the project has been mentioned at the end of the report.

References

[1] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, p. 521–530, 2012.

[2] H. Raza, G. Prasad, and Y. Li, "Adaptive learning with covariate shift-detection for non-stationary environments," in *2014 14th UK Workshop on Computational Intelligence (UKCI)*, pp. 1–8, Sep. 2014.

[3] H. Raza, G. Prasad, and Y. Li, "Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments," *Pattern Recognition*, vol. 48, no. 3, p. 659–669, 2015.

[4] H. Raza, H. Cecotti, Y. Li, and G. Prasad, "Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface," *Soft Computing*, vol. 20, no. 8, p. 3085–3096, 2015.

[5] A. Chowdhury, H. Raza, Y. K. Meena, A. Dutta, and G. Prasad, "Online covariate shift detection-based adaptive brain–computer interface to trigger hand exoskeleton feedback for neuro-rehabilitation," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, pp. 1070–1080, Dec 2018.

[6] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 1433–1440, Curran Associates, Inc., 2008.

[7] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[8] J. Wen, R. Greiner, and D. Schuurmans, "Correcting covariate shift with the frank-wolfe algorithm," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[9] S. J. Reddi, B. Poczos, and A. Smola, "Doubly robust covariate shift correction," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[10] M. Sugiyama, M. Krauledat, and K.-R. MÃžller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 985–1005, 2007.

[11] "Your machine learning and data science community." url=https://www.kaggle.com/.

[12] "Santander customer transaction prediction." url=https://www.kaggle.com/c/santander-customer-transaction-prediction/data.

[13] "Google analytics customer revenue prediction." url=https://www.kaggle.com/c/ga-customer-revenue-prediction/data.

[14] georsara1, "Covariate shift check-up –updated." url=https://www.kaggle.com/georsara1/covariate-shift-check-up-updated, Oct 2018.

PLAN

The milestone 1 of the project is achieved by submitting this proposal. The main plan now is to understand the methods of covariate shift adaptation from Professor H. Raza in detail, understand the literature in a much proper way and implement the concepts in python. Major work is appropriate implementation of CS detection and adaptation technique.
I suppose that the subsequent work on designing classifiers and evaluating performance would not take much time based on past experience.

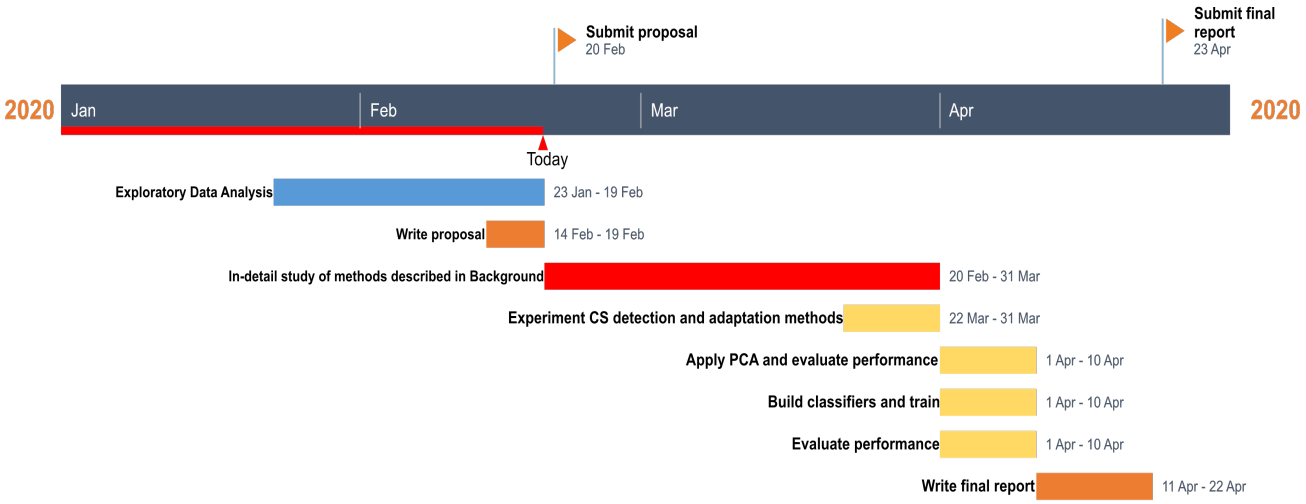

Fig. 1. Gantt Chart - Project Plan

The project has been uploaded in a folder named *assignment* in the github repo mentioned below;
Project Github link: https://github.com/surajghuwalewala/$CE888_{D}ata_{S}cience_{a}nd_{D}ecision_{M}aking.git$