Learning under Covariate Shift in the Data (Supervised Learning)

Khaleda Begum (kb19250@essex.ac.uk)

Abstract— **Data shift like covariate shift is very common in real world data. Supervised Learning with data under covariate shift requires the minimization of the data shift between training and test data before applying to leaning model. Through making changes in features of dataset with different methods the covariate shift is minimized. Before applying covariate shift minimization technique it is necessary to detect covariate shift among the data. In this article two real world datasets were taken from kaggle website. Histogram Intersection and Kullback Divergence method was applied to detect the covariate shift in real world dataset. The data shift visualization was done through histogram where the data distribution difference was shown also. After that learning model was developed with the real world dataset using classifier and features importance was estimated so that significant features can be used in learning model. For both datasets the learning model was retrained with features excluding least important column with covariate shift. This steps did not change the learning accuracy significantly. My source code for this experiment is publicly available at Github at "https://github.com/rroxy08/CE888/assignment1/" [1]**

I. INTRODUCTION

Covariate shift is defined as something that occurs "when the data is generated according to a model P(y—x)P(x) and where the distribution P(x) changes between training and test scenarios" [2] . In supervised learning process in a simple machine learning model there are two steps: training and testing. In training process, samples in training data are taken as input in which features are learned by learning algorithm or learner and build the learning model. In the testing process, learning model uses the execution engine to make the prediction for the test or production data. Tagged data is the output of learning model which gives the final prediction or classified data ' [3]. When developing methods of supervised learning, it is commonly assumed that samples used as a training set and data points used for testing the generalization performance follow the same probability distribution. But this common assumption is not fulfilled in real-world applications of machine learning such as robot control, brain signal analysis, and bioinformatics ' [4]. Hence in supervised learning before using any data, the data shift detection is needed. If there is presence of any data set shift then proper method should be applied to minimize the shift distribution so that the data can be learned in best way with less classification error. The goal of this article is to find out how data with covariate shift can be learned under supervised learning model. For that reason two different datasets from keggle were selected. First the covariate shift in the train and test data is detected and then the features of the data are adapted in such a way that the data shift is minimized so that the learner model can learn with less classification error.

II. Background

Since data shift is common in real life data, different methods for learning data under covariate shift have been widely studied and proposed. Basic steps of these works are to detect the covariate shift and then adapting the learning process by minimizing data shift using suitable methods. Since covariate data shift is the difference of data distribution among train and test data, it can be detected using various methods like histogram visualization, histogram intersection [5], Kullback-Leibler divergence method [6] etc. One of the mostly used way of detecting and minimizing covariate shift in data is to determine and use feature importancee weight. Different methods have been proposed to use importance weight to minimize the covariate shift; like Kernel Density Estimation (KDE) method , Discriminative Learning process, Kernel Mean Matching(KMM), Kullback Leblier Importance Estimation Procedure (KLIEP), Least Squares Importance Fitting (LSIF), Unconstrained Least Squares Importance Fitting (uLSIF) [7] . [8] proposed importance weighted cross validation (IWCV), procedure that can be applied for unbiased classification under covariate shift. In [9] a direct importance estimation method was proposed (that does not involve density estimation) called Kullback–Leibler Importance Estimation Procedure (KLIEP) and it is based on the minimization of the Kullback–Leibler divergence of data distribution. An exponential weighted moving average (EWMA) model has been proposed in [10] to detect covariate shift. In [11] Exponential weighted moving average (EWMA) model was used to detect covariate data shift and transductive-inductive learning model was used to adapt to covariate shift learning.

III. Methodology

For training a learning model with data under covariate shift, the experiment was done in two steps. First the datasets were inspected for covariate shift. For detecting covariate shift , the columns of both train and test data were visualized through histogram plotting. Also histogram intersection method and kullback leibler divergence method was used to detect the column data with covariate shift [12] . In Kullback Divergence modela a threshold is defined and the distance is computed using a sample based approximation. In case of values greater than the threshold, covariate shift is present. In case of being the distance lower than the threshold covariate shift can be discarded [13] . similarly for histogram intesection method the intersection value near 1 means data distribution is close [5].

Histogram Intersection calculation method:

$$d(p,q) = 1 - (SUM(min(p_i,q_i))/SUM(q_i))$$

and Kullback leibler divergence calculation method:

$$D_KL(p,q) = SUM(p_i log(p_i/q_i))$$

Then the features of the datasets were processed to minimize the shift between two distributions (train and test). after preparing the features the data were tested through classification learning model to test if the data is learned in best way than before. In this article two datasets were taken from kaggle for experiment. The two kaggle data sets are:

1) Dataset1: Sberbank Russian Housing Market [14] .
2) Dataset2: Porto Seguro's Safe Driver Prediction [15] .

The learning model was developed using random forest classifier. Feature importance for each dataset was calculated. Then again the learner model was fitted with data excluding covariate shift data with least feature importance. Excluding these column didn't change the learning accuracy significantly.

A. Dataset1:

The 'Russian Housing Market' dataset [14] is taken from Kaggle repository. This dataset has 291 attributes and one target variable "price doc". The training set has 30471 samples and the test set has 7662 samples. The dataset was pre-processed for filling the missing values with mean values in case of 'int' data types and with mode values in case of 'object' and 'float' data types, since the columns like 'build year', 'state' was float type. The string values in the attributes were converted to neumerical values like 0,1,2 etc.

B. Dataset2:

The 'Porto Seguro's Safe Driver Prediction' dataset [15], taken from Kaggle repository has 59 attributes and one target variable "target". The training set has 595212 samples and the test set has 892816 samples. The dataset was pre-processed for filling the missing values with mean values in case of 'int' and 'float' data types and with mode values in case of 'object' data type. The string values in the attributes were converted to neumerical values like 0,1,2 etc.

IV. Results

A. Dataset-1:

1) *Visualization of train and test data::* For visualization of data distribution difference, the samples from each corresponding column of both train and test data are plotted using histogram chart. Since the dataset-1[14] has 291 columns/ attributes, it is not possible to show the 291 histograms in this article. Hence only some columns' data distribution are shown in fig:1 and 2 along with their histogram inersection value and Kullback Leibler Divergence value.

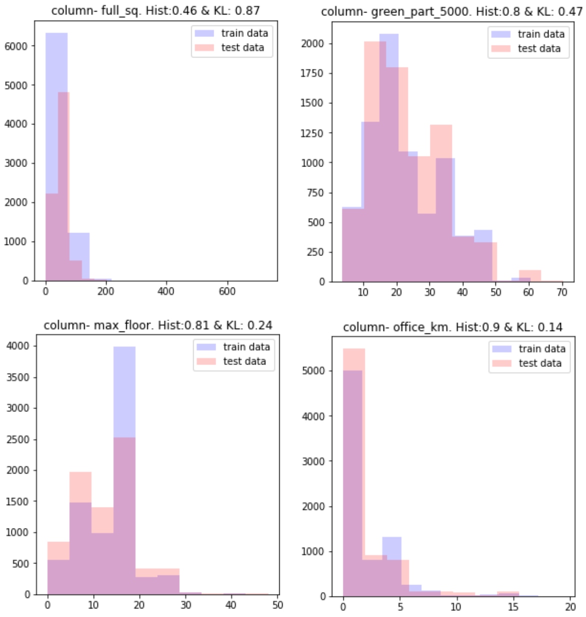


Fig. 1: Histogram of columns of dataset-1

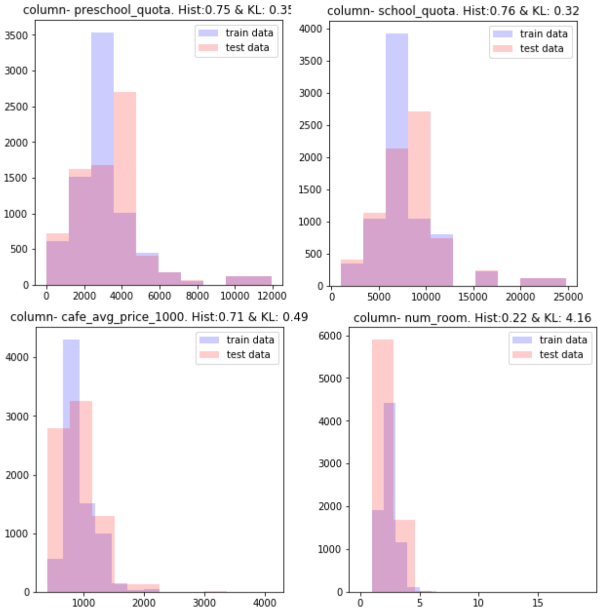


Fig. 2: Histogram of columns of dataset-2

2) *Applying Covariate Shift detection method::* From the histogram visualization of dataset-1 it was seen that there are some data shift in the train and test data. For histogram intersection value, if the value is more than 0.90, the data are almost closely distributed. Similarly the for KL divergence value, less than 0.02, the data are closely distributed. Hence for detecting covariate data shift in the 290 columns('id' column was excluded) data of both train and test data, threshold values for histogram intersection and KL divergence was set 0.9 and 0.02 respectively. With histogram intersection method 32 and KL divergence method total 28

columns were found to have different distribution of data.

B. Dataset-2

1) Visualization of train and test data:: Similarly the samples from each corresponding column of both train and test data are plotted using histogram chart for the dataset-2[15] that has 59 columns/ attributes. Only 8 columns' including data shift are shown in fig:3 and 4 along with their histogram inersection value and Kullback Leibler Divergence value.

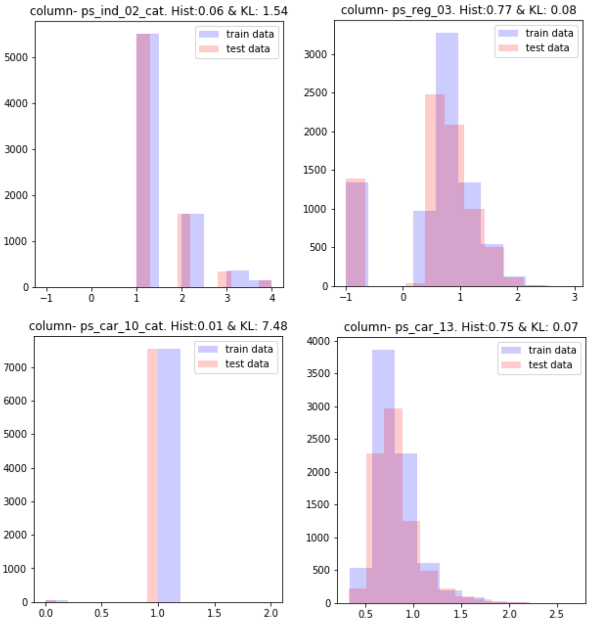


Fig. 3: Histogram of columns of dataset-2

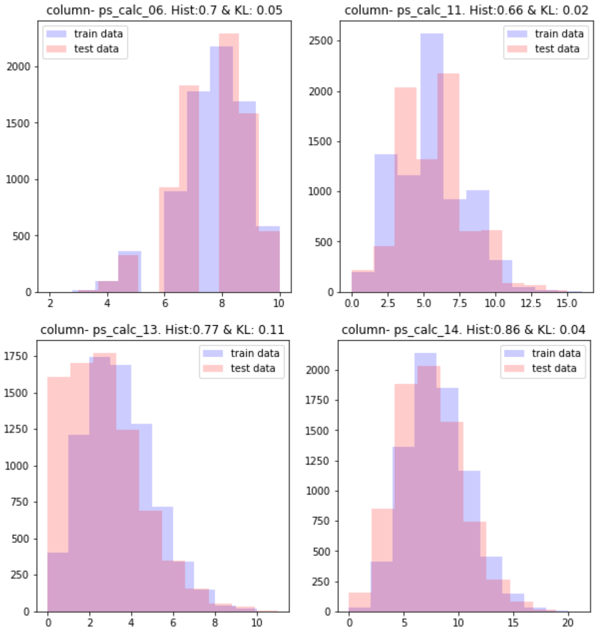


Fig. 4: Histogram of columns of dataset-2

2) Applying Covariate Shift detection method:: For dataset-2, the threshold for histogram intesection value was set 0.9 and the Kullback Divergence value was set 0.02. With these threshold value 8 columns were detected having covariate shift more than acceptable level.

A learning model was developed using classifier and its initial learning accuracy was calculated. For dataset-1 the accuracy was very low. but for dataset-2 the accuracy was .962. Then for both dataset features importance was calculated. The detected columns with covariate shift and having least importance was removed from the dataset. Then again learning model was trained and tested. For both datasets the learning accuracy didnot change.

V. Discussion

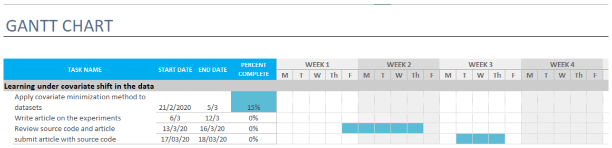
From the Histogram chart and the two data shift detection method it is seen that, covariate shift is present among the samples of the columns between train and test data. There are many ways of reducing the covariate shift which can be applied on the dataset to minimize the shift and increase learning model's accuracy. Initially the importance of the features were estimated to find the least important feature to the classifier so that features for the learning model can be selected. Some features with covariate shift and least feature importance value was removed from the train data. After removing them the leaning model was developed again which didnot show any significant improvement of learning score. But removing features is not always good option for learning model, as it may reject valuable information. So other shift minimization methods should be used.

VI. Conclusions

Covariate shift in data is the difference between input data distribution. Histogram chart is the easiest way for covariate shift visualization. But for large number of columns/attributes visualizing data is a time consuming process. That's why data distribution comparison methods like Histogram Intersect method and Kullback Leibler Divergence method was used to detect covarite shift. Then the data was applied to develop a learning model. After estimating feature importance , some covariate data features with least importance was removed from train data and again the learning model was trained. That did not improve the learning accuracy . So different covariate shift minimization methods can be used on features to minimize the covariate shift. At the same time it should be compared how the learning model learns with the new data after shift minimizing.

VII. Plan

So far in this article only the covariate shift detection was done. Also learning accuracy was tested to ignoring the least important features with covariate shift. Next step could be applying different covariate shift minimization technique on data to see which method can minimize it in best way. Also the after applying each method the data should be applied to learning model to compare learning accuracy. By this we can conclude how to learn a data with covariate shift using supervised learning model.



GANTT CHART

TASK NAME	START DATE	END DATE	PERCENT COMPLETE	WEEK 1					WEEK 2					WEEK 3					WEEK 4				
				M	T	W	Th	F	M	T	W	Th	F	M	T	W	Th	F	M	T	W	Th	F
Learning under covariate shift in the data																							
Apply covariate minimization method to datasets	21/2/2020	5/3	15%																				
Write article on the experiments	6/3	12/3	0%																				
Review source code and article	13/3/20	16/3/20	0%					■	■	■	■	■											
submit article with source code	17/03/20	18/03/20	0%												■	■	■						

Fig. 5: GAntt chart

VIII. Source Code

Source code for two keggle datasets experiments [14][15] implementation is available at [1].

References

[1] "github repository." https://github.com/rroxy08/CE888/assignment1/.

[2] 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, pp. 521–530, 2012.

[3] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, 12 2017.

[4] M. Sugiyama and M. Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[5] S. Ando, "Histogram intersection for change detection." http://blog.datadive.net/histogram-intersection-for-change-detection/. Accessed: 2020-02-15.

[6] Delas, "Covariate shift in machine learning inference." https://jorditg.github.io/machinelearning/covariate_shift_in_machine_learning_inference, 2018. Accessed: 2020-02-15.

[7] N. G. Nair, P. Satpathy, J. Christopher, *et al.*, "Covariate shift: A review and analysis on classifiers," in *2019 Global Conference for Advancement in Technology (GCAT)*, pp. 1–6, IEEE, 2019.

[8] 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.

[9] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 4, pp. 699–746, 2008.

[10] 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, IEEE, 2014.

[11] H. Raza, H. Cecotti, Y. Li, and G. Prasad, "Learning with covariate shift-detection and adaptation in non-stationary environments: Application to brain-computer interface," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2015.

[12] E. Bazan, P. Dokládal, and E. Dokladalova, "Quantitative analysis of similarity measures of distributions," 2019.

[13] J. de la Torre, "Covariate shift in machine learning inference." http://www.neural-solutions.com/2018/03/09/covariate_shift_in_machine_learning_inference.html/. Accessed: 2020-02-15.

[14] "Sberbank russian housing market." https://www.kaggle.com/c/sberbank-russian-housing-market/. Accessed: 2020-02-15.

[15] "Porto seguro's safe driver prediction." https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/. Accessed: 2020-02-16.