Detecting Distributional Shift Responsible for Predictive Model's Failure*

1st Dipanwita Sinha Mukherjee

Enterprise Analytics & Data Science
Artificial Intelligence - Center of Excellence
Wells Fargo International Solutions Private Limited
Bangalore, India
Dipanwita.Sinhamukherjee@wellsfargo.com

2nd Divyanshy Bhandari

Enterprise Analytics & Data Science
Artificial Intelligence - Center of Excellence
Wells Fargo International Solutions Private Limited
Bangalore, India
Divyanshu.Bhandari@wellsfargo.com

3rd Naveen Gururaja Yeri

Head of Enterprise Analytics & Data Science
Artificial Intelligence - Center of Excellence
Wells Fargo International Solutions Private Limited
Bangalore, India
naveen.yer@wellsfargo.com

Abstract—Any predictive software deployed with this hypothesis that test data distribution will not differ from training data distribution. Real time scenario does not follow this rule, which results inconsistent and non-transferable observation in various cases. This makes the dataset shift, a growing concern. In this paper, we've explored the recent concept of Label shift detection and classifier correction with the help of Black Box shift detection(BBSD), Black Box shift estimation(BBSE) and Black Box shift correction(BBSC). Digits dataset from "sklearn" and "LogisticRegression" classifier have been used for this investigation. Knock out shift was clearly detected by applying Kolmogorov–Smirnov test for BBSD. Performance of the classifier got improved after applying BBSE and BBSC from 91% to 97% in terms of overall accuracy.

Index Terms—Dataset shift, Label shift, KS-test, Classifier, Dimensionality reduction

I. INTRODUCTION

Machine Learning based powering software systems are now widely used in many industries, enabling vision systems in self-driving cars and social networks, automating language translation, strengthening recommendation engines used by online platforms, providing assistance in radiology and advancing medical sciences, assisting best performing speech recognition software, and many more. In short, these software systems are playing important role to drive significant decisions while leveraging their predictions [1]. But at the same time, due to lack of detailed clarity, how these ML software works, it is very hard

to test and maintain. Therefore, modern ML systems are brittle and subtle changes in the data distribution [2], [3] can affect or destroy the performance of the models, even shifts in the label distribution can significantly suppress accuracy. Also, detecting shift in high-dimension cases are still an open ended topic.

Suppose, we trained a learner (f) on some images of chest X-rays of previous year's data (Source distribution $\mathbf{P}$) to predict whether a patient have tuberculosis or not [4]. Assume, that the training set contains 0.8% positive cases and the model approximately predict the same in a reliable manner. Suddenly, in last week (Target distribution $\mathbf{Q}$) the model prediction rises to 7%. This raise the concern of whether our trained classifier is predicting accurately or not (The target distribution itself may got changed).

Shifts broadly can be classified as *covariate shift* [5]: where $p(y|x)$ remains fixed and *label shift* [6]: where $p(x|y)$ remains fixed. In above mentioned pandemic situation, findings of a person having tuberculosis given cough $p(y|x)$ as symptoms may rise but the manifestation of the disease $p(x|y)$ can't change. This leads to our interest of investigation about label shift, which is under explored. Mathematically it can be formulated as:

$$q(y, x) = q(y)p(x|y) \quad (1)$$

The objective is to identify the shift if it exist by checking the similarity of the source distribution

$\mathbf{P}$ (from where training data is sampled) and target distribution $\mathbf{Q}$ (from where real world data is sampled). Our work inspired from Lipton [7] who shows that checking whether source distribution differ from target distribution or not can be identified by checking that whether $p(f(x)) \neq q(f(x))$. We use their Black Box Shift Estimation (BBSE) concept to estimate label shift using a black box predictor(f). Afterwards, a systematic method of correcting for label shift has been computed through importance weighted ERM. In short, our work has been focused on:

- Detect distributional shift
- Quantify it and
- Correct the model so that it can perform well on unseen changed data

II. PROBLEM FORMULATION

In this setup, we use $x \in X = \mathbb{R}^d$ and $y \in Y$ represents the features and label variables respectively. Y is in discrete space $1, 2, \dots, k$ and $p(y)$ denotes the p.m.f. of $y \sim P$ (source distribution) where $q(x)$ is the p.d.f. of $x \sim Q$ (target distribution).

We have samples with labelled $((x_1; y_1); (x_2; y_2); \dots (x_n; y_n) \sim P)$ and unlabelled data $(x'_1; \dots; x'_m \sim Q)$. Detecting shift between these two samples can be seen as hypothesis testing where null hypothesis is $H_0 : q(y) = p(y)$ and the alternative is $H_1 : q(y) \neq p(y)$. H_0 can be verified with two sample statistical test (such as Kolmogorov–Smirnov test, Anderson-Darling test, Maximum Mean Discrepancy test etc.) on samples from $p(\hat{y})$ and $q(\hat{y})$ where $\hat{y} = f(x)$ and $f : X \rightarrow Y$. This shift detection is followed by Shift estimation (estimating $w(y)$ and $q(y)$) and Shift correction. Estimating task requires three assumptions to consider:

- Conditional distribution for source and target distribution are same

$$p(x|y) = q(x|y) \quad \forall x \in X, y \in Y.$$
- For every $y \in Y$ with $q(y) > 0$, it requires $p(y) > 0$.
- The confusion matrix $C_p(f)$ of the predictor(f) should be invertible.

III. PRIOR WORK

Although its wide applicability, learning under label shift with unknown $q(y)$ is not well explored. Storkey [8] infer about posterior from unlabeled prior $p(y)$ which requires to estimate $p(x|y)$, difficult in high-dimensional datasets. Chan & Ng [9] also requires estimating $p(x|y)$ to infer about $q(y)$. Zhang [10] works on kernel mean matching for this label shift

problem. In summary, all of them have investigated heavily to estimate $q(x)$ the ratio $q(x)/p(x)$ or $p(x|y)$. In real world scenario, with increasing dataset size and dimensionality, this estimation are becoming difficult. Shimodaira and Gretton [11], [12] proposes different methods (ERM, Kernel mean matching etc.) to correct models via re-weighted training samples.

The most celebrated work by Lipton [7] has detected and quantify the label shift. They proposed Black Box Shift Estimation (BBSE) to estimate the test distribution $q(y)$. BBSE exploits any black box predictor model to reduce dimensionality prior to shift correction. Rabanser [13] extend the idea of Lipton [7] by exploring different dimensionality reduction techniques with Black Box shift detection followed by subsequent two-sample testing to detect dataset shift.

IV. PROPOSED TECHNIQUE

In this work, the proposed method from Lipton [2] which works for any kind of predictor(f) has been experimentally investigated. Inspired from their concept, in this work also we've used black box dimensionality reduction technique followed by two-sample hypothesis testing. Jointly, they called as BBSD(black box shift detection). Once the shift got detected, BBSE (black box shift estimation) and BBSC(black box shift correction) have been used to correct the black box classifier as pictured in Fig.1. Detailed description regarding each part has been mentioned below:

Fig. 1: Dataset shift detection pipeline

A. Dimensionality Reduction Techniques—Label Classifier (BBSD):

As a dimensionality reduction technique, different classifiers can be used, hence it is named as Label classifier. Here, we have tried with Logistic Regression, Support vector machines and Multilayer perceptron classifier from ‘sklearn’ library to train from source dataset, though all the results are from Logistic Regression only. Recently, Stephan [13] have showed

that among different dimensional reduction techniques (PCA, Sparse random projector, Autoencoders, Label classifier etc.), Label classifier named as BBSD outperforms as DR (dimensional reduction) technique for multiple univariate testing case.

B. Statistical Hypothesis Testing – Kolmogorov-Smirnov test (BBSD):

As stated previously also, that assumption of null hypothesis $H_0 : q(y) = p(y)$ can be tested by running various two sampled test. Here, Kolmogorov-Smirnov has been used which is a nonparametric and used to decide that if a sample comes from a reference probability distribution (one-sample test) or two samples (two-sample test) are coming from the same distribution or not. The advantage of running KS test on x over $f(x)$ has advantage of one-dimensional representation. KS test returns with p value and test statistics. Depending on these values, shift detection decision needs to be taken. In other words, a shift will exist when there is statistically significant difference in distribution between source and target data.

C. Black Box Shift Estimation (BBSE):

In this stage, the ratio $w_l = q(y_l)/p(y_l)$ for each label l , has been estimated. This is calculated on the assumption that the obtained confusion matrix belong to classifier is invertible. Also, in order to do that, $q(y_l)$ also need to be estimated, which is difficult as don't have samples from $q(y_l)$.

D. Black Box Shift Correction (BBSC):

Importance-weighted empirical risk minimization (ERM) method was used by Shift correction [7]. ERM [14] is based on a weighted loss function. If the imported base classifier supports, the 'class_weight' attribute then the predictions can be improved via the ERM.

V. RESULTS AND DISCUSSION:

We experimentally investigated the utility of proposed method with the help of Digits dataset from 'Scikitlearn' library. It is an 1797, $8 * 8$ grayscale images containing handwritten digits from 0–9. The image from the dataset has been shown below (Fig. 2).

Fig. 2: Digits Image

Knock-out shift has been incorporated for simulating the label shift effect. To do that, first data has been split into train (80%) and test (20%), afterwards a fraction of data from training set is knocked out. The observation goes with the intuition that with increasing fraction of knock out data, the difference between the distribution $p(y)$ and $q(y)$ increases which makes the detection easier.

A. Findings of the Study

LogisticRegression class from sklearn is used as a Black Box(f) classifier and is trained with 'liblinear' solver. Afterwards, a two-sample KS test has been conducted for statistical identification of whether the distribution $f(Trainset)$ and $f(Testset)$ are the same or not. Hence, BBSD reports a p value obtained from KS-test which was observed to decrease with increasing knock out fraction(δ) and reported in Table I. Threshold value for shift detection has been set for '0.05', hence there is a prominent shift for the below mentioned tabulated cases.

Knock-out fraction(δ)	p Value
0.5248	0.001759
0.7342	$2.7683e^{-05}$
0.8776	$5.5511e^{-16}$

TABLE I: p value against knock-out fraction

The changes in the marginal distribution due to the knock-out shift is profoundly visible and in Fig. 3 with fraction 0.7342 compared to original plot Fig. 4 . BBSD successfully detects this distributional change with p value $2.7683e^{-05}$, far below the threshold value.

Fig. 3: After applying shift

Fig. 4: Before applying shift

Fig. 5: Confusion matrix under label shift

Fig. 6: Classification Report under Label shift

It is clear (Fig. 7 And Fig. 8), that after applying correction through BBSC, classification performance has been improved. The reported accuracy is now 97% for Precision, Recall and F1 score (Fig. 8).

In the next phase, BBSE is used to estimate the ratio w . The whole dataset is divided into train and test, where train part was used to train the classifier(f), test part was used to estimate w . This estimator also corrects the label-shift via importance-weighted ERM [15] through a systematic process, hence called as BBSC. In short, BBSC corrects the classifier through importance-weighted empirical risk minimization approach. Fig. 5 and Fig. 6 is the plot of confusion matrix and classification report of the above-mentioned classifier under dataset shift scenario. The accuracy is showing 91% for Precision, Recall and F1 score.

Fig. 7: Confusion matrix after correction under label shift

Fig. 8: Classification Report after correction under label shift

VI. CONCLUSION

The proposed label shift detection approach [7] has been investigated here for image dataset. Source and Target distribution are fed to Black Box models which reduces the dimensionality of the input dataset. Afterwards, the reduced representations are analyzed using two-sample KS Test. The identification of dataset shift was correctly reported with varying shift intensity. The overall utility of this process can be summarized as:

- Accuracy does not depend on the data dimensionality. With the help of Black box model, high dimensional data is converted to lower dimensional representation.
- All three methods (BBS, BBSE, BBSC) works with any kind of Black box models. In other words, it does not depend on whether the ML model is Logistic Regression, Neural Net, SVM etc. Also, can be incorporated with deep neural net architecture.
- Therefore, this shift detection techniques can be used with any ML model as a pipeline to detect and correct the model due to generality.

There is a scope to investigate this shift detection techniques for other kind of dataset with deep neural network architectures.

REFERENCES

- [1] S. Angra and S. Ahuja, "Machine learning and its applications: A review," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, 2017, pp. 57–60.
- [2] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *arXiv preprint arXiv:1906.02530*, 2019.
- [3] B. Turhan, "On the dataset shift problem in software engineering prediction models," *Empirical Software Engineering*, vol. 17, no. 1, pp. 62–74, 2012.
- [4] S. Sathitranacheewin, P. Sunanta, and K. Pongpirul, "Deep learning for automated classification of tuberculosis-related chest x-ray: dataset distribution shift limits diagnostic performance generalizability," *Heliyon*, vol. 6, no. 8, p. e04614, 2020.
- [5] F. Alberge, C. Feutry, P. Duhamel, and P. Piantanida, "Detecting covariate shift with black box predictors," in *2019 26th International Conference on Telecommunications (ICT)*. IEEE, 2019, pp. 324–329.
- [6] J. Guo, M. Gong, T. Liu, K. Zhang, and D. Tao, "Ltf: A label transformation framework for correcting label shift," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3843–3853.
- [7] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *International conference on machine learning*. PMLR, 2018, pp. 3122–3130.
- [8] A. Storkey, "When training and test sets are different: characterizing learning transfer," *Dataset shift in machine learning*, vol. 30, pp. 3–28, 2009.
- [9] Y. S. Chan and H. T. Ng, "Word sense disambiguation with distribution estimation," in *IJCAI*, vol. 5. Citeseer, 2005, pp. 1010–5.
- [10] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International Conference on Machine Learning*. PMLR, 2013, pp. 819–827.
- [11] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.
- [12] W. Zaremba, A. Gretton, and M. Blaschko, "B-tests: Low variance kernel two-sample tests," *arXiv preprint arXiv:1307.1954*, 2013.
- [13] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *arXiv preprint arXiv:1810.11953*, 2018.
- [14] S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton, "A unified view of label shift estimation," *arXiv preprint arXiv:2003.07554*, 2020.
- [15] M. Sugiyama, B. Blankertz, M. Krauledat, G. Dornhege, and K.-R. Müller, "Importance-weighted cross-validation for covariate shift," in *Joint Pattern Recognition Symposium*. Springer, 2006, pp. 354–363.