Covariate dataset shift detection with convex hulls

Kelvin Sanchez
Johns Hopkins Whiting School of Engineering
Baltimore, Maryland USA
ksanche9@jh.edu

***Abstract*—Learning Enabled Cyber Physical Systems need protection from learning data that may be harmful to users who interact or rely on them. Without a quick and accurate method of detecting covariate shifts, LE-CPS may suffer catastrophic failure when learning autonomously. Prior work has shown that it is possible to use convex hulls for data analysis and may be fast enough to determine if shifts have occurred on datasets. Such a convex hull system can be used to detect the drop in performance and allow for reorientation to prevent harmful decision making by the autonomous system. Early tests show the reliability of convex hull differences to determine if covariate shifts have occurred and can be explored for other benefits. These benefits may further expand into the realm of computer vision for anomaly detection, where the unique shapes of features may detect adversarial examples.**

Code for this experiment can be found at https://github.com/Kelvin-Sanchez/Convex-Hulls.

***Index Terms*—convex hulls, polygon, area, classifier, machine learning**

I. Introduction

LE-CPS are susceptible to dataset shifts, which could lead to loss in performance. Such a loss in performance on these systems can lead to unwanted behaviors or harmful events for users as seen with medical models that diagnose patients from different populations. Models trained on one population, may not be transferable to another resulting in incorrect diagnoses. To prevent such errors it is essential to have a reliable method of covariate dataset shift detection. Convex hulls may be the solution needed to solve the problem, as they can generate a unique shape for each dataset. Convex hull algorithms enclose a dataset with the smallest possible polygon. The unique shape then allows for direct comparisons between datasets that have undergone covariate shift, and certain metrics may provide insight into what has occurred to the features to cause a drop in performance.

Traditionally, statistical measures are used to determine if a dataset has become shifted; however, they can be computationally expensive due to the large size of datasets. We explored the method of covariate shift detection with a preliminary experiment to show the potential of convex hulls in dataset analysis. In the future, convex hulls may also provide protection to LE-CPS models performing image processing. Using the convex hulls of images or objects may provide a way of detecting anomalous or adversarial examples by taking the convex hulls of objects and then using them as a basis for examining new objects to determine if an adversarial example has been encountered.

II. Problem Definition

Current processes for dataset shift detection may not always capture changes in systems to prevent catastrophic failure. As seen with modern LE-CPS incidents, models are susceptible to failure when encountering new environmental factors and these failures can come at the cost of human life. With [1] models trained on medical datasets were providing inaccurate diagnoses for patients due to the different populations(input datasets). It is therefore important we begin to develop methods of detecting changes in input data that could alter a model's performance. This problem, however, is not limited to features in a dataset, as the same occurs for LE-CPS models trained on image classification. It is not rare today to see news about self driving vehicles crashing into other objects in its environment due to a failure of object recognition. Convex hulls may also provide an effective way of detecting these anomalous objects or events with unique polygons.

III. Related Work

To further understanding of convex hull usage in dataset shift detection previous work was explored. Konstantin et al. [1] showed it was possible to test the transferability of a model trained to detect acute respiratory distress syndrome. Two different population datasets had their corresponding convex hulls calculated then compared. They concluded this using convex hulls of two datasets from different patient populations finding low overlap. A different researcher, Anatoly [2], demonstrated multiple methods of using convex hulls to analyze data for machine learning algorithms. The methods of interest take into account the points caught within a convex hull area of overlap or difference to take approximations of convex hulls to calculate proximity to each other. Both papers provide excellent foundational information for avenues to explore using convex hulls, and their research created the foundation for this continued work.

IV. Experimental Design

A. System Components

- Two dimensional dataset with corresponding labels, in this case the synthetic set generated with the help of Google Gemini.
- Scikit-Learn provides the MLPClassifier model for the proof of concept.

- Shapely allows for the convex hull areas to be measured so the difference can be taken.
- Matplotlib pyplot provides a graphical visualization of the convex hulls and the points of data from the datasets.

B. Equations

As part of the experiment to test whether or not convex hulls can be used for covariate shifts, the synthetic data had to be created using some form of relationship. To keep the experiment as simple as possible, a linear relationship between the wrist size(ws) and finger length(fl) to determine head size(hs) was used. That linear relationship is defined as follows:

$$0.8 * fl + 0.8 * ws = hs \tag{1}$$

C. System Proposal

Models should have a convex hull calculated when they are performing as intended or initially set. This initial convex hull will act as a base or normal shape for the trained model. As new information is learned or added to the dataset, a new convex hull can be generated. This new convex hull can be utilized prior to training or after performance has decreased. The convex hulls can then be compared by taking their difference or utilizing other metrics to find changes to determine what may be occurring on the covariate level. An increase in area of difference should correlate with a drop in performance.

D. Experimental Setup

To begin the experiment, a simple dataset was required for training of a classifier. A synthetic dataset using a linear relationship between wrist size, finger length, and head size generated as defined by (1). With the linear relationship calculations, the resulting value was then binned into groups 'very small', 'small', 'medium', 'large', 'very large.'

These were generated using provided parameters that defined the size of the dataset, the random seed used for the generation of random values, and mean finger length and wrist size. Paired Gaussian distribution sets of 20,000, 25,000, 50,000, and 100,000 were generated with a sibling dataset having a covariate shift. The random seeds for both generated sets were different to avoid generating the same dataset. The finger length and wrist size mean values were altered to shift the set.

A Scikit-learn MLPClassifier model was trained on one of the initial sets of data points and had its convex hull calculated with the use of SciKit-learn and Shapely, which allowed for the calculation of area for a dataset (the same was done for the shifted set). The shift is done by adding or subtracting 2 to the mean of finger length, and doing the same to the wrist size. This was done to examine if the MLPClassifier's performance changes with a shifted input set. To demonstrate the visual representation, Matplotlib was used to display the datasets alongside their corresponding convex hull as seen in Fig. 1. Performance was measured with the original and shifted set to find the difference in classifications for F1 scores, classification summary, and accuracy. Not all measures are provided, as they were inconsistent and provided little value due to instability as will be explained in the results.

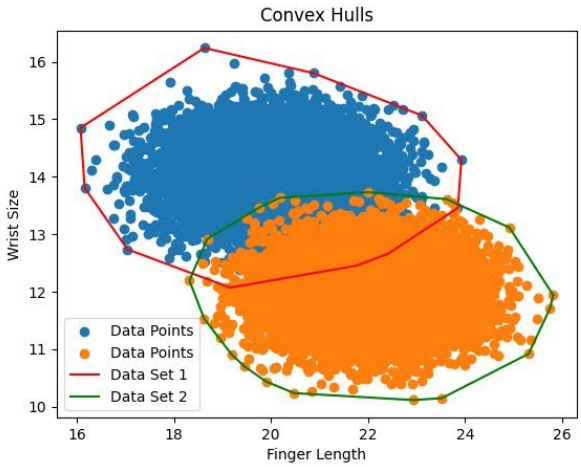

Fig. 1. Convex hulls of original vs shifted dataset.

V. Results

We expected to detect a shift in features after examining a drop in performance on a trained model, and we did observe this in the experiments conducted. Table I shows the various sample sizes generated. The accuracy scores for the MLPClassifier using the unshifted dataset performed around 95% for each set. When tested against the shifted set, performance levels dropped as expected. The convex hulls and their corresponding areas were calculated for both datasets. The difference of areas was calculated by having the shifted dataset convex hull area subtracted from the original datasets'. This measure in area appears to correlate with accuracy decreases; however, a statistical approach should be used to determine the correlations between the two. This correlation calculation was difficult to definitively calculate due to the changing datasets between runs and the outliers inflating the areas.

Convex hulls were accurate in detecting the covariate shift, however, it may not be faster than traditional statistical analysis. A comparison between the two should be conducted to determine if performance is slower, however, that lies outside the current time scope. It may otherwise be beneficial to also use the approximate the proximity between the two convex hulls as mentioned by Anatoly [2] to compete with statistical methods in the case they are faster.

A reproducibility issue was run into with this experiment, aside from the correlation values changing, which may also have resulted from the datasets changing for each run. To combat this, a static dataset can be generated by removing the data_generation function from the test loop or saving the synthetic data to an external CSV file for testing. Another problem faced was the lack of samples at each size or bin in the data_generation function. This lack of samples resulted in precision values of 0 and therefore incorrect results from

Scikit-learn's classification summary. The correlation value, when corrected with an unchanging dataset from run to run, is expected to be between 0 and -1, otherwise known as an inverse relationship. A non-robust relationship can be seen when examining the results table. This inverse relationship is expected because, as the area of difference increases, datapoints exist outside the normal range encountered by the model. This leads to a loss of accuracy or ability to generalize by the MLPClassifier, which causes a decrease in accuracy. When the errors faced with the experiment are corrected, the relationship is expected to be consistent and greater than 0.6.

TABLE I
ACCURACY CHANGES WITH EACH SAMPLE SET

Sample sizes	Initial Acc.	Shifted Acc.	Acc. Delta	CH diff.
20000	97.79	87.31	-10.48	16.97
25000	98.53	91.22	-8.31	16.47
50000	94.83	70.75	-24.09	17.37
100000	95.60	65.59	-30.00	18.55

VI. CONCLUSION

Convex hulls show promise for their use in covariate dataset shift detection. The experiment reached some level of difficulty since the existence of outliers in datasets may cause a large increase in area. This in turn, would create area difference calculations that are not entirely accurate to the cluster of data points as a whole. Outliers then make correlation difficult to measure consistently. The next step would be to determine outliers and remove them from a dataset for accurate area calculation. Even with this setback such a system can be used to determine why LE-CPS may suffer from a degrading of performance as new data is introduced.

VII. FUTURE WORK

One area of interest is the application of convex hulls to image processing as done by computer vision to combat Adversarial Examples. With the unique shape of objects generated, it may be possible to detect anomalies within images by comparing the convex hulls. With object detection, the human body could have its convex hulls measured for feet, legs, torso, arms, and head. If within a dark environment, the lower half of a person is visible, then the convex hulls may provide LE-CPS methods of recognizing known convex hulls, possibly preventing harm.

Currently LE-CPS systems with camera sensors utilize models that examine pixels of images and do not immediately recognize features as humans do [3]. That is to say, current models may use sole pixels as primary features when classifying a new data point, while people, depending on the conditions, use the overall shapes of features to make a classification. When faced with an Adversarial Example, the shift away from the typical classification can be detected by the anomalous shapes encountered by the perturbations faced as compared to typical images with similar classifications. Applying such a method to camera systems in the form of convex hulls may aid in removal of reliance on pixels. Convex hulls should be further explored to determine what benefits they may provide over traditional statistical methods of data analysis and for their applications to image processing.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] A. P. Nemirko, "Convex Hull Proximity Estimation for Machine Learning Problems," Pattern Recognition and Image Analysis, vol. 32, no. 3, pp. 616–621, Sep. 2022, doi: https://doi.org/10.1134/s1054661822030282.

[3] A. Ilyas et al., "Adversarial examples are not bugs, they are features," arXiv.org, https://arxiv.org/abs/1905.02175 (accessed Dec. 3, 2024).