

A Framework for Monitoring Classifiers' Performance: When and Why Failure Occurs?

David A. Cieslak¹ and Nitesh V. Chawla¹

¹Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA

Abstract. Classifier error is the product of model bias and data variance. While understanding the bias involved when selecting a given learning algorithm, it is similarly important to understand the variability in data over time, since even the *One True Model* might perform poorly when training and evaluation samples diverge. Thus, the ability to identify distributional divergence is critical towards pinpointing when fracture points in classifier performance will occur, particularly since contemporary methods such as ten-folds and hold-out are poor predictors in divergent circumstances. This article implements a comprehensive evaluation framework to proactively detect breakpoints in classifiers' predictions and shifts in data distributions through a series of statistical tests. We outline and utilize three scenarios under which data changes: sample selection bias, covariate shift, and shifting class priors. We evaluate the framework with a variety of classifiers and datasets.

1. Introduction

Consider the fundamental task of data mining: given a training sample of data, formulate a model which optimizes some measurement criteria, typically accuracy. This model is then applied to an as yet unseen set of testing examples. Depending on the nature of the data, a practitioner might select a model generated through decision trees algorithms, Bayesian methods, calculating nearest neighbors, or support vector machines. Typically an empirical validation approach is used such as ten-fold cross-validation or leave-one out validation on the training set. Structural risk minimization might be used if the Vapnik-Chervonenkis dimension of the model space is known [22].

Assuming that the expression for the *One True Model* for data is within the set of Turing machines, then it is possible to express a **well-calibrated classifier**: the proper class occurrence rate is mapped correctly for each unseen example. [3] further suggests that any reasonable performance metric should be optimized by this one true model and no other model should yield better performance.

Received October 29, 2007

Revised December 24, 2007

Accepted January 29, 2008

Unfortunately, this task makes several fundamental assumptions, namely the “stationary distribution assumption” [24] in the machine learning literature and “non-biased distribution assumption” [28] in the data mining community.

Definition 1. The **Stationary or Non-Biased Distribution Assumption** [24] states that for each and every training set instance and test set instance is identically and independently drawn from the common distribution $Q(x, y)$.

Previous work [4, 5, 6, 28] has already introduced instances violating this assumption through injection bias in data. In addition to bias, a shift in class priors or a covariate shift may alter the data distribution [20]. In these cases, even the *One True Model* may become irrelevant when applied to future instances should the data distribution change substantially and unpredictably. In this paper, we explore two issues within the context of this problem. First, *can we identify changes in performance attributable to a change in the data distribution?* Second, *can we detect the presence of shift between two distributions of data?*

Generally, we try to determine generalization error based on a training set for a set of classifiers in order to determine which will generally perform best. However, both theoretical and empirical methods can be limited in the presence of such distributional divergences. The structural risk minimization bound established as a function of the VC dimension makes the critical stationary distribution assumption. Thus, implying that the bounds may not hold in the scenarios containing distribution drifts [24, 5]. The empirical methods comprising of ten-fold cross-validation, bootstrap, leave-one out, etc. generate empirical measures on the generalization performance of a classifier. It is obvious that these measures are limited as they are generated from the validation set, which is derived from a similar distribution as the training set. These measures, by no means, reflect the effective generalization in the presence of changes in testing set distributions. This presents the challenge of establishing a landscape of classifiers’ performance across different data assumption. [14] notes the fact that often typical assessment methods cannot produce reliable and unbiased assessments of their predictions’ quality.

Thus, our paper focuses on the following critical components relevant to an application of knowledge discovery and data mining process: a) detection of deviation in the predictive estimates over the testing set as compared to the validation set; b) identification of causes for such a drift in distribution that is what feature(s) are responsible for the testing population to change. We believe these issues are pervasive in the real-world deployment and evaluation of data mining solutions. We choose to compare performance between the validation set and the testing set, as the validation set is utilized at the time of training to estimate the “generalization” performance. The goal then is to see whether the predictions on the unseen testing set in the future are reflective of the observations on the validation set. What is the uncertainty in the predictive estimates?

Contribution This paper outlines a statistical framework, as depicted in Figure 1, to identify the fracture points in predictive distributions and alteration to the feature space, where a fracture is considered as the points of failure in classifiers’ predictions – deviations from the expected or the norm. We consider changes in data distribution by injecting scenarios of sample selection bias, class prior shift, and covariate shift. We approach the problem in two stages. In stage one, we detect whether there is a statistically significant shift in the predictive distributions. We propose to use the Kruskal-Wallis test [7] to isolate a change in data through the distribution of probabilities generated by a learning algorithm. Note that the tests are unsupervised as we will not be aware of the actual testing set classes. Thus, we compare the posterior probability distributions between the validation set and testing set. If this test indicates that there is indeed a change to the

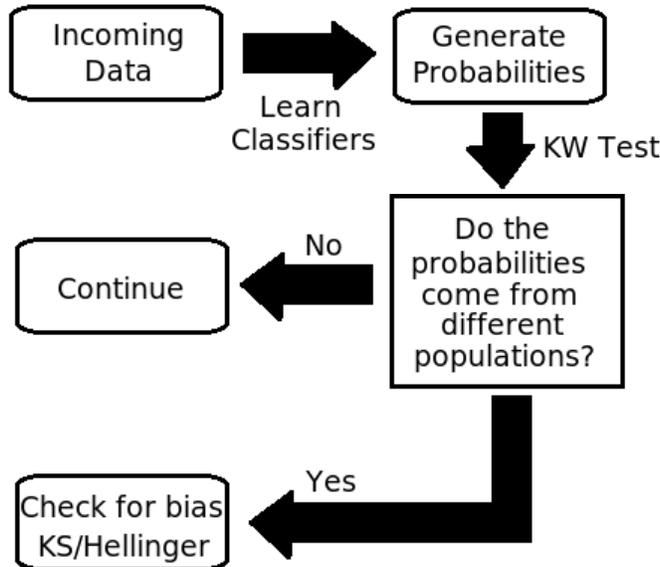


Fig. 1. The proposed Data Drift Identification and Response Framework

predictive distribution, a practitioner may then use a series of unsupervised statistical measures based on the Kolmogorov-Smirnov Test [12, 23] and Hellinger distance [2] to indicate the presence or absence of a change in the feature space. With this information, a practitioner becomes aware of alteration in the data distribution and is equipped to make informed classifier choices or take additional corrective steps. We use four different classifiers and nine different datasets to assess the utility of this framework. Thus, the key questions that we address in the paper are: *a) How to detect fracture in the predictive distributions on the testing set?* and *b) How to detect the feature(s) responsible for the introduction of distributional divergence to the testing set?*

We would also like to point out that this framework can be used to construct a sensitivity index for different classifiers during training. That is, one can simulate different distributional change during validation and observe the variation in the performance of classifiers over these changes. Accordingly, the most generalizable classifier can be chosen, as demanded by an application or domain.

The remainder of this paper is divided as follows: Section 2 defines distributional change and our treatment of the same in this paper. Section 3 describes the datasets and classifiers used in the paper. Section 4 presents a case study on performance of classifiers after distributional change. Section 5 then identifies how changes may be detected between data samples; it also provides a thorough description of the statistical methods used in our work. Section 6 discusses the results of the experiments in this article and Section 7 draws conclusions to the work presented in this paper.

2. Changes in Data Distribution

We use three methods to establish a violation in the stationary distribution assumption. The first of these is *sample selection bias* [4, 5, 10, 28]. Suppose that we consider

examples (x, y, s) drawn independently from a distribution D where the domain is $X \times Y \times S$, with X being the feature space, Y is the class label space, and S is a binary space for which the variable s indicates the example is in the training when $s = 1$ and is not in the training set when $s = 0$. Operating in this environment, the following cases emerge regarding the dependency of s on (x, y) [10, 17].

Definition 2. The **missing completely at random (MCAR)** sample selection bias occurs when s is independent of both x and y . We thus state that $P(s = 1|x, y) = P(s = 1)$, thus the sample bias depends on a factor totally independent from the feature vector x and class label y . This implies that the training and testing sets are derived from the same distribution. The stationary distribution assumption theoretically holds under MCAR, but we include it in our paper for completeness.

Definition 3. Sampling bias is **missing at random (MAR)** if s depends on x but conditional on x is independent of y , thus, we may state $P(s = 1|x, y) = P(s = 1|x)$. Therefore, sampling is feature dependent as the sampling probability varies according to the feature vector x , but is independent to the class label y . This situation can occur if the testing set is thresholded on one or more known features.

Definition 4. **Missing not at random (MNAR)** bias occurs when there is no independence assumption between x, y , and s . This scenario essentially introduces the sample selection bias, as the cause of distributional shifts may be unknown. That is, one may not have access to the feature leading to the censoring in the dataset. We may state the tautology of $P(s = 1|x, y) \neq P(s = 1|x)$. Thus, at any particular feature x , the distribution of observed y in the training set is different from the observed y in the testing set: $P(y = 1|x, s = 1) \neq P(y = 1|x, s = 0)$.

We establish the biases as follows. For MCAR, the examples are removed uniformly at random; we remove 25% and 50% of the examples in the testing sets. We also use MAR and MNAR by removing the top 25% and 50% of values along one feature. We first sort the dataset based on one particular feature and then remove the top 25% or the top 50% of examples conditioned on that particular feature value (thresholding). In the case of MNAR, the remaining examples have the chosen feature masked as ‘‘Unknown’’ or missing. By masking the feature as unknown or missing, we are able to inject the ‘‘latent’’ MNAR bias. We generate separate MAR and MNAR biased testing distributions for each feature within the dataset. The reported results are aggregates to indicate the ‘‘average case’’ arising from introduction of biases. For fairness, an equivalent number of MCAR samples were generated; thus, MCAR results are similarly aggregated. [8] also studies classifier performance as a function of input quality degradation; however, their analysis does use very different methods to alter data distributions. [8] also centers on Knowledge-based systems which are designed to mimic human problem-solving through artificial intelligence using a stored information on a particular subject. This study focuses on classifiers, rather than Knowledge-based systems.

In Figure 2 we view a single, hypothetical data distribution at $t1$. This snapshot is the first that is observed and is used in training a classifier. In the left snapshot, we view the distribution of data across a single feature as separated two classes. In the right snapshot, we view the totality of the data distribution. At some point $t2$, which is after $t1$, the data distribution in Figure 3 is observed. We note that the difference between Figures 2 & 3 is a result of feature bias. Here, we note that at $t2$, the top quarter of values has been removed from the distribution, yielding a ‘‘cut’’ in the sample. This affects the overall shape of $P(X)$, as seen in the right snapshot of $t2$.

The second method used to induce distributional change is *covariate shift* (CS) [20]. The previous method affects how examples in the test set are sampled from the data

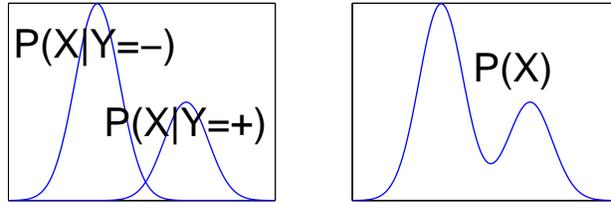


Fig. 2. Depiction of the original data distribution at the initial time t_1 .

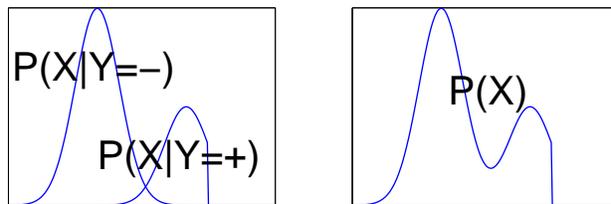


Fig. 3. Depiction of the data distribution after it has undergone feature bias at t_2 .

distribution, but here those examples are fundamentally altered. To incorporate covariate shift, we begin by calculating the standard deviation for a single feature of a dataset. The value for this feature increased by N percent for each testing sample example. Thus, CS_{25} has increased each testing example by 25% of one standard deviation. This article uses $N \in \{25, 50, 100, 200\}$. As with sample selection bias, the results reported in this study are an aggregate by applying covariate shift to each feature to represent the “average case” arising from the introduction of covariate shift. We formally define covariate shift as follows.

Definition 5. Covariate shift (CS) occurs when the data distribution generating the feature vector x and its related class label y changes as a result of a latent variable t . Thus, we may state that covariate shift has occurred when $P(y|x, t_1) \neq P(y|x, t_2)$.

In Figure 4, we observe how the original data distribution at t_1 in Figure 2 changes to the shape of the data distribution at t_2 as a product of covariate shift. We note how distribution has shifted to its new position at the solid line relative to its original position

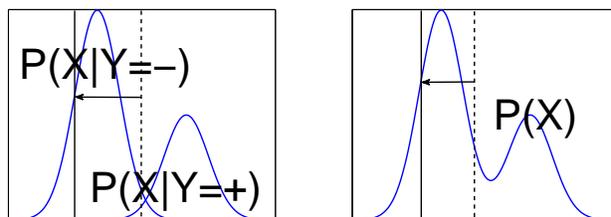


Fig. 4. Depiction of the data distribution after covariate shift at t_2 .

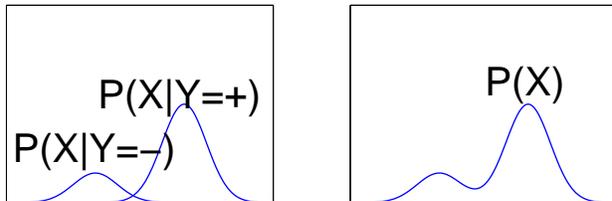


Fig. 5. Depiction of the data distribution after shifting priors at $t2$.

at the dashed line. While we note the forms of the joint distributions remain unaltered, their absolute position has changed significantly.

The third method used to violate the stationary distribution assumption is *shifting priors*. Unlike the previous two methods, which alter the feature space directly and the class distribution only indirectly, this method changes the class distribution directly and the feature space only indirectly. Random undersampling is used to induce varying class skew ratios, such that $P(+)=\{0.02, 0.05, 0.1, 0.2, 0.3, \dots, 0.7\}$. For example, a hypothetical dataset has 2000 examples from class $-$ and 1000 examples of class $+$ in the testing set. To evaluate on $P(+)=0.5$, 1000 class $-$ examples are randomly removed from the evaluation set. We formally define shifting priors as follows.

Definition 6. Shifting priors occurs when s depends on y but the conditional on y is independent of x , thus, we may state $P(s=1|x,y)=P(s=1|y)$. Therefore, sampling is dependent on the class label and independent of the feature vector x .

We note the effects of shifting priors from $t1$ to $t2$ by comparing Figures 2 & 5. In the left snap-shots of $t2$, we view the distribution of data across a single feature as separated two classes and note that the relative balance of classes has reversed between $t1$ and $t2$. The right snap-shots contain the joint distribution of all examples and we note that affecting the ratios of $P(+)$ to $P(-)$ yields an indirect, yet significant shift in the distribution of $P(X)$ generated.

3. Datasets and Classifiers

This paper uses several common UCI [18] and real-world datasets, summarized in Table 1. These datasets vary extensively in both size and distribution, offering many different domains. Page, Pendigits, Phoneme, Satimage, and Segment come from the UCI Machine Learning repository [18]. The Oil dataset contains a set of oil slick images based on live data [13]. Compustat represents real world finance data and may contain natural bias as the training and testing samples come from different two-year periods, while Mammography comes from studying calcifications in the medical domain [25]. E-State consists of electrotopological state descriptors for a series of compounds from the National Cancer Institute’s Yeast AntiCancer drug screen [9].

For the experiments conducted, we used C4.5 Decision Trees, Naive Bayes, k -Nearest Neighbor (where $k=5$), and Support Vector Machines. Each classifier formed probability estimating models. Decision trees were trained as Probability Estimation Trees (PETs) [19]. k -Nearest Neighbor formed predictive probabilities as the proportion of the classes for the set of nearest neighbors. For SVM, the SVM^{light} software [1] was used with default parameters to form probabilistic predictions. Naive Bayes naturally forms probabilities. We restrained ourselves to default parameters for all classifiers

Dataset	Examples	Features
Compustat	(7,400, 2,958, 3,299)	20
E-State	(2,662, 1,064, 1,596)	12
Mammography	(5,593, 2,236, 3,354)	6
Oil	(470, 188, 279)	49
Page	(2,738, 1,094, 1,641)	10
Pendigits	(5,497, 2,198, 3,297)	16
Phoneme	(2,702, 1,081, 1,621)	5
Satimage	(3,218, 1,287, 1,930)	36
Segment	(1,155, 462, 693)	19

Table 1. Datasets used in this study. Column Examples indicates the number of examples given as (*training, validation, testing*)

to establish an even playing-field. These algorithms are among the most influential data mining algorithms in the research community [26].

4. Case Study: Impact of Distributional Shifts

Various factors can be responsible for introducing distributional divergences in the testing set. The feature space could be biased through a number of methods, causing the classifier to generate inappropriate predictive distributions. In some cases, divergence occurs as a result of collecting separate sub-populations governed by independent feature and class probability density functions within a single distribution. An example is the frequencies of measured wingspans of one species of bird found on two independent tropical islands. Temporal distance may also incorporate divergence: the rules governing data may change slightly or drastically over time. Such divergences can occur in various applications such as marketing and credit scoring, as the targeted population can change over time.

We now present a case study across classifiers and different datasets to demonstrate the effect of changes on the testing set. We use the Friedman test to statistically validate whether the predictions in the testing set start to significantly differ from the validation set once the different distributional shifts, as defined in the previous Section, are introduced. We will discuss the Friedman test before presenting our results. The goal of this section is to establish the premise of the framework that the classifiers behave differently when distributions change in the testing set.

4.1. Friedman Test

The Friedman test is a non-parametric statistical test developed by the U.S. economist Milton Friedman [16]. The Friedman test is used for two-way analysis of variance by ranks. This two-way test assumes that all data comes from populations having the same continuous distribution, apart from possibly different locations due to column and row effects and that all observations are mutually independent. An example Friedman test

evaluation is of n welders using k welding torches with the ensuing welds were rated on quality. Is there one torch that produced better welds than the others?

X is a matrix such that observances are placed in columns and samples are stored across rows. $r(x_{ij})$ is then the rank within block (i.e. within its row). The average rank per sample is calculated as

$$r_j = \sum_{i=1}^k r(x_{ij}) \quad (1)$$

k is the number of samples and n represents the number of examples in each sample. With the above ranking, calculate the following:

$$\chi^2 = \frac{12}{kn(k+1)} \sum_{j=1}^k r_j^2 - 3n(k+1) \quad (2)$$

with χ^2 as an associated p-value. This is the p-value for the null hypothesis that the column medians are essentially the same. When the p-value is very low, this indicates that this is likely not the case and the null hypothesis is void.

To apply Friedman, we begin by first randomly partitioning the dataset into 50% for training, 20% for validation, and 30% for testing. Each classifier learned on the corresponding training set is then applied to the natural validation and testing samples, resulting in probabilistic predictions on both sets. This formed the Base results for the stationary distribution assumption, that is both validation and testing sets were derived from the same distribution.

4.1.1. Sample Selection Bias

We introduced the three biases — MCAR, MNAR, and MAR — as follows. Considering a feature for each data set at a time, we injected the corresponding amounts of biases as discussed in the previous section. This resulted in as many testing sets as the number of features for each dataset and bias combination. This allowed us to avoid the dominance of results by any one feature in particular. We applied the same classifiers learned on the training set to each of the generated biased testing sets resulting in probabilistic predictions.

Then, we formed 100 bootstraps on each (validation and testing) set of probabilistic predictions for each dataset and calculated accuracies on each. The Friedman test was then used to test the null hypothesis: there is no statistically significant difference between the validation and testing set accuracies for a dataset. Figure 6 shows the resulting p-values. The p-values for a given amount of bias are the averages of the p-values from the application of that particular bias to each feature in the dataset. Thus, it reflects the summarized p-value given a bias, dataset, and classifier. The convention in the figure is: the x-axis domain shows the different testing biases, including the Base stationary distribution. Each bias has a cluster of four lines representing the different classifiers. The y-axis shows the range of p-value across all the datasets for each classifier. As the p-value decreases, the hypothesis is more strongly rejected.

Figure 6 shows a compelling trend. If we run along the x-axis, we observe that the range drops as we go more towards heavily biased testing sets. This confirms the premise that the performances of classifiers will suffer in non-stationary environments. Among the classifiers, decision trees and k -nearest neighbor seem to be less sensitive

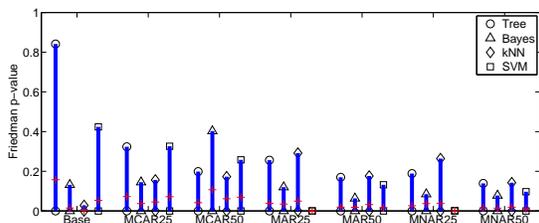


Fig. 6. Friedman Test p-values across all datasets undergoing bias. +'s represent the average p-value.

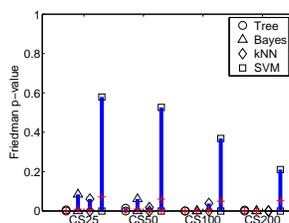


Fig. 7. Friedman Test p-values across all datasets undergoing covariate shift. +'s represent the average p-value.

to distributional biases as compared to SVM and Naive Bayes. Since the y-axis reflects the range over datasets, we observe that some datasets lead to a complete failure of predictive estimates (p-value of approximately 0). Nevertheless, within 85% confidence all the classifiers fail for all the datasets at MNAR-50. This is a strong demonstration of the fragility of classifiers in changing distributions, hence the forming the main motivation of our work.

4.1.2. Covariate Shift

We test for a performance fracture from covariate shift in an identical fashion as before. The covariate shift is imposed on each feature individually; thus, there are as many covariate samples as generated under the sample selection bias experiments. We generate a 100 bootstrap sample for each and test according to Friedman, forming an aggregate across all features. Figure 7 shows a similarly compelling trend. Moving right along the x-axis reflects an increase in the degree of covariate shift. As with bias, this confirms the premise that classifier performance suffers in a non-stationary environment. Additionally, performance degrades as the degree of shift increases. Throughout, SVM appears the most resilient to covariate shift. Under moderate shift, such as CS25 and CS50, Naive Bayes performs well on some datasets. We note that in general, the classifiers are failing (i.e. have a low p-value) under covariate shift. While SVM and Naive Bayes are resilient to some instances of covariate shift, the average case indicates that overwhelmingly a classifier will fail under covariate shift.

4.1.3. Shifting Priors

To incorporate a shift in class priors, we use random undersampling on the testing distribution to impose a desired class skew by selecting a $P(+)$ \in $\{2, 5, 10, 20, \dots, 70\}$. 100 bootstraps of probabilities are generated, which in turn form populations of 100 accu-

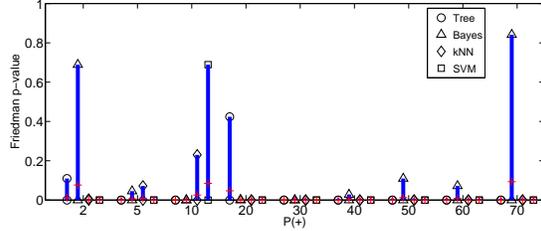


Fig. 8. Friedman Test p-values across all datasets shifting class priors. +’s represent the average p-value.

racies. These sets of accuracies are then compared using the Friedman test. Figure 8 shows the effects of prior shift on the Friedman values for produced accuracy. The far right represents distributions with the highest $P(+)$. As before, the premise is confirmed that the performances of classifiers will suffer in non-stationary environments. All classifiers suffer heavily as the $P(+)$ increases. This is intuitive, since accuracy is partially class driven. Classifiers tend to perform less accurately on small classes; thus, the relative class balance reversal dictates that the accuracies should generally be driven down as $P(+)$ increases. The highest p-values are returned for $P(+)$ values of 2, 5, and 10. As we have observed in Table 1, these values are nearest to the original class distributions. This matches the intuition that produced accuracies should be similar when the training and testing distributions, and hence $P(+)$ values, are similar. We do notice an odd fluctuation in some of the p-values from Naive Bayes when $P(+)$ has the values of 50, 60, and 70. We note that these values occur from the Compustat, E-State, Oil, and Phoneme datasets and are 0 on all other datasets. In these cases, performance is conserved either fortuitously or because the distributions are relatively similar, despite the shift in priors. This indicates that there is really no cause of alarm for these data sets, even if the priors have shifted. That is the classifier is performing as expected, even if the testing set is now carrying a distributional bias towards the (former) minority class. Our results in the subsequent sections assert this point, establishing a keen highlight for this framework. This further demonstrates how classifiers might fracture as a result of a changing distribution, and it is highly data dependent. Thus it is increasingly important to have such a framework in place.

We note that one can directly use this framework to induce distributional change during the validation process. This can result in an immediate evaluation of sensitivity of different classifiers as the population drifts. Then, conditioned on the nature of the application, one can then choose a classifier that is most consistent, perhaps at the cost of some accuracy at the stationary distribution.

5. Detecting and Identifying Changes in Distribution

The goal of this work is to apply unsupervised methods to detect drifts in predictive and data distributions. Unsupervised methods are required as the class of testing data is presumed to be unknown at the time of prediction. The following subsections provide tests for finding bias through three separate tests: the Kruskal-Wallis Test in Section 5.1, the χ^2 test for nominal features in Section 5.2 and the Kolmogorov-Smirnov test for continuous features in Section 5.3, and Hellinger Distance in Section 5.4. Together, they provide a statistical framework as shown in Figure 1. We split the original data into the 50 : 20 : 30 training, validation, and testing proportion, respectively, as described

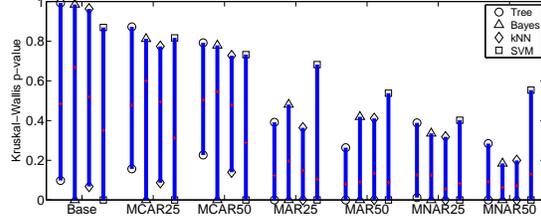


Fig. 9. Kruskal-Wallis Test p-values across all datasets undergoing selection bias. +’s represent the average p-value.

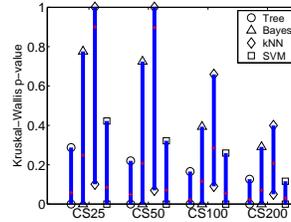


Fig. 10. Kruskal-Wallis Test p-values across all datasets after covariate shift. +’s represent the average p-value.

before. We introduce sample selection bias, covariate shift, and shifted classed priors to form a variety of testing samples. The results in Sections 5.1, 5.3, and 5.4 all represent the average values found across bias on all features. This reflects the “average case” feature becoming biased in a particular dataset.

5.1. Kruskal-Wallis Analysis of Generated Probability Estimates

Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method for testing equality of population medians among groups [7]. Unlike One-way ANOVA, no assumption regarding a normal distribution is made since the test is non-parametric. There is also no assumption that the population variables between compared groups are the same. This test calculates the following statistic

$$K = (N - 1) \frac{\sum_{i=1}^g n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^g \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2} \quad (3)$$

where n_g is the number of observations in group g , r_{gj} is the overall rank of observation j in group g , N is the total number of observations, \bar{r}_g is the average rank of the observations within group g , and \bar{r} is the average rank of all observations. The p-value is then calculated as

$$\Pr(\chi_{g-1}^2 \geq K) \quad (4)$$

This is the p-value of the null hypothesis that all samples are drawn from the same population or different populations of the same distribution. Therefore, this is a very useful test for determining if sets of probabilities are drawn from the same or different

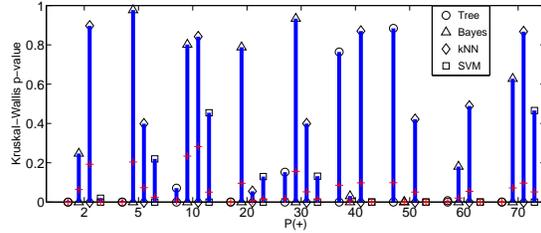


Fig. 11. Kruskal-Wallis Test p-values across all datasets when priors shift. +’s represent the average p-value.

distributions. Here it is applied as a comparison of the probabilities estimated on the validation set against the natural testing distribution and the six other biased distributions.

5.1.1. Sample Selection Bias

In Figure 9 we observe the calculated set of Kruskal-Wallis p-values under sample selection bias. Those generated in comparing the set of validation probabilities against those of the testing set and distributions formed through MCAR are quite similar, which is expected as there is similarity between the validation sample and the testing and completely randomly biased testing samples. However, there is a substantive difference to the MAR and MNAR biased sets. Under these sophisticated biases, the distribution of probability estimates differs significantly. With such a drastic change in the estimates, there should follow a fairly substantial change in the classifier performance. We also note that the values captured through Kruskal-Wallis are quite correlated to those found under the supervised (determining accuracy and rank-order requires known classes) Friedman test (Figure 6).

5.1.2. Covariate Shift

Figure 10 observes the KW p-values during covariate shift. As the shift increases, we note that there is a general drop in p-value. It is evident that as the covariate shift increases, the classifiers start shifting in the predictive distributions. Thus, it is increasingly likely that the pool of probabilities are derived from different populations since the data distribution is moving relative to the trained decision boundaries. Different classifiers are behaving differently again, and the range of the p-values indicates that the reaction of classifiers to different datasets is also very different. Thus, it is critical to instrument a monitoring framework for a dataset and a chosen classifier as the fracture points may occur at different points.

5.1.3. Shifting Priors

Additionally, we note that KW is quite effective in identifying a shift in class priors, as seen in Figure 11. While there are some odd spikes in the p-values as $P(+)$ increases, we note that the average case will warrant further investigation when $P(+)$ ≥ 20 , since the typical p-value is very low, if not zero. As seen in Figure 5, the relative occurrences of examples as a product of the feature vector can change drastically under a prior shift. Assuming that the classifier produces a relatively diverse set of probabilities, it then follows that the population of generated probabilities should change significantly as the distribution in feature space changes. KW can ascertain these differences between

populations and is therefore effective in identifying a change in distribution as caused by shifting class priors. We do note that there are some spikes in p-values, which correlate with the previous Friedman results in Figure 8. This shows that not all classifiers are failing under the same conditions for all datasets. That is, the data properties dictate different sensitivities in predictions. Again the same set of datasets (Compustat, E-state, Oil, and Phoneme) and classifiers that we observed in Friedman test demonstrate their robustness for shifting priors. As we will see under KS and Hellinger, on these datasets there is a distributional similarity in terms of the feature space. These leads to a similar set of predicted probabilities and the conserved results under Friedman.

5.1.4. Summary

With this information, it is both feasible and useful for the practitioner to initially train a model and predict probabilities on both the validation and testing data samples. Using Kruskal-Wallis, the practitioner may then determine whether the the sets of probabilities came from different populations. If so, it is then wise to use the tests in Sections 5.3 and 5.4 to attempt to determine bias type and isolate biased features.

5.2. χ^2 Test

χ^2 is a statistical test used to compare observed nominal data. This is useful in determining whether the distribution of observations within categorical data are dissimilar.

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^v \frac{\frac{n_{i,j}}{N} - \hat{n}_j}{\hat{n}_j} \quad (5)$$

when there are p populations and v values, $n_{p,v}$ represents the count of value v in population p , n_p is the count within population p , $N = \sum_{k=1}^p n_k$ and $\hat{n}_j = \sum_{k=1}^p n_{k,j} n_k / N$. We note that as we are comparing two distributions, $p = 2$. To determine a p-value with this test, degrees of freedom are also considered as

$$df = (p - 1)(v - 1) \quad (6)$$

Based on the found values of χ^2 and df , a look-up table is then used to determine a p-value. With this test, we may determine an appropriate p-value for nominal features.

5.3. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (often called the KS test) determines if there is divergence between two underlying one-dimensional probability distributions or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples [12, 23].

The two-sample KS test is particularly useful as a general nonparametric method of comparing two sample distributions as it detects divergence in both location and shape of the observed distribution functions. KS has an advantage over other statistical methods in that it makes no assumption on the distribution of data, which other methods such as Student's t -test make. However, other methods may be more sensitive if the distributional assumptions are met.

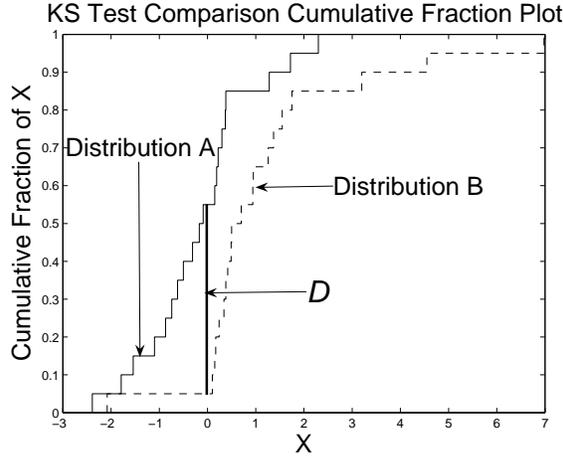


Fig. 12. An example KS test plot. Here, the distributions are significantly divergent as $D = 0.5$.

Dataset	Base	MCAR25	MCAR50	MAR25	MAR50	MNAR25	MNAR50
Compustat	0.250	0.250	0.240	0.385	0.517	0.353	0.492
E-State	0.000	0.000	0.007	0.125	0.174	0.083	0.136
Mammography	0.000	0.028	0.028	1.000	1.000	1.000	1.000
Oil	0.041	0.006	0.010	0.260	0.390	0.246	0.379
Page	0.300	0.190	0.100	0.560	0.590	0.511	0.544
Pendigits	0.000	0.004	0.027	0.461	0.523	0.425	0.492
Phoneme	0.000	0.000	0.000	0.520	0.600	0.400	0.500
Satimage	0.000	0.000	0.000	0.875	1.000	0.871	1.000
Segment	0.053	0.022	0.033	0.421	0.446	0.395	0.421

Table 2. Proportion of features failing the KS test at 95% confidence

Quite simply, KS makes use of a plot of the Cumulative Fraction Function. Suppose we have two distributions, such that $A = \{0.34, 0.94, 0.24, 1.26, 6.98, 0.95, 0.15, -2.08, 0.17, 1.55, 3.20, 0.50, 0.70, 4.55, 0.10, 0.49, 0.38, 0.42, 1.37, 1.75\}$ and $B = \{0.15, -0.62, -0.17, -0.31, -0.50, 0.38, 2.30, 0.37, -1.79, -0.87, 1.72, -0.09, -1.54, 0.30, -2.39, -0.74, 0.22, 1.28, 0.19, -1.10\}$. The KS test begins by sorting both sets of values independently. A single plot of both distributions is then generated. The x-axis contains the values of distribution. For each point x , the y-axis is calculated as the percentage of instances strictly smaller than x ; hence, it is the cumulative fraction of the data which is smaller than x . Figure 12 contains a plot of A against B . Using a planesweep, the KS test then calculates the maximum vertical deviation between the two distributions. For A and B , Figure 12 indicates this as D . In this case, the maximum vertical deviation is 0.5. We would like to state whether this value represents a significant distance. We calculate

Dataset	MCAR25	MCAR50	MAR25	MAR50	MNAR25	MNAR50
Compustat	0.046 ± 0.003	0.053 ± 0.004	0.391 ± 0.003	0.217 ± 0.017	0.361 ± 0.026	0.250 ± 0.021
E-State	0.000 ± 0.000	0.000 ± 0.000	0.175 ± 0.021	0.128 ± 0.016	0.104 ± 0.013	0.057 ± 0.007
Mammography	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Oil	0.001 ± 0.000	0.004 ± 0.000	0.554 ± 0.004	0.388 ± 0.002	0.476 ± 0.003	0.400 ± 0.003
Page	0.114 ± 0.028	0.156 ± 0.014	0.461 ± 0.069	0.399 ± 0.054	0.554 ± 0.069	0.479 ± 0.052
Pendigits	0.000 ± 0.000	0.069 ± 0.006	0.405 ± 0.003	0.362 ± 0.020	0.453 ± 0.001	0.412 ± 0.013
Phoneme	0.000 ± 0.000	0.000 ± 0.000	0.467 ± 0.016	0.349 ± 0.026	0.783 ± 0.069	0.650 ± 0.074
Satimage	0.000 ± 0.000	0.000 ± 0.000	0.037 ± 0.001	0.000 ± 0.000	0.051 ± 0.001	0.014 ± 0.000
Segment	0.002 ± 0.000	0.009 ± 0.001	0.311 ± 0.003	0.235 ± 0.011	0.235 ± 0.014	0.249 ± 0.006

Table 3. Average ϕ -correlation for feature failure under bias.

$$\chi^2 = \frac{4D^2n_1n_2}{n_1 + n_2} \quad (7)$$

where n_1 and n_2 are the number of examples in the two samples. Using $d = 2$ and the χ^2 calculation, the resultant p-value suggests whether there is a significant difference between the two distributions and may be compared against a desired confidence level. Within the context of monitoring classifiers' performance, we may use the KS test to determine if there is a significant distributional difference between the training and testing distributions for continuous features. When features are nominal, a χ^2 test is instead applied to determine p-value.

To do so, we must iterate through both distributions on a feature wise basis, and tabulate the number of failing features, which is why using Kruskal-Wallis on the probability distributions is a better first step.

5.3.1. Sample Selection Bias

Table 2 represents the proportion of features failing the KS test under each bias. Based on these results, we observe that Compustat, Page, and Segment contain some degree of natural bias between training and testing distributions. Of these, Compustat is the least surprising as its training and testing data come from two independent sets of financial information covering separate and sequential two year periods. For these three datasets, it is noted that MCAR actually reduces the failure proportion somewhat, likely because there are unusual values creating large maximum separations. The random bias removes these values and reduces the separation, hence dropping the feature failure rate. In the remaining datasets, MCAR very minimally increases the feature failure rate, if at all. It is observed that the more systematic biases MAR and MNAR increase the feature failure rate substantially. This indicates that the KS test may be used simply and quite effectively to detect a bias incorporated between two data distributions.

In addition to understanding the degree to which bias causes feature failure under the KS test, we seek to study the interaction of a particular feature failing on other features. This is important as the bias is applied to group of features simultaneously or the MNAR could actually be affecting a collection of features together. Restated,

Do features tend to fail independently or concomitantly? It is important to answer this for the sample selection bias scenario. To this end, a Failure Correlation Matrix F was constructed where $F_{i,j}$ represents the count for which features i and j fail under KS concomitantly. Based on the counts within F , the ϕ -correlation is calculated for each pairwise set of features as

$$\phi = \frac{F_{i,i}F_{j,j} - F_{i,j}F_{j,i}}{\sqrt{(F_{i,i} + F_{i,j})(F_{i,j} + F_{j,j})(F_{i,i} + F_{j,i})(F_{j,i} + F_{j,j})}} \quad (8)$$

as ϕ is a strong measure of the associativity of two dichotomies and discounts the effects of sample size. The average correlation per pairwise comparison is reported in Table 3. Values between 0.0 and 0.3 are considered to have little to no associativity, 0.3 to 0.7 have some associativity, and above 0.7 has very strong associativity.

The average ϕ -correlation is quite low, if not zero, for the baseline comparison and MCAR. Thus, there is little correlation between the failure of features, if failure occurs at all. As MAR and MNAR are introduced, there is a spike in ϕ -correlation. This is an expected result as there is some degree of covariance among the measured features; thus, a bias on one feature will to some degree incorporate a bias to related features. The exception to this trend is Mammography, which reports zero correlation categorically, as within each test either all or none of the features fail the KS test except for some MCAR trials for which failure occurred totally at random.

Once bias is suspected through the Kruskal-Wallis test on the set of predicted probabilities, the KS Test operates as a “quick” method to check for the existence of bias to see if a fairly high proportion of the features fail this test (in most cases, 30% feature failure appears to be a reasonable point to presume some bias as observed in Table 2). Table 3 reported the ϕ -correlation of the KS Test as capable of determining groups of features which tend to fail together. Suppose there is a high correlation of failure between two features. In the case that only one fails, one may assume a reasonable correlation between the two features and omit the failing feature during model training confident that the succeeding feature will account for much of the information contained within the failing one. As seen in Table 2, the KS Test struggles to isolate individual biased features. Thus, it is a good method to confirm the findings from Kruskal-Wallis. To more acutely determine degree and which features are biased, we turn to Hellinger Distance, as discussed in Section 5.4.

5.3.2. Covariate Shift

We now consider the covariate shift scenario. The proportion of features failing in this case are depicted in Table 4. We note that despite the relative increment (25%, 50%, 100%, or 200%), that this proportion remains static, a promising result. Due to the nature of the shifting mechanism, only one feature within the feature vector is altered per experiment. The increase in proportion of features failing in Table 4 relative to the baseline failure rate, as reported in Table 3 indicates that on average one additional feature is failing – unless that feature fails in the baseline. Therefore, the KS Test is quite sensitive to even minor covariate shift and is invaluable for detecting this form of distributional change.

5.3.3. Shifting Priors

Next, we apply the KS test to the changing priors scenarios. To do so, we calculate the KS p-value between training and evaluation samples along each feature. The proportion

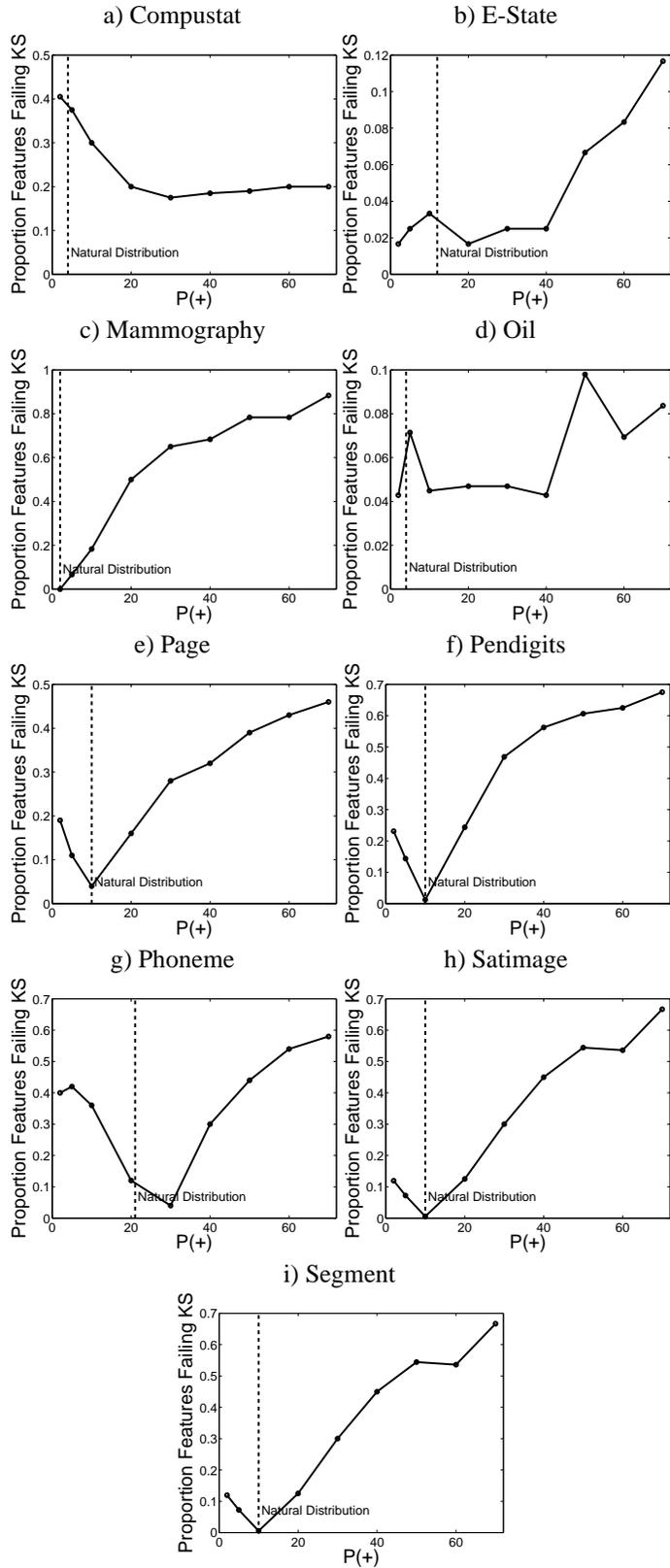


Fig. 13. Proportion of Features Failing KS as $P(+)$ changes.

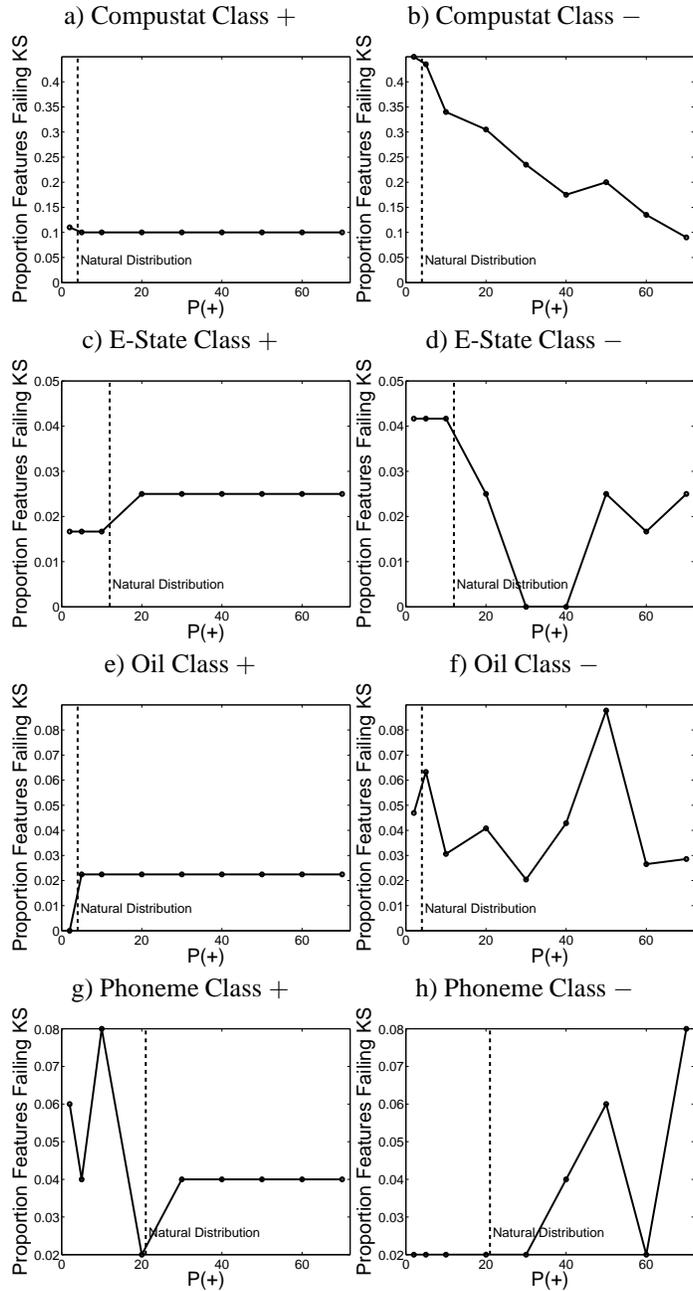


Fig. 14. Proportion of Features Failing KS as $P(+)$ changes, by class.

Dataset	CS 25	CS 50	CS 100	CS 200
Compustat	0.287	0.287	0.287	0.287
E-State	0.083	0.083	0.083	0.083
Mammography	0.167	0.167	0.167	0.167
Oil	0.060	0.060	0.060	0.060
Page	0.370	0.370	0.370	0.370
Pendigits	0.062	0.062	0.062	0.062
Phoneme	0.200	0.200	0.200	0.200
Satimage	0.028	0.028	0.028	0.028
Segment	0.102	0.102	0.102	0.102

Table 4. Proportion of features failing the KS test at 95% confidence

of features failing KS (i.e. $p\text{-value} < .05$) is depicted for each dataset in Figure 13. Here, we observe that the proportion of features failing KS tends to be the lowest when the relative priors most resemble the original distribution. As the priors diverge from the original, the proportion of features failing KS increases.

The exceptions to this occur in Compustat, E-State, Phoneme, and Oil. We examine these datasets further in Figure 14 by analyzing feature failure as a product of class. In these cases, we note that the change in features failing as a product of the positive class (Figures 14(a), (c), & (e)) is relatively small as compared to the negative class (Figures 14(b), (d), & (f)). This indicates that at low values of $P(+)$, the KS failure is primarily driven by the negative class. As $P(+)$ increases, the relative importance of the negative class is likewise reduced, which in turn leads to an increased impact on KS by the positive class. This explains the reversal in trend for Compustat and Oil in Figure 13. Additionally, Compustat, E-State, Phoneme, and Oil occasionally yield high p-values for KW at high values of $P(+)$, which corresponds to a similar conservation in the Friedman performance. This implies that there is no reason for alarm for these data sets, as the class demonstrating the distribution shifts is now undersampled in the testing sets, and does not induce failure of classifiers' performance.

5.3.4. Summary

We have thus demonstrated how the Kolmogorov-Smirnov Test may be used effectively in identifying the proportion of features which are significantly different within two data samples. In the case of bias, a more difficult bias usually causes a greater proportion of features to fail KS. In addition, KS may be combined with ϕ -correlation to determine how features fail independently and concomitantly. This test is also extremely sensitive to covariate shift. Finally, KS can be used to detect shifting class priors, although this method is susceptible when feature failure is a product of class, such as in Compustat, E-State, Phoneme, and Oil.

	<i>a</i>	<i>b</i>	<i>c</i>
<i>Pop1</i>	7	0	0
<i>Pop2</i>	0	10	2

Table 5. Example population data

5.4. Hellinger Distance

Hellinger Distance [2], also referred to as Bhattacharyya Distance [11], is a measure of distributional divergence. [15] concludes that for linear ordination, the Hellinger Distance offers a better compromise between linearity and resolution, as compared to similar metrics such as the χ^2 metric and the χ^2 distance. Hellinger distance has been used effectively within the ecological domain and is recommended for clustering or ordination of species abundance data [21]. This measure has also been used as a means of locating statistical outliers for fraud detection in insurance applications[27].

To apply this measure of density, we presume two independent distributions of data X and Y . Both X and Y contain p bins, where each bin contains the count of some logical subunits measured between X and Y . The Hellinger Distance between X and Y is then calculated by

$$Hellinger(X, Y) = \sqrt{\sum_{j=1}^p \left(\sqrt{\frac{X_j}{|X|}} - \sqrt{\frac{Y_j}{|Y|}} \right)^2} \quad (9)$$

Suppose that there exist two populations, *Pop1* and *Pop2*. The occurrence count of value a , b , and c within each population have been tabulated and are reported in Table 5.

Using (9), $Hellinger(Pop1, Pop2) = \sqrt{2}$, which happens to be the maximum possible Hellinger Distance. This is expected as *Pop1* and *Pop2* are completely divergent: there is no overlap in values a , b , and c .

Here, we outline a method for using Hellinger distance to quantify distributional divergence. Here, *Pop1* and *Pop2* are the Training and Testing distributions and the count values are feature value counts. In the case of continuous features, the Training and Testing sets are temporarily combined to construct 30 equi-width bins. Using (9), Hellinger distance is calculated for each feature and the average across all features is considered as the Hellinger distance. This method is outlined in Algorithm 1.

In these experiments, we first calculated the distances between both the training set features and the original testing set, which we call the baseline distance. This reflects the actual distances between the training and testing set features without any bias injection. Then, we introduced the different distribution changes in the testing set and calculated the corresponding distances from the training set features. We are primarily interested in the relative changes in the distances of the altered testing sets from the baseline.

5.4.1. Sample Selection Bias

The observed relative changes as a product of bias are summarized in Figure 15. The calculated distances tend to be relatively low between the base training and testing distributions and testing distributions generated through MCAR. There is a substantial increase in Hellinger Distance when an MAR or MNAR is at play. Thus, applying

Algorithm 1 *Calc_Hellinger***Require:** Number of bins b , Training set T , and Testing set E

```

1:  $Hellinger = 0, n_f =$  number of features in  $T$  &  $E$ 
2: for each feature  $f_i$  of  $T \& E$  do
3:    $h \leftarrow 0$ 
4:   if  $f_i$  is continuous then
5:     Combine  $T$  and  $E$  and discretize into  $b$  equi-width bins
6:     for  $j \leftarrow 1$  to  $b$  do
7:        $h+ = (\sqrt{|T_{f_i,j}|/|T|} - \sqrt{|E_{f_i,j}|/|E|})^2$ 
8:     end for
9:   else  $\{f_i$  is nominal $\}$ 
10:    for each feature value  $j$  of  $f_i$  do
11:       $h+ = (\sqrt{|T_{f_i,j}|/|T|} - \sqrt{|E_{f_i,j}|/|E|})^2$ 
12:    end for
13:   end if
14:    $Hellinger+ = \sqrt{h}/n_f$ 
15: end for
16: return  $Hellinger$ 

```

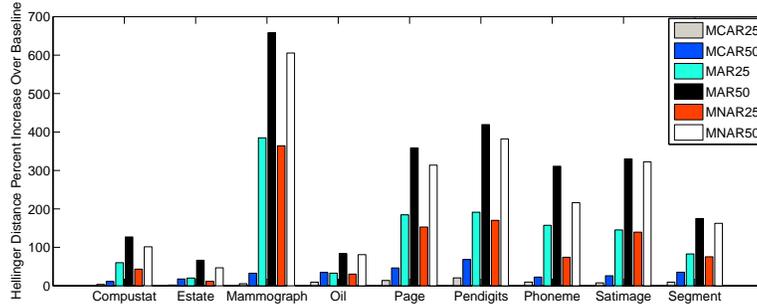


Fig. 15. Hellinger Distance detecting bias. From left to right each set of bars indicates the relative percentage change in Hellinger Distance between the original testing set and MCAR25, MCAR50, MAR25, MAR50, MNAR25, MNAR50 for each dataset.

Hellinger Distance is quite effective in differentiating between the relative level of bias sophistication.

Of additional interest is the skew of Hellinger distances produced. Table 6 demonstrates that there is typically a substantial negative skew to the set of distances calculated, meaning there is a tail of values below the mean. This is indicative that there is more data below the mean than would be expected in a normal distribution. There is generally a strong variability in the distances depending on the feature conditioned for injection of bias. This indicates that different feature thresholds have different effects on the testing set distributions, and it is important to consider the different scenarios. We note that Oil violates the typical trend, likely due to the extremely small size of this dataset. In general, Hellinger distance enables the isolation of features along which bias occurs.

Dataset	MCAR 25	MCAR 50	MAR 25	MAR 50	MNAR 25	MNAR 50
Compustat	-0.790	-0.751	-0.562	-0.554	-0.554	-0.545
E-State	-0.195	-0.055	-0.349	-0.618	-0.715	-1.115
Mammography	0.634	0.466	-0.332	-0.021	-0.179	0.274
Oil	0.957	0.911	0.868	1.007	1.131	1.183
Page	-0.609	-0.253	-0.239	-0.137	-0.103	-0.060
Pendigits	-0.271	-0.836	-0.283	-0.413	-0.161	-0.295
Phoneme	-0.171	0.049	-0.095	-0.002	0.527	0.448
Satimage	-0.097	-0.127	-0.008	-0.025	0.010	-0.001
Segment	-0.656	-0.445	0.134	0.092	0.306	0.272

Table 6. Skew of the average Hellinger Distance per feature

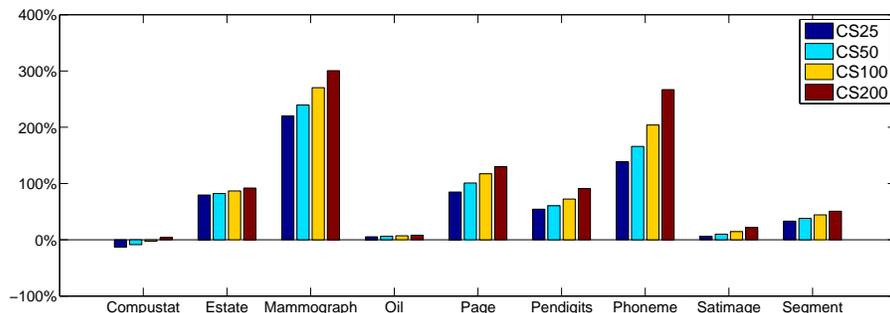


Fig. 16. Hellinger Distance detecting covariate shift. From left to right each set of bars indicates the relative percentage change in Hellinger Distance between the original testing set and CS25, CS50, CS100, and CS200 for each dataset.

5.4.2. Covariate Shift

We note that covariate shift generates the changes in Hellinger distance noted by Figure 16. While doubling bias can significantly increase Hellinger distance (as seen in Figure 15), increasing the degree of covariate shift only increases this distance a relatively minor amount beyond the given threshold. As under the KS test, this limited increase is somewhat expected, given that our mechanism for covariate shift only alters a single feature. That said, a change stemming from this type of covariate shift may go unnoticed by Hellinger distance since the relative increase is small, such as in the Compustat, Oil, Satimage, and Segment datasets. Hence, it is important to have the KS-test in conjunction for covariate shift, as it reflects the point in separation for one feature versus the sample selection bias that affects all the features conditioned on one feature.

5.4.3. Shifting Priors

We also consider how Hellinger distance changes as a result of a shift in class priors. The percent change in this distance as a product of $P(+)$ is presented in Figure 17. We

observe for each dataset that the percent change in Hellinger Distance is minimal at or near its natural distribution. As $P(+)$ moves further from natural, there is a continual increase in Hellinger distance. The change in $P(+)$ is altering the percent change; thus, Hellinger distance is highly effective in detecting a change in class priors since the alteration in the class ratio is affective the underlying feature distribution.

5.4.4. Summary

From these experiments, we note that Hellinger is able to corroborate the findings of KS and complements the differentiation and determination of biases. The KS Test is useful in determining if there is a significant maximal point of separation. Hellinger Distance is more refined in isolating bias since it is a method of comparing the relative densities of two distributions. MCAR is the lowest range, then MNAR, then MAR. We expect this ordering: MCAR is sampled at random and should fairly closely resemble the training set. MAR should produce the highest changes in Hellinger: the feature(s) generating bias have been observed and the distributional change will be reflected by this distance. MNAR is expected to produce results between MCAR and MAR since the feature MNAR biases along is hidden, but it is also reasonable to expect some level of correlation to the observed features. When only one feature is altered via covariate shift, the Hellinger distance might not indicate a distributional divergence. Since the KS test easily detected covariate shift and missed some class priors shift and Hellinger easily picked up class prior shift and missed some covariate shift, we recommend the coupled usage of KS test and Hellinger distance to isolate the biased features.

6. Discussion

This article has proposed an outline for a statistical framework using three tests to detect divergence between two data samples. This framework has been tested on three engineered data change scenarios. The first of these scenarios is sample selection bias, which removes a segment of the testing distribution base on one feature with which other features may covary. Under bias, there is a general performance failure according to the Friedman test. This is reflected by an identification of a divergence between probability populations under the KW test, particularly when the bias is the systematic MAR and MNAR. An increase in the degree of bias is reflected by lowered p-values. Likewise, the relative Hellinger distance corroborates the likely presence of bias in the testing distribution.

Our second change scenario directly affected a single feature per experiment, by physically altering its value. KW is effective in identifying the presence of covariate shift, and lower p-values are somewhat indicative of a larger degree of covariate shift. Suspecting a shift, the KS test then effectively confirms a distributional gap along a single feature, although not necessarily the degree of shift. The Hellinger distance is somewhat effective in identifying when covariate shift occurs, but we observe that relative Hellinger distances can be quite effective when establishing the degree of covariate shift.

Our final scenario is unlike the others in that the feature space is manipulated only indirectly as a byproduct of adjusting the class skew ratio. We note that for some instances, the performance is actually conserved. As we have seen through the KW and KS test, this occurs when the distribution is actually similar to the training sample. This is a compelling argument for continuous monitoring of classifiers' performance for different datasets, as they will reflect different sensitivities to changes in data class skew

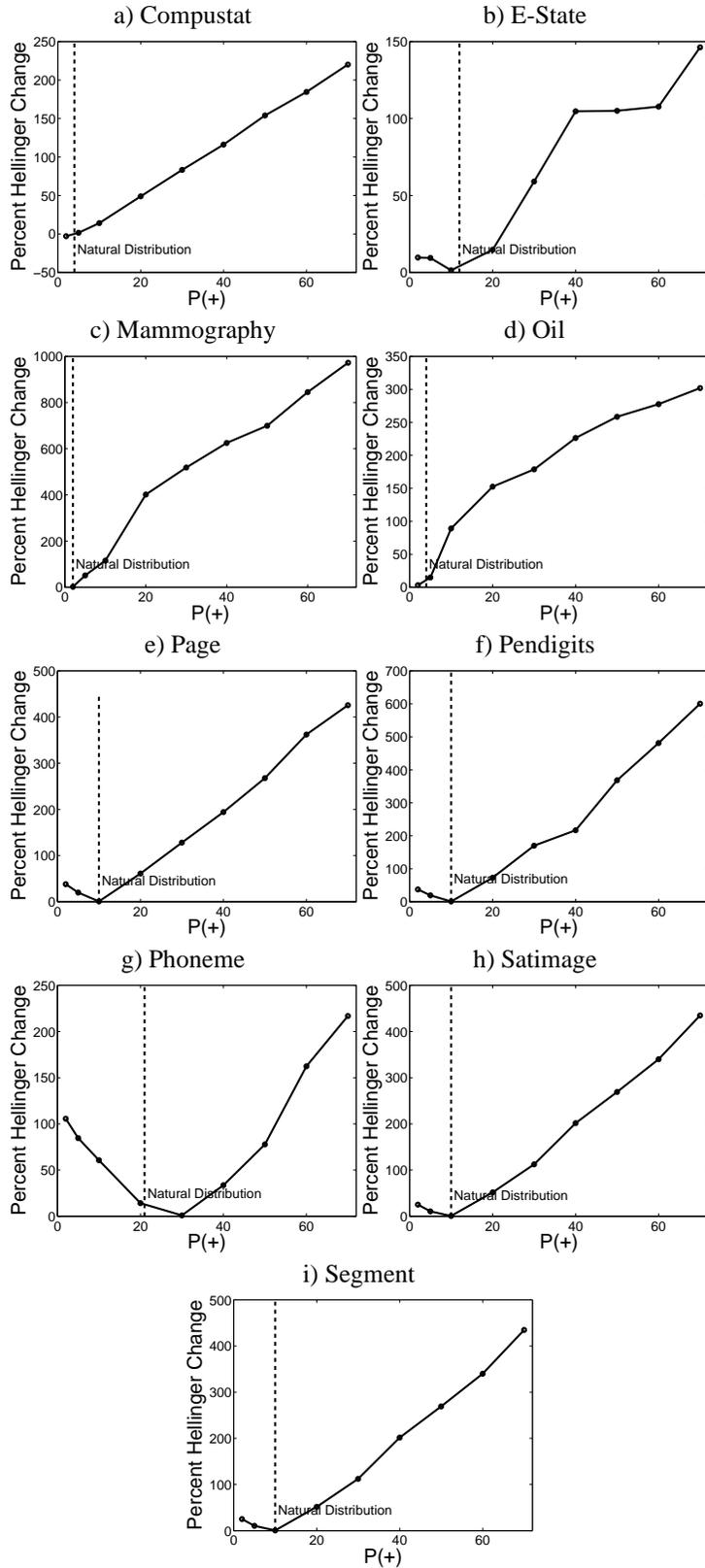


Fig. 17. Percent Change in Hellinger as $P(+)$ changes.

distribution. For instance, as we see for Compustat and Oil, the larger feature shifts are present in the majority class, which is heavily undersampled in the testing distribution because of feature shift. Now, since there are no drifts in the feature space and the former minority or positive class now dominates the testing set, the classifier performs as expected. No surprising elements are seen in the testing samples. In the general case of class prior shift, KW will indicate further investigation should be pursued and increased Hellinger distance in particular correlates as class skew deviates further from the natural level.

7. Conclusions

Data mining is presented with the challenge of drifts in data distribution between the training and testing samples. The basic assumption that the past is a reasonable predictor of future may not hold in different scenarios. This certainly hinders the performance of learning algorithms, as we have also demonstrated in this work. Thus, it becomes critical to identify and react to the changes in data distribution. To that end, we implemented a framework that comprised of a family of statistical measures. We showed that it is possible to proactively detect fractures in classifier performance. Our test suite comprised of a variety of classifiers and data sets with different characteristics.

Based on our observations, we make the following recommendations. Using *Kruskal-Wallis* on the distributions of validation and testing probabilities is useful as a first step. If the practitioner determines there is no significant difference between them, then it is possible to proceed as per typical. Otherwise, the practitioner can use the following steps to isolate altered data. Firstly, the *Kolmogorov-Smirnov* or the χ^2 Test can be used to detect independent feature failure and is quite strong at detecting bias and covariate shift. Through ϕ -correlation analysis, one can then determine the co-failure of features, which we have shown to be quite strong under sophisticated bias. This is relevant for sample selection bias as multiple features are being affected simultaneously by conditioning on one feature's bias. Then, *Hellinger* distance can be used to readily identify and differentiate the level of bias, even when the factor of bias is unmeasured, such as MNAR. However, we recommend using *Hellinger* in conjunction with the *KS* or χ^2 tests as the latter can allow for statistical significance tests and since it is a strong detector for shifting class priors.

We believe that a single statistical measure cannot be used in isolation, rather a family of measures should be used in conjunction to remain more confident in detecting fractures in classifier predictions. We also stress that it is increasingly important to instrument such a comprehensive framework for a continuous monitoring of classifiers' performance.

References

- [1] SVM^{light} Support Vector Machine. http://www.cs.cornell.edu/People/tj/svm_light/.
- [2] A. Basu, I. R. Harris, and S. Basu. Minimum distance estimation: The approach using density-based distances. In *Handbook of Statistics*, volume 15, pages 21–48, 1997.
- [3] R. Caruana and A. Niculescu-Mizil. Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 69–78, 2004.
- [4] N. V. Chawla and G. Karakoulas. Learning From Labeled And Unlabeled Data: An Empirical Study Across Techniques And Domains. *JAIR*, 23:331–366, 2005.
- [5] W. Fan and I. Davidson. ReverseTesting: An Efficient Framework to Select Amongst Classifiers under Sample Selection Bias. In *Proceedings of KDD*, 2006.

- [6] W. Fan, I. Davidson, B. Zadrozny, and P. Yu. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. In *5th IEEE International Conference on Data Mining*, 2005.
- [7] J. D. Gibbons. *Nonparametric Statistical Inference*, 2nd edition. M. Dekker, 1985.
- [8] P. Groot, A. ten Teije, and F. van Harmelen. A quantitative analysis of the robustness of knowledge-based systems through degradation studies. *Knowledge and Information Systems*, 7(2):224–245, 2004.
- [9] L. Hall, B. Mohny, and L. Kier. The Electrotological State: Structure Information at the Atomic Level for Molecular Graphs. *Journal of Chemical Information and Computer Science*, 31(76), 1991.
- [10] J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979.
- [11] T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communications*, 15(1):52–60, February 1967.
- [12] A. N. Kolmogorov. On the empirical determination of a distribution function. (*Italian*) *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.
- [13] M. Kubat, R. Holte, and S. Matwin. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, 30:195–215, 1998.
- [14] M. Kukar. Quality assessment of individual classifications in machine learning and data mining. *Knowledge and Information Systems*, 9(3):364–384, 2006.
- [15] P. Legendre and E. D. Gallagher. Ecologically Meaningful Transformations For Ordination Of Species Data. *Oecologia*, 129:271–280, 2001.
- [16] H. R. Lindman. *Analysis of variance in complex experimental designs*. W. H. Freeman & Co., San Francisco, 1974.
- [17] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.
- [18] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI Repository of Machine Learning Databases, 1998.
- [19] F. Provost and P. Domingos. Tree Induction for Probability-Based Ranking. *Machine Learning*, 52(3):199–215, September 2003.
- [20] J. Quinero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008.
- [21] C. Rao. A Review of Canonical Coordinates and an Alternative to Correspondence Analysis using Hellinger Distance. *Questio (Quaderns d'Estadistica i Investigacio Operativa)*, 19:23–63, 1995.
- [22] J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. A Framework for Structural Risk Minimization. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, 1996.
- [23] N. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. (*Russian*) *Bulletin of Moscow University*, 2:3–16, 1939.
- [24] V. Vapnik. *The Nature of Statistical Learning*. Springer, New York, 1996.
- [25] K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe, and W. P. Kegelmeyer. Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. *IJPRAI*, 7(6):1417–1436, 1993.
- [26] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.
- [27] K. Yamanishi, J. ichi Takeuchi, G. J. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Knowledge Discovery and Data Mining*, pages 275–300, 2004.
- [28] B. Zadrozny. Learning and Evaluating under Sample Selection Bias. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Author Biographies



David Cieslak received B.S. (2004) and M.S. (2006) degrees from the University of Notre Dame's Department of Computer Science and Engineering. He is currently a Ph.D. student at the Department of Computer Science of the University of Notre Dame and is a member of the Data, Inference Analysis, and Learning (DIAL) research group. His interests include class imbalance, performance analysis under changing data distributions, and data segmentation.



Dr. Nitesh V. Chawla is an Assistant Professor in the Department of Computer Science and Engineering at the University of Notre Dame. His core research in machine learning and data mining focuses on learning from unbalanced datasets, cost sensitive learning, distribution sensitive learning, massively parallel and distributed data mining, semi-supervised learning, social networks, and graph based data mining. His work also includes various applications of machine learning to systems, marketing, finance, medicine, and biometrics. His recent research has been supported from NSF and DOJ. He has received various awards for his research and teaching. He is currently the Associate Editor for the IEEE Transactions on SMC-B. He has also served on organizing and program committees for various conferences, workshops, and special issues.