IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2017, accepted July 27, 2017, date of publication August 3, 2017, date of current version August 29, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2735378

A Novel Online and Non-Parametric Approach for Drift Detection in Big Data

MOINAK BHADURI, JUSTIN ZHAN, CARTER CHIU, AND FELIX ZHAN
Big Data Hub, University of Nevada at Las Vegas, Las Vegas, NV 89154, USA
Corresponding author: Justin Zhan (justin.zhan@unlv.edu)

ABSTRACT A sizable amount of current literature on online drift detection tools thrive on unrealistic parametric strictures such as normality or on non-parametric methods whose power performance is questionable. Using minimal realistic assumptions such as unimodality, we have strived to proffer an alternative, through a novel application of Bernstein's inequality. Simulations from such parametric densities as Beta and Logit-normal as well as real-data analyses demonstrate this new method's superiority over similar techniques relying on bounds, such as Hoeffding's. Improvements are apparent in terms of higher power, efficient sample sizes, and sensitivity to parameter values.

INDEX TERMS Change point detection, non-parametric methods, Hoeffding's inequality, Bernstein's inequality, big data, anomaly detection.

I. INTRODUCTION

The nature of temporal progression of a continuous process often comes under scrutiny and changes in the pattern of evolution are frequently indicative of a shift in the possibly deterministic mechanisms that govern the process. Examples originate from a spectrum as rich as quality control (where the production manager might be interested in bounding the proportion of defective items generated), oceanography (where scientists might speculate an increase in sea surface temperatures), drug testing (where physicians could be interested in the pattern of recovery time due to a newly launched drug) and several others. Or one could also consider whether some in an incoming stream of emails could be categorized as spam. The quest that unifies these applications is the search for efficient algorithms that will warn one of a change in one or more of the fundamental properties of the system: properties such as the average level (representation of the most common value) or the amount of fluctuation involved. Assuming such a change point exists, the method that detects it sooner without being unnecessarily taxing on the sample size is favorable.

Irrespective of the actual learning method employed [2], [5], [14], it is prudent to restrict investigation to some data functional detectors such as the mean or the median [33] since these functions are often adequate reflectors of the drift in underlying probability distributions. Using one such functional, the present work will offer a technology that will sound an alarm if the mean level of a process changes significantly so that remedial measures can be promptly undertaken to bring the process back on track. The method relies on realistic assumptions, is easily implementable and is efficient with respect to some of its competitors.

The plan of the current work is simple: section 2 takes a tour of the established results in this area and leads up to a section on probabilistic notions. Section 4 details the working of the probability bound that is pivotal in this endeavor along with a necessary modification. The sections to follow will describe a new statistical test and algorithm while implementing those on real and simulated data sets. We shall conclude with a summary and an exploration of future avenues.

II. RELATED WORK

The problem of detecting drifts has been tackled by several authors and various methods of data storage and analyses have been proposed. A notable one among these methods being the *time window* approach, where a time stamp, defining age is associated with incoming examples. Different subcategories have been introduced, for instance by Gama and Rodrigues [15]. Reliable algorithms to preserve monitoring statistics across windows have also been studied [12]. For the specific purpose of detecting changes in population mean, methods such as ADWIN [5] or ADWIN2 [5] have been constructed, extending the above idea. Tools [2], [14] guaranteeing theoretical limits on the the maximum processing time of incoming instances are prevalent and these thrive on classical principles of online change point detection [30].

Baena-García *et al.* [2] used the discrepancy between two classification errors to detect gradual changes in an online stream.

The drawbacks of the window based approaches (such as the scarcity of a fixed or uniformly adjustable window size) are established by Lazarescu and Venkatesh [25]. For instance, works such as [17] or [6] take the window size as 100 while conducting real data analyses. So one strategy weighs the data or parts of the hypothesis according to their age or current utility [21], [22], [35]. Thus, first order integrated moving average model [8] dependent methods such as EWMA mean monitoring statistic are often useful, for instance when the mean changes in steps [10]. CUSUM [26] procedure is another method along similar lines. Most of these methods assume that the random variable under consideration is governed by some known parametric density [30] such as the normal. Chandola *et al.* [9] observes that such parametric statistical techniques estimate the model parameters under assumptions and that the quality of detection results depend hugely on the actual estimator used. Oomen and Rueda [32] discusses methods utilizing sliding windows to tackle non-stationary environments such as the ones where the underlying distribution changes with the number of incoming observations.

Decision tree based methods under the assumption of a persistent descent [31] and the ECDD [34] method where one uses an EWMA chart to monitor the misclassification rate have both recently been proposed. Klinkenberg [22] introduces ideas on global and local weighing through exponential weighing functions such as $w_\eta(x_t) = e^{-\eta t}$ to quantify gradual changes in the interest in a specific example while Cohen and Strauss [11] considers a sufficiently rich class of decay functions under exponential, polynomial and sliding window frameworks. Noting however that most of the parametric methods suffer from the drawback of unrealistic distributional assumptions, Zhan *et al.* [37], [38] proposed a method based on weak estimators. This is extremely useful when the stationarity assumption comes into question for instance, in connection to spam filtering. Most of the statistical techniques in this regard have been detailed in McGregor [28] and Kong *et al.* [23] and compared by Zhang *et al.* [39]. Other authors such as Metsis *et al.* [29] compares the performance of some naive Bayesian filters while researchers such as Wang *et al.* [36] explored the possibility of online linear classifiers. Performance of filters that minimize false positives (i.e. sounding an alarm in the absence of a shift) have been examined by Guzella and Caminhas [18] and Androutsopoulos *et al.* [1] designed cost effective measures that quantify the impact of false positives. Blanco *et al.* [7] introduced a more reasonable non-parametric method based on Hoeffding's inequality. While remaining within this liberating framework , in the sections to follow, we shall endeavor to improve this new method even further. The methods outlined here can be generalized to cases where the data arrive online but the labels are not straightforward to obtain [41] . We are aware that several authors would prefer to differentiate between notions of abrupt and gradual shifts in trend and that several such as [3], [16], and [19] have previously identified various types of gradual change. However, given our confidence that the proposed bound will efficiently pick up the wide array of changes with considerable ease, we shall not investigate too deeply into the issue of speed of change in this research.

III. PRELIMINARIES

A. NOTATIONS

A solid grounding in the rudiments of probability theory will be imperative for the discussions to follow. This section is designed as a brief refresher of the related ideas.

1) RANDOM VARIABLES

An abstract set S that nests all possible outcomes can almost always be attached to a random experiment. This set, termed the sample space, in turn generates a sigma algebra $\mathcal{F}$ of subsets that houses collections that are closed under complementation and countable unions. Members of $\mathcal{F}$ qualify as measurable sets or events and a set function $P(.)$ on $\mathcal{F}$ can then be defined which should attach numbers between 0 and 1 termed probabilities to members of $\mathcal{F}$. This collection $(S, \mathcal{F}, P)$, termed the probability space, serves as the foundation stone for any probabilistic exercise.

Meaningful mathematics often results from the addition of another layer of complexity: a mapping that connects S to a space enjoying topological niceties such as *Polishness*, a space such as the real line $\mathcal{R}$. A random variable X is precisely that sort of a measurable mapping when the target space is $\mathcal{R}$. To be formal, if an image space (Q, B_Q) can be conceived where B_Q is some σ-algebra on Q, then $X : S \Rightarrow \mathcal{R}$ is a measurable map from $(S, \mathcal{F}, P)$ to (Q, B_Q, P_X). Put differently, X satisfies the following property:

$$X^{-1}(B) := \{\omega \in S : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in B_Q \tag{1}$$

which states that the inverse images of target Borel sets should be legitimate events. For the present exercise, $(\mathcal{R}, B_\mathcal{R})$ will serve as the target space, with $B_\mathcal{R}$ being the Borel sigma algebra on $\mathcal{R}$. Uncertainty is inherited from the parent space to the target space through $B = (-\infty, x]$ in (1):

$$P(X^{-1}(-\infty, x]) = P_X(-\infty, x] = F_X(x)$$

where F_X is traditionally termed as the distribution or the law of X. At times when F_X is differentiable, we can equivalently talk about the probability density function (p.d.f) of X, notationally, $f_X(x)$, with the following understanding:

$$P_X(X \in B) = \int_B f_X(x)dx \quad \forall B \in B_\mathcal{R}$$

The modeling problem at hand dictates the parametric form of $f_X(x)$. An indiscriminately overused choice is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\theta}{\sigma})^2}, \quad x > 0. \tag{2}$$

which describes the normal (μ, σ^2) distribution. Our simulation purposes will call for $Beta(m, n)$ and $Logit - normal(\mu, \sigma)$ choices of $f_X(x)$.

If $f_X(x)$ exists, then describing the behavior of X becomes relatively simple. For instance, one could talk about the average or the expected value $E(X)$ of the variable X, through:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

which quantifies the center of gravity of the probability distribution and is of pivotal importance to the idea of drift detection. The variable of variability around the mean is captured for instance by:

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx$$

an idea largely ignored by some of the methods to follow. Established techniques often capture a drift in the online process through the sample average $\bar{X}$ and properties such as $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$ and $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i)$ (for independently and identically distributed variables) will be useful. Finally, it is intuitively acceptable and mathematically simple to prove that:

$$Var(X + a) = Var(X)$$

the addition of a deterministic constant does not affect the variance structure of a random variable.

2) TESTING HYPOTHESES AND POWER ANALYSIS

In stark contrast to the more widely used field of statistical estimation, where coming up with a reasonable guess of the parameter of interest is of primary importance, modelers are often confronted with the dilemma of choosing one of two competing statements, both typically framed in terms of the target parameter. The one which we would not want to reject unless there is some compelling evidence against it, is termed the null H_0, while the other one is the alternative H_1. Two types of errors naturally abound: falsely rejecting the null, which is usually of severe consequence, and is termed Type-I error. Then there is also falsely accepting H_0, termed as Type-II error. Due to the uncertainty inherent in the value of the parameter, it is not possible to determine whether one of those two errors actually happened. One can only attach probabilities to them. Owing to an inverse relationship between the two types of errors, one usually fixes on the upper limit on the probability of Type-I error and tries to minimize the probability of Type-II error.

One of the most prevalent methods to quantify the quality of a testing procedure is through its power function. This curve records the probability of making a correct decision (i.e rejecting the null when it is false) as a function of changing parameter values. A better test usually manifests itself through a higher power curve compared to the one constructed from an inferior competitor. Such a crucial indicator is usually available in closed, tractable forms for parametric densities such as the normal, but in the dearth of such structure, one has to observe recourse to approximations through simulations.

The following exercise will offer a case in point: let us imagine that under the traditional normal sampling where observations are governed by p.d.f (2), we are interested in the following testing for large values of the mean when the population standard deviation σ is known:

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta > \theta_0, \tag{3}$$

Statistical methods prove that the optimum test will be the one that rejects H_0 in favor of H_1 if $\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}}\tau_\alpha$ where τ_α is the upper α point of a $N(0, 1)$ distribution, and n is the sample size. The probability of rejecting H_0 under θ-sampling, captured by the power function $\pi(\theta)$ is:

$$\pi(\theta) = P_\theta(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}}\tau_\alpha) = 1 - \Phi(\frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + \tau_\alpha) \tag{4}$$

where $\Phi(.)$ represents the 'law' of the standard normal deviate. This is shown in Fig. 1 below.

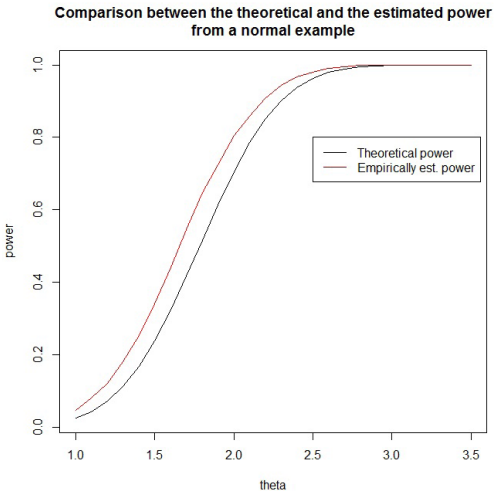


FIGURE 1. Comparison between theoretical and estimated power for a greater than type alternative under Normal sampling.

The sampling distribution result: $\bar{X} \sim N(\theta, \frac{\sigma^2}{n})$ under random draws from (2) is crucial in the derivation and in case this happens to be unknown to a user, an empirical power function can be generated through simulations as follows: With $\theta_0 = 1, \sigma = 2$ and $n = 25$ for instance, one can simulate 25 vectors, each of a sufficiently large size (say 10000) from a $N(1, 4)$ density, calculate the average and extract the 95th percentile from this approximate null distribution. This point should serve as the cutoff $\theta_0 + \frac{\sigma}{\sqrt{n}}\tau_\alpha$ in the formula above. To empirically estimate the power at some arbitrary θ, one can repeat the process with sampling from a $N(\theta, 4)$ density and calculate the proportion of times the means exceed the cutoff. This follows due to the long term relative frequency interpretation of probability and Fig. 1 assures one of the closeness to the exact curve. This exercise should motivate one in certain situations (such as the ones to follow) when the

closed form power expressions are incalculable or when we have a realistic, distribution free assumption on the random draws.

B. HOEFFDING'S BOUND

One of the most celebrated concentration inequalities in the one due to Hoeffding [20] and several forms of the upper bound are prevalent in literature. The one that will be most apt for our purpose can be stated as:

Theorem 1 (Hoeffding's Inequality): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0, 1]$ almost surely $\forall i = 1, 2, .., n$. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon > 0$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-2n\epsilon^2} \tag{5}$$

Authors such as Blanco *et al.* [7] have exploited this bound to propose tests on online drift detection. Though its simple form appeals to many, one can observe that the dispersion structure containing crucial information about the spread of the variable has been ignored. This observation, in part, motivates our proposal.

IV. OUR APPROACH

At the outset, we would like to recall an inequality due to Bernstein and see how it improves the Hoeffding's bound. This improved inequality will eventually pave the way for an efficient test to be described later. Thus, the present section will be devoted to a survey of the Bernstein's bound and of the problems it may present.

A. INTRODUCTION TO BERNSTEIN'S INEQUALITY

In a way similar to Hoeffding's several variants of Bernstein's inequality can be formulated. The version most pertinent to our case is:

Theorem 2 (Bernstein's Inequality): Let $Y_1, Y_2, \ldots, Y_n$ be mean zero independent random variables with $Y_i \leq 1$ almost surely $\forall i = 1, 2, .., n$. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var(Y_i)$. Then for any $\epsilon > 0$:

$$P(\bar{Y} \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\sigma^2 + \frac{\epsilon}{3})}} \tag{6}$$

Using $Y_i = X_i - E(X_i)$, we note that $E(Y_i) = 0$ and $Y_i \leq 1$ a.s. Thus the assumptions are satisfied and we additionally have:

$$Var(Y_i) = Var(X_i) \tag{7}$$

and thus,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n}\sum_{i=1}^{n} Var(X_i) \tag{8}$$

In terms of the original variables, we have:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\sigma^2 + \frac{\epsilon}{3})}} \tag{9}$$

We shall use this form in the analyses to follow.

B. A COMPARISON OF TWO COMPETING INEQUALITIES

In inference, the parameters that we are not interested in are termed nuisance parameters and their presence often pose problems for creating confidence intervals or rejection regions. (6) and (9) above depict such a situation with σ. Popoviciu's inequality detailed below, offers an avenue to get around this issue.

Theorem 3 (Popoviciu's Inequality): If X is a random variable such that its p.d.f. is supported on [a,b], i.e. if the variable takes on values in this interval almost surely, then:

$$Var(X) \leq \frac{1}{4}(b - a)^2 \tag{10}$$

We shall assume that the interval of interest is [0,1]. No generality is lost this way since an arbitrary compact support can be converted to [0,1] through adequate rescaling. Bernstein's inequality modified through Popoviciu's is a seemingly logical competitor to Hoeffding's bound. The following theorem creates this modified bound:

Theorem 4 (Popoviciu-Modified Bernstein's Bound): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0, 1]$ almost surely $\forall i = 1, 2, \ldots, n$. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon > 0$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\frac{1}{4} + \frac{\epsilon}{3})}} \tag{11}$$

Proof: Using (9), we know that under similar conditions,

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\sigma^2 + \frac{\epsilon}{3})}}$$

Now using (10) with $a = 0, b = 1$, we have $Var(X_i) \leq \frac{1}{4}$ whereupon (7) and (8) ensure:

$$\sigma^2 \leq \frac{1}{4}$$
$$\Rightarrow e^{-\frac{n\epsilon^2}{2(\sigma^2 + \frac{\epsilon}{3})}} \leq e^{-\frac{n\epsilon^2}{2(\frac{1}{4} + \frac{\epsilon}{3})}}$$

as required.

Although this provides for a way to make the bound implementable, it unfortunately is not sharper than the Hoeffding's bound proposed above in (5) as is evidenced by the following graph (Fig. 2):

The red curves indicate the modified Bernstein's bounds remain consistently higher than the black curves representing the corresponding Hoeffding's bounds. Hence, in our quest for non-trivial sharper bounds, we impose the following two mild assumptions on the distribution generating the X values:

Assumption 1: The p.d.f. is unimodal.

Assumption 2: The p.d.f. is unimodal with mode M.

Usually knowledge about the mechanism being studied will be enough to check whether these assumptions will go through in a particular instance. For instance, at a smoke shop, the distribution of cigarettes bought will certainly not be unimodal. This is because smoking patterns are rather different among heavy, moderate and occasional smokers. On the other hand, error measurements from an industrial

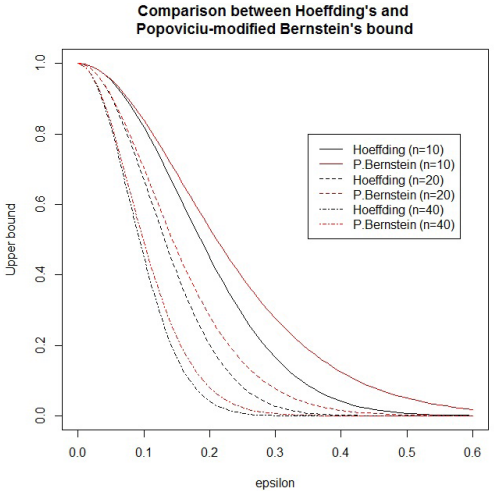


FIGURE 2. Comparison between Hoeffding's and Popoviciu-modified Bernstein's bound for different sample sizes.

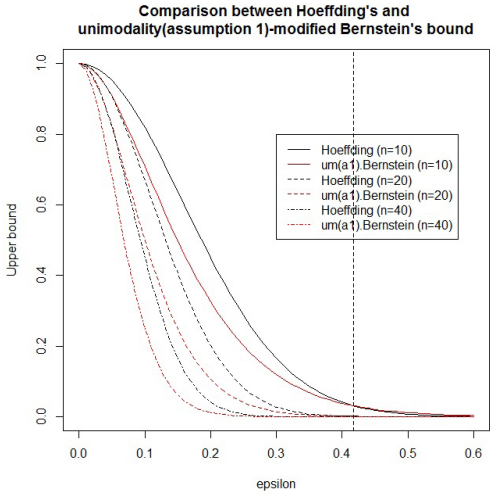


FIGURE 3. Comparison between Hoeffding's and unimodality (assumption 1)-modified Bernstein's bound for different sample sizes.

experiment for instance, tend to accumulate around an attractor, making the assumption of unimodality more realistic. Under assumption 1 and the overarching assumption of the variables being bounded on [0,1], Dharmadhikari and Joag-Dev [13] note that:

$$Var(X_i) \leq \frac{1}{9} \tag{12}$$

and under assumption 2, that:

$$Var(X_i) \leq \frac{1 - M(1-M)}{9} \tag{13}$$

These two observations will be of pivotal importance in proposing better bounds on the tail probabilities.

Theorem 5 (Modal Assumption 1-Modified Bernstein's Bound): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$ almost surely $\forall i = 1, 2, \ldots, n$ and assume that the p.d.f of these variables are unimodal. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon > 0$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\frac{1}{9}+\frac{\epsilon}{3})}} \tag{14}$$

Proof: Similar to Theorem 4.

Theorem 6 (Modal Assumption 2-Modified Bernstein's Bound): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$ almost surely $\forall i = 1, 2, \ldots, n$ and assume that the p.d.f of these variables are unimodal with mode M. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon > 0$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\frac{1-M(1-M)}{9}+\frac{\epsilon}{3})}} \tag{15}$$

Proof: Similar to Theorem 4.

As our next exercise, we would like to exhibit the sharpness of each of the modal assumption modified Bernstein's bound with the Hoeffding's bound through:

Theorem 7 (Sharpness Over the Hoeffding's Bound): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$ almost surely $\forall i = 1, 2, \ldots, n$ and assume that the p.d.f. of these variables are unimodal. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon \in [0, \frac{5}{12}]$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\frac{1}{9}+\frac{\epsilon}{3})}} \leq e^{-2n\epsilon^2} \tag{16}$$

Proof: If we assume $e^{-\frac{n\epsilon^2}{2(\frac{1}{9}+\frac{\epsilon}{3})}} \leq e^{-2n\epsilon^2}$, taking logarithms on both sides:

$$-\frac{n\epsilon^2}{2(\frac{1}{9}+\frac{\epsilon}{3})} \leq -2n\epsilon^2$$
$$\Rightarrow 1 \geq 4(\frac{1}{9}+\frac{\epsilon}{3})$$
$$\Rightarrow \epsilon \leq \frac{5}{12} = 0.4167$$

We note that the value 0.4167 is not too strict a limitation in view of the fact that most prevalent statistical practices keep the probability bounds at levels such as 0.05 or 0.1. What follows is a graphical confirmation of the claim above (Fig. 3).

The limit 0.4167 can be pushed further to the right if one intends to use assumption 2.

Theorem 8 (Sharpness Over the Hoeffding's Bound): Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$ almost surely $\forall i = 1, 2, \ldots, n$ and assume that the p.d.f. of these variables are unimodal with mode M. With $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and any $\epsilon \in [0, \frac{5}{12} + \frac{1}{3}M(1-M)]$:

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2(\frac{1-M(1-M)}{9}+\frac{\epsilon}{3})}} \leq e^{-2n\epsilon^2} \tag{17}$$

Proof: Similar to Theorem 7.

Note that the limit coincides with the one obtained in Theorem 7 under extreme skewness, i.e when $M = 0$ or when $M = 1$.

V. TEST BASED ON BERNSTEIN'S BOUND

A rigid parametric set of assumptions such as Gaussianity often pave the way foe tests to capture erratic movements in

the average or the dispersion structure of an ongoing process. But unfortunately, such assumptions often prove unrealistic and far-fetched, which in turn calls for the development of distribution free tests, along lines similar to the one proposed by Blanco *et al.* [7], detailed below:

Theorem 9 (Test Based on Hoeffding's Inequality, Blanco et al. [7]): If $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are independent random variables with values in [0,1], then to test

$$H_0 : E(\bar{X}) \leq E(\bar{Y}) \quad vs \quad H_1 : E(\bar{X}) > E(\bar{Y}) \tag{18}$$

the critical region would be $\bar{X} - \bar{Y} \geq \epsilon_\alpha$ where:

$$\epsilon_\alpha = \sqrt{\frac{1}{n} ln \frac{1}{\alpha}} \tag{19}$$

with α being the pre-decided level of significance.

Proof: This can be readily checked using Theorem 1 by setting $Z_i = X_i - Y_i \; \forall i = 1(1)n$ thereby converting a two-dimensional problem into a one- dimensional form.

Blanco *et al.* [7] used the region above to propose a limit to the probability of Type-II error and have generalized this test to its weighted counterpart. Owing to the sharpness evidenced in the modal assumption modified Bernstein's inequality over Hoeffding's, it stands to reason that a similar test using the better inequality will generate a better critical region. Such a realization leads to the following:

Theorem 10 (Test Based on Bernstein's Inequality): If $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are independent random variables with values in [0,1], then to test

$$H_0 : E(\bar{X}) \leq E(\bar{Y}) \quad vs \quad H_1 : E(\bar{X}) > E(\bar{Y}) \tag{20}$$

the critical region (stricter than the one shown previously) can be taken to be $\bar{X} - \bar{Y} \geq \epsilon^*_\alpha$ where ϵ^*_α can be solved from the quadratic:

$$n\epsilon^2_\alpha - \frac{2}{3}\epsilon_\alpha ln(\frac{1}{\alpha}) - \frac{2}{9} ln(\frac{1}{\alpha}) = 0 \tag{21}$$

Proof: This follows by setting $e^{-\frac{n\epsilon_\alpha{}^2}{2(\frac{1}{9}+\frac{\epsilon_\alpha}{3})}} = \alpha$ and solving for α.

Such a critical region would then lead to confidence intervals or bounds on probabilities of misclassification. We shall however, not dwell on these matters, but rather focus on formalizing the result above through an algorithm.

A. NON-PARAMETRIC DRIFT DETECTION ALGORITHM

A straightforward approach to the application of the modified Bernsteins bound to drift detection is presented here. Indeed, the proposed inequality is not limited to any one particular algorithm. It can be used in any implementation based on a statistical test by substituting the test used to detect drift with the one described above. In this approach, the general idea is to compare the expected value of two contiguous halves of a new stream of data such that when their difference exceeds the threshold determined by Bernsteins bound, we can be confident that concept drift has appeared and update the classifier accordingly.

Algorithm 1 N-PDDA: Drift Detection Method Based on the Modified Bernsteins Bound

Input: A stream of $x_1, x_2, x_3, \ldots$ where $x_i \in [a, b] \forall i$.
Parameter: $w \in (0, 1)$: fraction of drift detection threshold
used for warning level, $\alpha_W \in (0, 1]$: confidence for warning
level, α_D *in*$(0, 1]$: confidence for drift level.
Ensure: STATE ∈ STABLE, WARNING, DRIFT
/* variable declarations */
$\bar{X}$: mean of values computed from x_1 to $x_{n/2}$
$\bar{Y}$: mean of values computed from $x_{n/2+1}$ to x_n
$\epsilon^*_{\alpha_D}, \epsilon^*_{\alpha_W}$: critical values used.
for all new x_i in stream **do**
update $\bar{X}$and $\bar{Y}$ considering new value x_i
calculate $\epsilon^*_{\alpha_D}, \epsilon^*_{\alpha_W}$
if $E(\bar{X}) \leq E(\bar{Y})$ is rejected at level α_D
then
STATE ← DRIFT
reset Xand Y
goto "for all"
else
if $E(\bar{X}) \leq E(\bar{Y})$ is rejected at level α_W
then
STATE ← WARNING
else
STATE ← STABLE
end if
end if
end for

Algorithm 1 revolves around three states that describe the current condition of the data: STABLE, which indicates that no drift appears to be present, WARNING, when concept drift may be present, and DRIFT, when we are confident that concept drift is present. The state is determined by the previously described statistical test, in which WARNING is signaled when the difference exceeds a predefined fraction of the critical value, and DRIFT is signaled when the critical value is exceeded. The algorithm does not explicitly address the actions taken as a result of the state, which are open to implementation. The presence of the intermediary WARNING state can be used to allow an implementation to buffer new observed data to train a new classifier. Then, in the case that concept drift is eventually confirmed via the DRIFT state, the new classifier can be instated. However, should the state return to STABLE after having been in WARNING, we can safely assume a false alarm and discard the buffered data.

In each iteration corresponding to the presence of a new x_i on stream, the algorithm updates the values of $\bar{X}$ and $\bar{Y}$ with the new data taken into account. The set of values used to compute each mean together comprise the entire set of new data processed since concept drift was last detected, such that each set contains a contiguous half exactly equal in size.

In other words, X contains values x_1 to $x_{n/2}$ and Y contains values $x_{n/2+1}$ to x_n. There are implementation-level concerns involving adjusting the sets and accounting for odd amounts of data that can be addressed as trivial. Once the means are updated, we can determine the state of the data by applying the statistical test. If the difference exceeds the critical value ϵ_α^*, we have identified concept drift, in which case a new classifier is installed, all variables are reset, and the algorithm restarts on a new stream of data. If the difference does not exceed the critical value but does exceed the warning level, the WARNING state is instead invoked, possibly prompting the collection of data and training of a potential replacement classifier. Barring these events, the state is STABLE.

VI. SIMULATION STUDIES

This section will employ both Hoeffding's and Bernstein's methods on data drawn from parametric distributions such as Beta and Logitnormal and will attempt to find a winner, a most efficient method. Such an effort will motivate our real data analyses to follow. The $Beta(m, n)$ density of the first kind is given by the following p.d.f.:

$$f(x|m,n) = \frac{1}{B(m,n)} x^{m-1}(1-x)^{n-1} \quad , x \in [0,1] \qquad (22)$$

where $B(m, n)$ is the usual Beta function given by $B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$. Methods to simulate observations from this density exist on softwares such as R which we shall use to draw samples of varying sizes. Another useful density to look at will be the Logitnormal (μ, σ) density given by:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x(x-1)} e^{-\frac{(logit(x)-\mu)^2}{2\sigma^2}}, \quad x \in [0,1] \qquad (23)$$

where $logit(x) = log\frac{x}{1-x}$. The theorems described above can be used on data coming from either of these distributions since they are bounded by [0,1].

A. THE SIMULATED DATA SETS

We first re-parametrize the Beta density as $Beta(m, n+\delta)$ with δ serving as the tuning parameter, so that changes in δ will change the distribution in a recognizable way as depicted in Fig. 4. It is not hard to show that $E(X) = \frac{m}{m+n}$ under $Beta(m, n)$ sampling.

To generate the empirical distributions shown above, we have held m and n at 3 and 4 respectively and varied δ over 0,3,6. Higher values of δ thus leads to easier differentiation between the baseline population (say X, corresponding to $\delta = 0$) and the alternate population (say, Y). Almost all reasonable change detection tools should sound alarms for large vales of δ, but those that are able to do so even for small values of δ (i.e. when X and Y samples are extremely similar) are of course, preferable. The power functions discussed later will shed the necessary light on this property.

The Logitnormal density does not admit closed form representation for its mean, but changing μ can have a drastic effect on the skewness structure as shown in the adjoining graph. So for the power analyses to follow, σ shall be held constant

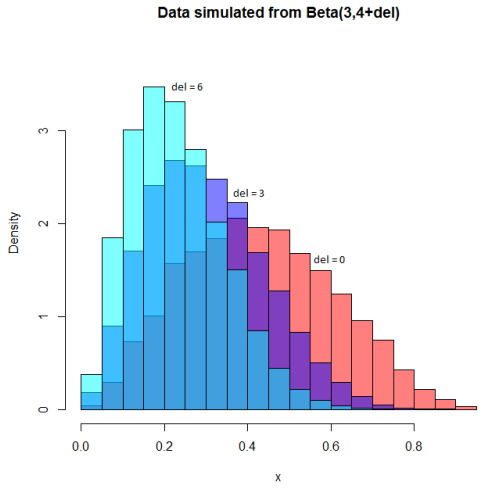


FIGURE 4. Changes in the empirical distribution with changing δ, Beta case.

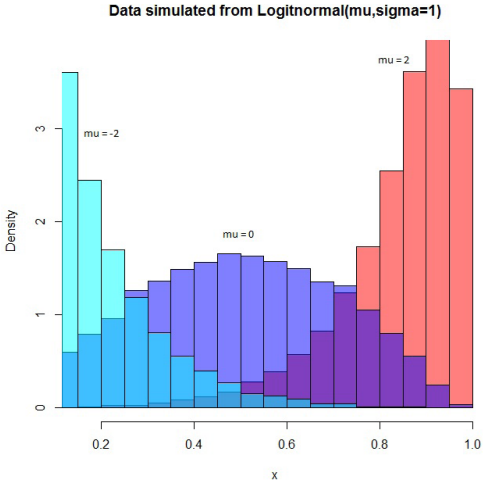


FIGURE 5. Changes in the empirical distribution with changing δ, Logitnormal case.

at 1 and μ shall change to replicate the different Logitnormal densities (Fig. 5).

B. POWER COMPARISONS

We can recall that power of a test is the probability of reaching a correct conclusion, formally defined through $1-Prob(Type-IIerror)$. This quantity needs to be maximized and to get reasonable reliable estimates, we have generated the X values from a $Beta(3, 4)$ density and the Y values from $Beta\ (3, 4+\delta)$ densities with different choices of δ. In view of the analytical expression for the mean, the original testing:

$$H_0 : E(\bar{X}) \leq E(\bar{Y}) \quad vs \quad H_1 : E(\bar{X}) > E(\bar{Y}) \qquad (24)$$

boils down to testing:

$$H_0 : 1 + \frac{\delta}{7} \leq 1 \quad vs \quad H_1 : 1 + \frac{\delta}{7} > 1 \qquad (25)$$

Choosing $\alpha = 0.05$ and different values of the sample size n, we evaluate the critical region using (21) with simulations of strength 1000, find the proportion of times the statistic

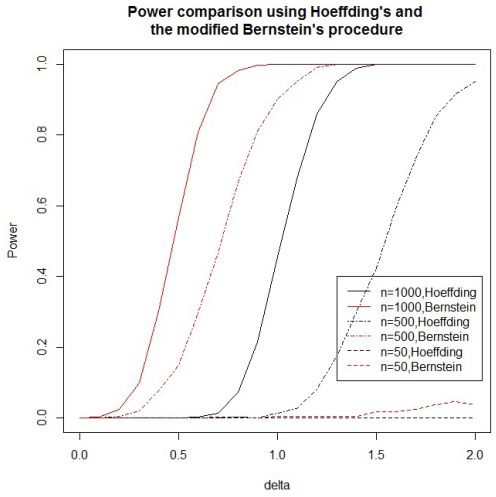


FIGURE 6. Power comparison between the two competing processes, Beta case.

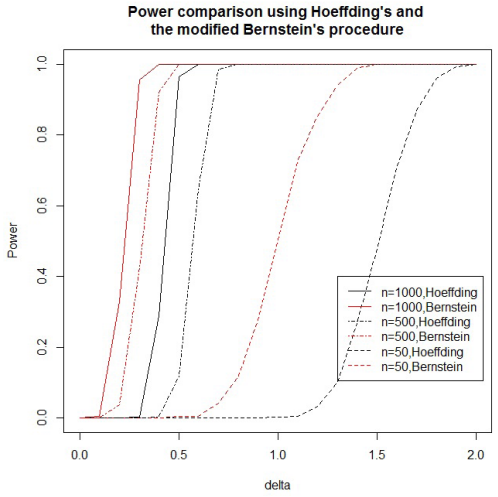


FIGURE 7. Power comparison between the two competing processes, Logitnormal case.

TABLE 1. Summary of data sets.

Dataset	Size	Sample Frequency	Mean	SD
Ozone Level Detection	2261	Day	2.314949	0.9229078
River Hirnant Flow	2928	Half-Hour	2.368867	2.93536
Air Quality	7674	Hour	2.15275	1.453252
Occupancy Detection	9752	Minute	21.00177	1.020693
Saugeen River Flow	23741	Day	30.06429	39.1638
Household Power Consumption	2049280	Minute	1.091615	1.057294

$\bar{X} - \bar{Y}$ gets trapped in the critical region. This serves as an estimate of the power. The following graphs record our findings Figs. 6 and 7:

We can observe that for either case, for every choice of the sample size, the new method proposed through the modified version of the Bernstein's inequality is generating greater power than the established method using Hoeffding's bound. The new power curves are consistently steep even for small choices of the tuning parameters (synonymous to a difficult detection), without a taxing necessity for a large sample size.

VII. REAL DATA EXPERIMENT

To follow the simulation studies, this section involves the application of the proposed drift detection algorithm on several real-world datasets.[1] Six time-series datasets with sizes ranging from a couple thousand values to over two million were selected. Fig. 8 displays the time-series plots

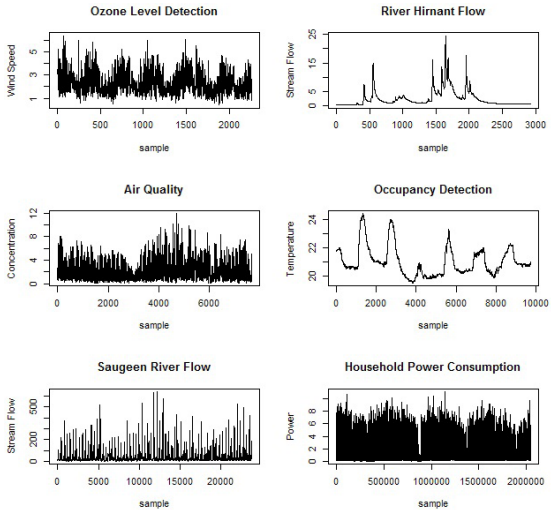


FIGURE 8. Time series generated from six different data sets.

of each dataset, and Table 1 contains their basic attributes. Variations in their appearance and characteristics, as can be seen, serve to provide a more complete analysis of the effectiveness of the tests. The presence and type of concept drift in these datasets cannot be determined, but for the purposes of our experiment it is sufficient to compare the volume of drift detected by Hoeffdings and modified Bernsteins inequality tests.

As is apparent, some of these are more volatile compared to others and not much can be said definitely of the presence and type of concept drift in the data sets, but the value of the analysis lies in the comparison of the number of drifts detected throughout a dataset by the two competing inequality tests. The results of the experiments can be observed in Table 2.

It is instantly observable that the Bernsteins inequality test detected more instances of drift than the Hoeffdings inequality test for every dataset. In some cases, the difference is impressively pronounced; for datasets like Occupancy

[1] http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection
https://datamarket.com/data/set/232a/half-hourly-precipitation-and-stream-flow-river-hirnant-wales-uk-november-and-december-1972
http://archive.ics.uci.edu/ml/datasets/Air+Quality
http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection
https://datamarket.com/data/set/235a/mean-daily-saugeen-river-flows-jan-01-1915-to-dec-31-1979
http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

TABLE 2. Drift detection.

Dataset	Drifts Detected (Hoeffding's)	Drifts Detected (Bernstein's)
Ozone Level Detection	30	262
River Hirnant Flow	10	17
Air Quality	247	708
Occupancy Detection	13	228
Saugeen River Flow	265	281
Household Power Consumption	4721	35507

Detection, the modified Bernsteins inequality proved to be demonstrably more sensitive in detecting drift. In the streamflow datasets (River Hirnant Flow and Saugeen River Flow) the difference in drifts detected was much smaller. This is likely attributable to the sharply defined movements seen in the graphs, which caused both inequality tests to identify drifts simultaneously. For other datasets with more noise and less definition, the finer sensitivity of the modified Bernsteins bound proved to key in on drift that Hoeffdings was unable to, to tremendous effect. It seems quite conclusive that the application of the modified Bernsteins inequality presents a marked improvement over Hoeffdings inequality in drift detection in real-world data.

VIII. CONCLUSIONS AND FUTURE WORK

Our current work advocates the use of a new technology modifying an infrequently used concentration inequality termed Bernstein's bound through a realistic assumption on unimodality to detect drifts in the mean level of an online process. While established methods such as the ones using Hoeffding's inequality ignores the dispersion structure of the data, our method does not with the interesting consequence of an improved power function. Real and simulated data analyses confirm that even small changes in the mean level will be picked up with a greater degree of reliability. Efficient detection of rare and risky events will be particularly easier owing to its capability to handle small sample sizes. Dependence on minimal assumptions such as boundedness and unimodality will make it amenable to a large variety of real world situations.

Fruitful work can be done from here from toward different avenues of research. Concentration inequalities putting bounds on the deviation of mean from its expected value arise with considerable frequency in probabilistic literature. Bercu *et al.* [4] for instance, is an excellent source in this regard. As shown by Zheng [40], under assumptions similar to ours,

$$P(\bar{X} - p > t) \leq w^{-n(p+t)}(q + pw - \frac{\sigma^2(w-1)^2}{2w})^n \qquad (26)$$

where

$$p = E(\bar{X}), \quad q = 1 - p, \; p_i = E(X_i),$$
$$\sigma^2 = \sum_{i=1}^{n}(p_i - p)^2/n,$$
$$A = (q - t)(p - \frac{\sigma^2}{2}), \quad B = -(p + t)(q + \sigma^2),$$
$$C = \frac{\sigma^2}{2}(1 + p + t), \quad w = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

or as shown by Krafft and Schmitz [24] using infinite series expansions:

$$P(\bar{X} - \mu > t) \leq e^{-nL_1(t)}, \quad 0 < t < 1 - \mu \qquad (27)$$
$$P(\bar{X} - \mu > t) \leq e^{-nL(1/2,t)},$$
$$0 < t < 1 - \mu \leq \frac{1}{2} \; or \; 0 < t < \frac{1}{2} - \mu \qquad (28)$$

where

$$L_1(t) = 2t^2 + \frac{4}{9}t^4 + \frac{2}{9}t^6$$

and

$$L(1/2, t) = 2t^2 + \frac{4}{3}t^4 + \frac{32}{15}t^6 + \sum_{k=4}^{\infty}\frac{(2t)^{2k}}{2k(2k-1)}.$$

Each of these has inherent limitations: in the first instance, for example, the right hand bound is itself a function of the parameter of interest. In the second, apart from the forbidding numerical computations, the upper bound is split up into ranges which might pose considerable hindrances to real applications. However, if methods can be devised to circumnavigate these (for instance, through dynamically changing estimates on the right hand side of (26)), then it would be interesting to compare the power functions to the one we have got here.

That apart, a generalized weighted version of the method using the McDiarmid [27] approach can be formulated along lines similar to Blanco *et al.* [7] where recent observations will carry more weight towards detecting a change. Similar ideas can be extended to the context of checking drifts in the variance structure. For the time being however, it is hoped that the significant improvement in power will encourage data modelers to adopt the new methodology and generate fruitful research along the veins suggested and detailed above.

REFERENCES

[1] I. Androutsopoulos, J. Koutsias, K. Chandrinos, and C. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (ACM)*, New York, NY, USA, 2000, pp. 160–167.

[2] M. Baena-García, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno, "Early drift detection method," in *Proc. 4th Int. Workshop Knowl. Discovery Data Streams*, 2006, pp. 77–86.

[3] P. L. Bartlett, "Learning with a slowly changing distribution," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 243–252.

[4] B. Bercu, B. Delyon, and E. Rio, *Concentration Inequalities for Sums and Martingales* (Springer Briefs in Mathematics). Springer, 2015, doi: 10.1007/978-3-319-22099-4.

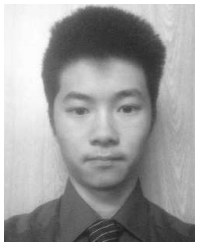
[5] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 443–448.

[6] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, Sep. 2010.
[7] I. F. Blanco, J. del Campo-Ávila, G. Ramos-Jiménez, R. M. Bueno, A. Ortiz-Díaz, and Y. Caballero-Mota, "Online and non-parametric drift detection methods based on hoeffding's bounds," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 810–823, Mar. 2015.
[8] G. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. San Francisco, CA, USA: Holden-Day, 1990.
[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.
[10] A. Chen and E. Elsayed, "Design and performance analysis of the exponentially weighted moving average mean estimate for processes subject to random step changes," *Technometrics*, vol. 44, no. 4, pp. 379–389, 2002.
[11] E. Cohen and M. J. Strauss, "Maintaining time-decaying stream aggregates," *J. Algorithms*, vol. 59, no. 1, pp. 19–36, 2006.
[12] M. Datar, A. Gionis, P. Indyk, and R. Motwani, "Maintaining stream statistics over sliding windows," *SIAM J. Comput.*, vol. 31, no. 6, pp. 1794–1813, 2002.
[13] S. W. Dharmadhikari and K. Joag-Dev, "Upper bounds for the variances of certain random variables," Florida State Univ., Tallahassee, FL, USA, Tech. Rep. M-807, AFOSR 89–240.
[14] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 3171. Heidelberg, Germany: Springer, 2004, pp. 286–295.
[15] J. Gama and P. Rodrigues, *Learning From Data Streams: Processing Techniques in Sensor Networks*, 1st ed. New York, NY, USA: Springer-Verlag, 2007.
[16] J. Gama, I. E. Žliobait, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, p. 44, Dec. 2014.
[17] J. Gama, R. Sebastião, and P. Rodrigues, "On evaluating stream learning algorithms," *Mach. Learn.*, vol. 90, no. 3, pp. 317–346, 2013.
[18] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10206–10222, 2009.
[19] D. P. Helmbold and P. M. Long, "Tracking drifting concepts by minimizing disagreements," *Mach. Learn.*, vol. 14, no. 1, pp. 27–45, 1994.
[20] W. Hoeffding, " Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[21] I. Katakis, G. Tsoumakas, and I. Vlahavas, "An ensemble of classifiers for coping with recurring contexts in data streams," in *Proc. 18th Eur. Conf. Artif. Intell. Conf.*, 2008, pp. 763–764.
[22] R. Klinkenberg, "Learning drifting concepts: Example selection vs. example weighting," *Intell. Data Anal.*, vol. 8, no. 3, pp. 281–300, 2004.
[23] J. Kong, B. Rezaei, N. Sarshar, V. Roychowdhury, and P. Boykin, "Collaborative spam filtering using E-mail networks," *Computer*, vol. 39, no. 8, pp. 67–73, 2006.
[24] O. Krafft and N. Schmitz, "A note on Hoeffding's inequality," *J. Amer. Statist. Assoc.*, vol. 64, no. 327, pp. 907–912, 1969.
[25] M. M. Lazarescu and S. Venkatesh, "Using multiple windows to track concept drift," *Intell. Data Anal.*, vol. 8, no. 1, pp. 29–59, 2004.
[26] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: Properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–29, 1990.
[27] C. McDiarmid, "On the method of bounded differences," *Surv. Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.
[28] C. McGregor, "Controlling spam with spamassassin," *Linux J.*, vol. 153, no. 9, pp. 1–6, 2007.
[29] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes—Which Naive Bayes?" in *Proc. 3rd Conf. Email Anti-Spam (CEAS)*, 2006, pp. 125–134.
[30] D. Montgomery, *Introduction to Statistical Quality Control*. New York, NY, USA: Wiley, 2001.
[31] M. Núñez, R. Fidalgo, and R. Morales, "Learning in environments with unknown dynamics: Towards more robust concept learners," *J. Mach. Learn. Res.*, vol. 8, pp. 2595–2628, Apr. 2007.
[32] B. Oommen and L. Rueda, "Stochastic learning-based weak estimation of multinomial random variables and its applications to pattern recognition in non-stationary environments," *Pattern Recognit.*, vol. 39, no. 3, pp. 328–341, 2006.
[33] G. J. Ross, D. K. Tasoulis, and N. M. Adams, "Nonparametric monitoring of data streams for changes in location and scale," *Technometrics*, vol. 53, no. 4, pp. 379–389, 2011.
[34] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 191–198, 2012.
[35] H. Wang, W. Fan, P. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 226–235.
[36] B. Wang, G. J. F. Jones, and W. Pan, "Using online linear classifiers to filter spam emails," *Pattern Anal. Appl.*, vol. 9, no. 4, pp. 339–351, 2006.
[37] J. Zhan, B. J. Oommen, and J. Crisostomo, "Anomaly detection in dynamic systems using weak estimators," *ACM Trans. Internet Technol.*, vol. 11, no. 1, p. 3, 2011.
[38] J. Zhan, J. Oommen, and J. Crisostmo, "Anomaly detection in dynamic social systems using weak estimators," in *Proc. IEEE Int. Conf. Comput. Sci. Eng.*, Los Alamitos, CA, USA, 2009, pp. 18–25.
[39] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques," *ACM Trans. Asian Lang. Inf. Process.*, vol. 3, no. 4, pp. 243–269, 2004.
[40] S. Zheng, "An improved Hoeffding's inequality," Dept. Mathematics, Missouri State Univ., Springfield, MO, USA, Tech. Rep. [Online]. Available: http://people.missouristate.edu/songfengzheng/Pubs/ImprovedHoeffding.pdf
[41] I. E. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 27–39, Jan. 2013.

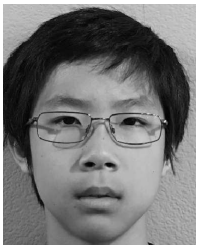
MOINAK BHADURI is currently pursuing the Ph.D. degree in statistics with the Department of Mathematical Sciences, University of Nevada, Las Vegas.

His research interests include point processes, repairable systems, anomaly detection, and clustering.

JUSTIN ZHAN is currently the Director of Big Data Hub and a Faculty Member with the Department of Computer Science, Howard R. Hughes College of Engineering, University of Nevada, Las Vegas.



CARTER CHIU is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Nevada, Las Vegas. He is a member of the Big Data Hub with the University of Nevada, Las Vegas.



FELIX ZHAN is a member of the Big Data Hub. His research interests include big data analytics and cyber security.