# Correcting the Usage of the Hoeffding Inequality in Stream Mining *

Pawel Matuszyk, Georg Krempl, and Myra Spiliopoulou

Otto-von-Guericke University Magdeburg, Germany,
{pawel.matuszyk,georg.krempl,myra}@iti.cs.uni-magdeburg.de

**Abstract.** Many stream classification algorithms use the Hoeffding Inequality [6] to identify the best split attribute during tree induction. We show that the prerequisites of the Inequality are violated by these algorithms, and we propose corrective steps. The new stream classification core, `correctedVFDT`, satisfies the prerequisites of the Hoeffding Inequality and thus provides the expected performance guarantees. The goal of our work is not to improve accuracy, but to guarantee a reliable and interpretable error bound. Nonetheless, we show that our solution achieves lower error rates regarding split attributes and sooner split decisions while maintaining a similar level of accuracy.

## 1 Introduction

After the seminal work of Domingos and Hulten on a very fast decision tree for stream classification [1], several decision tree stream classifiers have been proposed, including CVFDT [7], Hoeffding Option Tree [9], CFDTu [11], VFDTc [3], as well as stream classification rules (e.g. [8, 4]). All of them apply the Hoeffding Bound [6] to decide whether a tree node should be split and how. We show that the Hoeffding Inequality has been applied erroneously in numerous stream classification algorithms, to the effect that the expected guarantees are not given.

We propose `correctedVFDT`, which invokes the Inequality with correct parameter settings and uses a new split criterion that satisfies the prerequisites. Thus, `correctedVFDT` provides the expected performance guarantees. We stress that our aim is not a more accurate method, but a more *reliable* one, the performance of which can be properly interpreted.

The paper is organised as follows. In the next section, we present studies where problems with the usage of the Hoeffding Bound have been reported and alternatives have been proposed. In section 3, we explain why the usage of the Hoeffding Inequality in stream classification is inherently erroneous. In section 4 we propose a new method that alleviates these errors, and in section 5, we prove that it satisfies the prerequisites of the Hoeffding Inequality and thus delivers the expected performance guarantees. In section 6, we show that our approach

* Part of this work was funded by the German Research Foundation project SP 572/11-1 IMPRINT: Incremental Mining for Perennial Objects.

has competitive performance on synthetic and real data. Section 7 summarizes the findings and discusses remaining open issues.

## 2 Related Work

Concerns on the reliability of stream classifiers using the Hoeffding Bound have been raised in [9]: Pfahringer et al. point out that "Despite this guarantee, decisions are still subject to limited lookahead and stability issues." In Section 6, we show that the instability detected in [9] is quantifiable.

Rutkowski et al. [10] claim that the Hoeffding Inequality [6] is too restrictive, since (A) it only operates on numerical variables and since (B) it demands an input that can be expressed as a sum of the independent variables; this is not the case for Information Gain and Gini Index. They recommend McDiarmid's Inequality instead, and design a 'McDiarmid's Bound' for Information Gain and Gini Index [10]. However, as we explain in Section 3, the most grave violation of the Hoeffding Inequality in stream classification concerns the independence of the variables (prerequisite B). This violation of the Inequality's assumptions is not peculiar to the Hoeffding Inequality, it also holds for the way the McDiarmid Bound uses the McDiarmid Inequality. Replacing one Inequality with another does not imply that the prerequsite is satisfied. Hence, we rather replace the split criterion with one that satisfies its prerequisites. We concentrate on the Hoeffding Inequality in this work. The McDiarmid Inequality is more general indeed and we can study it in future work. For the purposes of stream classification, though, the Hoeffding Inequality seems sufficient, because restriction (A) is irrelevant: the split functions return real numbers anyway.

## 3 Hoeffding Bound - Prerequisites and Pitfalls

The Hoeffding Inequality proposed by Wassily Hoeffding [6] states that for a random variable $Z$ with range $R$, the true average of $Z$, $\overline{Z}$, deviates from the observed average $\widehat{Z}$ not more than $\varepsilon$, subject to an error-likelihood $\delta$:

$$|\overline{Z} - \widehat{Z}| < \varepsilon \text{ , where } \varepsilon = \sqrt{\frac{R^2 \cdot \ln(1/\delta)}{2n}} \tag{1}$$

where $n$ is the number of instances. Inequality 1 poses following PREREQUISITES:

1. The random variables must be identically distributed and almost surely bounded; the variable ranges are used when computing the bound.
2. Random observations of the variables must be independent of each other.

In many stream classification algorithms, $Z$ is the value returned by the function computing the 'goodness' of a split attribute. Given a significance level $\delta$, the Hoeffding Inequality states whether the instances $n$ seen thus far are enough for choosing the best split attribute. This is mission-critical, since wrong splits

(especially for nodes close to the root) affect the performance of the classifier negatively. In the presence of drift, this may also lead to uninformed decisions about discarding or replacing a subtree. We show that stream classification methods violate the prerequisites of the Hoeffding Inequality (subsection 3.1) and that the decision bound is wrongly set (3.2).

### 3.1 Violation of Prerequisites

Domingos and Hulten [1] proposed Information Gain (IG) and Gini Index (GI) as exemplary split functions appropriate for the Hoeffding Bound: at each time point, the data instances in the tree node to be split are considered as the observations input to the Hoeffding Inequality, and the values computed upon them by IG/GI are assumed to be averages.

*Violation 1:* The Hoeffding Inequality applies to *arithmetical* averages only [6]. IG and GI "can not be expressed as a sum $S$ of elements" (i.e. "of the independent variables") [10]. We do not elaborate further on this issue, since it is obvious.

*Violation 2:* The variables, i.e. the observations used for the computation of the split criterion must be independent (PREREQ. 2). However, consider a sliding window of length 4 and assume the window contents $w_1 = [x_1, x_2, x_3, x_4]$ and then $w_2 = [x_3, x_4, x_5, x_6]$, after the window has moved by two positions. Obviously, the window contents overlap. When a function like IG computes a value over the contents of each window, it considers some instances more than once. Thus, the computed values are not independent.

### 3.2 A Decision Bound that Cannot Separate between Attributes

Domingos and Hulten specify that the Hoeffding Bound should be applied as follows, quoting from [1], second page, where $G$ is the split function:

> "Assume $G$ is to be maximized, and let $X_a$ be the attribute with highest observed $\overline{G}$ after seeing $n$ examples, and $X_b$ be the second-best attribute. Let $\Delta\overline{G} = \overline{G}(X_a) - \overline{G}(X_b) \geq 0$ be the difference between their observed heuristic values. Then, given a desired $\delta$, the Hoeffding bound guarantees that $X_a$ is the correct choice with probability $1 - \delta$ if $n$ examples have been seen at this node and $\Delta\overline{G} > \epsilon$." [1] , [2]

*Claim.* The Hoeffding Bound does *not* provide the guarantee expected in [1].

*Proof.* Assume that the split candidates are $X, Y$ with IG values $G_X$ and $G_Y$, observed averages $\widehat{G_X}, \widehat{G_Y}$ and real averages $\overline{G_X}, \overline{G_Y}$ (cf. Figure 1). Considering

[1] In [1], this text is followed by a footnote on the "third-best and lower attributes" and on applying Bonferroni correction to $\delta$ if the attributes in the node are independent.

[2] Note: we use $\varepsilon$ instead of $\epsilon$, $\overline{Z}$ for the true average and $\widehat{Z}$ for the observed one.

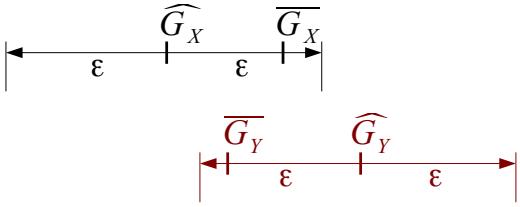


**Fig. 1.** Observed vs real averages of two random variables: the observed averages differ by more than $\varepsilon$, but the Hoeffding Bound does not guarantee that $G_Y$ is superior.

$n$ observations in range $R$ (of the split test), the probability that the real average $\overline{Z}$ deviates from the observed one $\widehat{Z}$ by more than $\varepsilon$ is bounded by Ineq. 1 [6]:

$$Pr(\widehat{Z} - \overline{Z} \geq \varepsilon) \leq exp(\frac{-2n\varepsilon^2}{R^2}) \tag{2}$$

In Figure 1, we see that $\widehat{G_Y}$ is greater then $\widehat{G_X}$ by more than $\varepsilon$, but this does not hold for the real averages $\overline{G_Y}$ and $\overline{G_X}$. Hence, a span of one $\varepsilon$ is not sufficient to guarantee separation of the gain values.

This claim holds also when we consider $G_X - G_Y$ as a single random variable $\Delta G$ (as done in [1]): the range of $\Delta G$ is the sum of ranges of $G_X$ and $G_Y$, again requiring a change of the decision bound. We give the correct bound in 4.1.

## 4 New Method for Correct Usage of the Hoeffding Bound

Our new core `correctedVFDT` encompasses a correction on the decision bound, and a new split function that satisfies the prerequisites of [6] (cf. section 3).

### 4.1 Specifying a Proper Decision Bound

Domingos and Hulten define $\Delta G = G_Y - G_X$ as a random variable with range $R = logc$ (for Information Gain IG, $c$ is the number of classes) and check whether $\widehat{\Delta G} - \overline{\Delta G}$ exceeds $\varepsilon$ [1], where $\varepsilon$ is a positive number. However, this definition of $\Delta G$ assumes that it is already non-negative, i.e. there exists some non-negative constant $k$, so that $|G_Y - G_X| \geq k$ holds.

Assume that there exists a $k > 0$ so that the true average [3] $E(|G_Y - G_X|)$ is $\geq k$. The absolute value is a convex function and $|G_Y - G_X|$ does not follow a degenerate distribution, so Jensen's inequality holds in its strict form, i.e.:

$$E(|G_Y - G_X|) > |E(G_Y - G_X)| \equiv |E(G_Y) - E(G_X)| \tag{3}$$

[3] We temporarily change the notation from $\overline{Z}$ to $E(Z)$ for better readability.

So, we cannot conclude that $|\overline{G_Y} - \overline{G_X}| \geq k$, i.e. even if the true average of $|G_Y - G_X|$ exceeds some positive value, we cannot say that $Y$ is superior to $X$.

We must thus perform *two* tests with the Hoeffding Inequality, (1) for $\Delta G_1 := G_Y - G_X$ under the assumption that $\Delta G_1 \geq 0$, *and* (2) for $-\Delta G_1 := G_X - G_Y$, assuming that $\Delta G_1 < 0$. Equivalently, we can perform a single *modified test* on a variable $\Delta G := G_Y - G_X$ that ranges over $[-\log c; +\log c]$, i.e. it may take negative values! Consequently, the new range of the variable $\Delta G$ that we denote as $R'$ is twice as high as the original range $R$. To apply the Hoeffding Inequality on such a variable, we must reset the decision bound to:

$$\varepsilon' = \sqrt{\frac{R'^2 \cdot \ln(1/\delta)}{2n}} = \sqrt{4\frac{R^2 \cdot \ln(1/\delta)}{2n}} = 2 \cdot \sqrt{\frac{R^2 \cdot \ln(1/\delta)}{2n}} \tag{4}$$

i.e. to twice the bound dictated by Ineq. 1. Then, the correctness of the split decision is guaranteed given $\delta$. Alternatively, we can keep the original decision bound and adjust the error-likelihood to $\delta^4$. Further, a larger number of instances is required to take a split decision. We study both effects in Section 6.

### 4.2 Specifying a Proper Split Function

Functions like Information Gain cannot be used in combination with the Hoeffding Inequality, because they are not arithmetic averages [10]. We term a split function that is an arithmetic average and satisfies the two prerequisites of the Hoeffding Inequality (cf. Section 3) as *proper*.

For a proper split function, we need to perform the computation of the expected quality of a node split on each element of the node independently. We propose *Quality Gain*, which we define as the improvement on predicting the target variable at a given node $v$ in comparison to its parent $Parent(v)$, i.e.

$$QGain(v) = Q(v) - Q(Parent(v)) \tag{5}$$

where the quality function $Q()$ is the normalized sum:

$$Q(v) = \frac{1}{|v|} \sum_{o \in v} oq(o) \tag{6}$$

and $oq()$ is a function that can be computed for each instance $o$ in $v$. Two possible implementations of $oq()$ are: *isCorrect*() (Eq. 7), whereas $Q()$ corresponds to the conventional accuracy, and *lossReduction*() (Eq. 8) that can capture the cost of misclassification in skewed distributions:

$$isCorrect(o) = \begin{cases} 1, & \text{if o is classified correctly} \\ 0, & \text{is misclassified} \end{cases} \tag{7}$$

$$lossReduction(o) = 1 - misclassificationCost(o) \tag{8}$$

We use *isCorrect*() to implement $oq()$ hereafter, and term the so implemented $QGain()$ function as *AccuracyGain*. However, the validation in the next Section

holds for all implementations of $oq()$. In the research regarding split measures the misclassification error has been indicated as a weaker metric than e.g. information gain [5]. Our goal is, however, not to propose a metric that yields higher accuracy of a model, but one that can be used together with the Hoeffding Bound without violating its prerequisites and thus allowing for interpretation of the performance guarantees given by this bound. In Section 6.2 we show that this metric is competitive to information gain in terms of accuracy and it reveals further positive features important for a streaming scenario.

## 5 Validation

We first show that our new split function satisfies the prerequisites of the Hoeffding Inequality. Next, we show that no correction for multiple testing is needed.

### 5.1 Satisfying the Assumptions of the Hoeffding Bound

*Quality Gain*, as defined in Eq. 5 using a quality function as in Eq. 6, satisfies the PREREQUISITES of the Hoeffding Inequality. PREREQ 1 (cf. Section 3) says that the random variable has to be almost surely bounded. The implementations of $oq()$ in Eq. 7, range in $[0, 1]$ and the same holds for the quality function $Q()$ in Eq. 6 by definition. Hence PREREQ 1 is satisfied.

PREREQ 2 (cf. Section 3) demands independent observations. In stream mining, the arriving data instances are always assumed to be independent observations of an unknown distribution. However, as we have shown in subsection 3.1, when Information Gain is computed over a sliding window, the content overlap and the combination of the instances for the computation of entropy lead to a violation of PREREQ 2. In contrast, our *Quality Gain* considers only one instance at each time point for the computation of $Q()$ and builds the arithmetical average incrementally, without considering past instances. This ensures that the instances are statistically independent from each other. The *Quality Gain* metric uses those independent instances to compute the goodness of a split. The result of this computation depends, however, on the performance of the classifier. Since, we consider a single node in a decision tree, the classifier and the entire path to the given node remains constant during the computation of the Hoeffding Bound. Consequently, all instances that fall into that node are conditionally independent given the classifier. This conditional independence of instances given the classifier allows us to use the Hoeffding Bound upon our split function.

### 5.2 Do We Need to Correct for Multiple Testing?

As explained in subsection 4.1, the split decision of `correctedVFDT` requires two tests on the same data sample: we compute $\varepsilon$ for the best and second-best attributes. Since the likelihood of witnessing a rare event increases as the number of tests increases, it is possible that the $\alpha$-errors (errors of first type) accumulate.

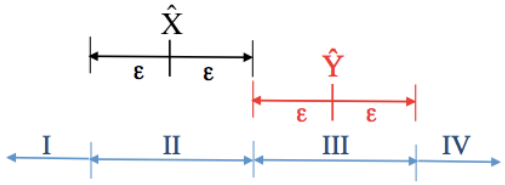


**Fig. 2.** When stating that $Y$ is superior to $X$ with confidence $1-\delta$, the error likelihood is $\delta$; error and non-error areas are represented by numbers I - IV.

To verify whether a correction for multiple tests (e.g. Bonferroni correction) is necessary, we consider the different possible areas of value combinations separately. The areas, enumerated as I-IV, are depicted in Figure 2.

Figure 2 depicts a situation where the Hoeffding Bounds of attributes $X$ and $Y$ are separable, and allow us to state with confidence $1-\delta$ that $Y$ is superior to $X$. There is a chance of $\delta$ that this statement is wrong. We distinguish three cases for variable $X$ (and equivalently for $Y$):

**Case (1):** the true average $\overline{X}$ is indeed in the $\varepsilon$-vicinity of $\widehat{X}$: $\widehat{X}-\varepsilon \leq \overline{X} \leq \widehat{X}+\varepsilon$ (area represented by II in Figure 2)
**Case (2):** $\overline{X}$ is left to the $\varepsilon$-vicinity of $\widehat{X}$: $\overline{X} < \widehat{X}-\varepsilon$ (area I)
**Case (3):** $\overline{X}$ is right to the $\varepsilon$-vicinity of $\widehat{X}$: $\overline{X} > \widehat{X}+\varepsilon$ (areas III and IV)

According to the Hoeffding Inequality, the likelihood of the Case (1) is $1-\delta$; we denote this case as `normal` or (n). We assume that the likelihood of error $\delta$ is distributed symmetrically around the $\varepsilon$-vicinity of $\widehat{X}$, hence the likelihood of Case (2) and of Case (3) is equal to $\delta/2$. In Case (2), the real average $\overline{X}$ is at the left of the $\varepsilon$-vicinity, hence the split decision would be the same as in Case (1). Therefore, we mark Case (2) as `not_harmful` (nh). In contrast, Case (3) for variable $X$ may lead to a different split decision, because we would incorrectly assume that $\overline{X}$ is higher than it truly is. This is represented by areas III and IV in Figure 2. We mark Case (3) as `harmful` (h).

In Figure 3 we show all possible combinations of cases and their likelihoods. This tree depicts the likelihood of the outcome of each combination; the middle level corresponds to the first test, the leaf-level contains the outcomes after the first and the second test. For instance, the left node on the middle level denotes the `not_harmful` (nh) error of the first test. At its right we see the `normal` case (n) with likelihood $1-\delta$. The leaf nodes of the tree represent the likelihood of outcomes after performing two tests: green nodes correspond to the `not_harmful` outcomes (n), (nh); red ones are potentially `harmful` (h); the blue ones contain both `harmful` and `not_harmful` outcomes.

Even if we consider all blue solid nodes as `harmful`, the sum of the likelihoods of `harmful` outcomes (cf. Eq. 9) is still smaller than $\delta$, hence a correction for multiple tests (e.g. Bonferroni correction) is not necessary.

$$\frac{\delta^2}{4} + \frac{\delta}{2}(1-\delta) + \frac{\delta^2}{4} + \frac{\delta}{2}(1-\delta) + \frac{\delta^2}{4} = \delta - \frac{\delta^2}{4} \tag{9}$$

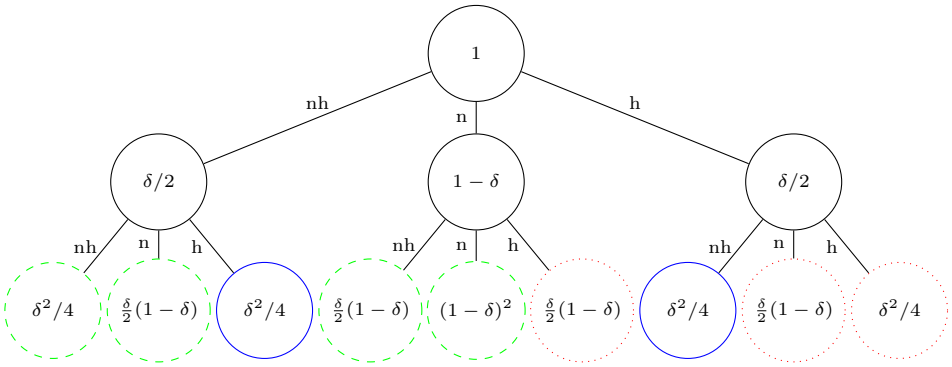


**Fig. 3.** Likelihood of all possible test outcomes. The middle level of the tree stands for outcomes of the first test. The leafs correspond to likelihood of outcomes after performing two tests. Green dashed leafs stand for `no error` (n) or `not_harmful` error (nh). Red dotted ones denote `harmful` error (h). Blue solid leafs combine `harmful` and `not_harmful` errors, so they have no label.

## 6 Experiments

We evaluate our `correctedVFDT` with *oq*() implemented as *isCorrect*() (Eq. 7), i.e. with *AccuracyGain* as our split function (cf. end of Section 4). We measure the impact of the modifications to VFDT [1] on classifier performance.

To quantify the impact of the improper use of the Hoeffdingin Inequality we use two indicators: the number of *Incorrect Decisions* and the average number of instances (*Average n*) considered before taking a split decision. For this experiment, a dataset with known ground truth is necessary. The artificial dataset and the experiment are described in 6.1.

When experimenting on real data, we quantify the performance of the stream classifier as *Avg. Accuracy* and the tree size as *Avg. # Nodes*. For this experiment, presented in subsection 6.2, we use the Adult dataset from the UCI repository[2].

### 6.1 Experimenting under Controlled Conditions

For the experiment under controlled conditions, we generate a dataset with a discrete multivariate distribution, as described in Table 1. The dataset has two attributes: $A_1$ with three discrete values in $\{A, B, C\}$, and $A_2$ with two discrete values in $\{D, E\}$. The target variable takes values from $\{c_1, c_2\}$.

In this experiment, we simulate a decision tree node and observe what split decision is taken in it. Since the distribution of the dataset is known , the attribute that each split function should choose at each moment is known. As we show in Table 2, we consider VFDT with IG - denoted as 'InfoGain' (cf. first two rows of Table 2 below the legend) for a decision bound of $1\varepsilon$ and $2\varepsilon$, and we compare with our `correctedVFDT` with *Accuracy Gain* - denoted as 'AccuracyGain' (cf. last two rows of Table 2), again for $1\varepsilon$ and $2\varepsilon$. This means that

| $A_2$ | | D | | E | |
|---|---|---|---|---|---|
| | A | $c_1 : 0.0675$ | $c_2 : 0.1575$ | $c_1 : 0.0675$ | $c_2 : 0.1575$ |
| $A_1$ | B | $c_1 : 0.1350$ | $c_2 : 0.0900$ | $c_1 : 0.1575$ | $c_2 : 0.0675$ |
| | C | $c_1 : 0.0450$ | $c_2 : 0.0050$ | $c_1 : 0.0450$ | $c_2 : 0.0050$ |

**Table 1.** Joint probability distribution of the synthetic dataset

| Setup | Incorrect Decisions | Average n | Alternative Hypothesis | p-value |
|---|---|---|---|---|
| InfoGain, $1\varepsilon$ | 25738 | 117.31 | P(incorrect decision) $> 0.05$ | $< 2.2e-16$ |
| InfoGain, $2\varepsilon$ | 1931 | 1671.53 | P(incorrect decision) $< 0.05$ | $< 2.2e-16$ |
| AccuracyGain, $1\varepsilon$ | 3612 | 17.68 | P(incorrect decision) $< 0.05$ | $< 2.2e-16$ |
| AccuracyGain, $2\varepsilon$ | 22 | 37.45 | P(incorrect decision) $< 0.05$ | $< 2.2e-16$ |

**Table 2.** Results of 100 000 repetitions of decision process on a split attribute at a node in a decision tree. We compare VFDT with 'InfoGain' to `correctedVFDT` with 'AccuracyGain' for the incorrect invocation of the Hoeffding Inequality (decision bound $1\varepsilon$) and for the correct invocation (decision bound $2\varepsilon$). For the performance indicators 'Incorrect Decisions' and 'Average n' lower values are better. The last column shows the results of the significance test on the deviation of the measured error from the theoretically permitted one, depicted in the 'Alternative Hypothesis' column, where the error-likelihood $\delta$ of the Hoeffding Bound is set to 0.05.

we consider both the erroneous decision bound $1\varepsilon$ and the corrected invocation of the Inequality with $2\varepsilon$ (cf. 4.1) for both VFDT and `correctedVFDT`.

In Table 2 we show the results, aggregated over 100,000 runs. In the second column, we count the 'Incorrect Decisions' over a total of 100,000 decisions. The third column 'Average n' counts the number of instances seen before deciding to split a node. The confidence level of the Hoeffding Inequality was set to $1 - \delta = 0.95$, hence only 5,000 (5%) incorrect split decisions are theoretically permitted. We run a binomial test to check whether the difference between the observed error and the expected one is significant (column before last) and return the computed $p$-value (last column of Table 2).

The original VFDT ($1^{st}$ row below legend in Table 2) exceeds the theoretical threshold of 5000 Incorrect Decisions by far. The corrected invocation of the Hoeffding Inequality (subsection 4.1) reduces the number of incorrect decisions by 92.497% (cf. $2^{nd}$ row in Table 2), but at the cost of increasing the number of instances required to make a split from 117.31 to 1671.53. This means that the learner would wait approximatively 10 times longer to take a split decision and would abstain from possibly good split decisions. In contrast, `correctedVFDT` makes less incorrect decisions and decides sooner, as seen in the last two rows of Table 2. The $3^{rd}$ row shows that even with the incorrect decision bound, `correctedVFDT` makes less incorrect decisions than the theoretic threshold. Best results are achieved for the correct decision bound of course ($4^{th}$ row): only 22 of the total 100,000 decisions are wrong, corresponding to an improvement of

| Setup | Incorrect Decisions | Average n | Alternative Hypothesis | p-value |
|---|---|---|---|---|
| InfoGain, $1\varepsilon$ | 14034 | 347.55 | P(incorrect decision) $> 0.01$ | $< 2.2e-16$ |
| InfoGain, $2\varepsilon$ | 339 | 2872.24 | P(incorrect decision) $< 0.01$ | $< 2.2e-16$ |
| AccuracyGain, $1\varepsilon$ | 1062 | 22.42 | P(incorrect decision) $< 0.01$ | 0.9757 |
| | | | P(incorrect decision) $> 0.01$ | 0.02617 |
| AccuracyGain, $2\varepsilon$ | 2 | 49.7 | P(incorrect decision) $< 0.01$ | $< 2.2e-16$ |

**Table 3.** Results analogous to those in Table 2, but with a confidence level of 0.99.

99.915 %. At the same time, our method for $2\varepsilon$ needs only 2.24% of the instances needed by VFDT, i.e. `correctedVFDT` converges much sooner than VFDT.

To ensure that these results are statistically significant we present the results of the binomial tests. The alternative hypothesis in the $4^{th}$ column in Table 2 differs from row to row. In the first row, the alternative hypothesis says that the number of incorrect decisions will be higher than the theoretic bound (by the Hoeffding Inequality); the p-value in the last column states that the alternative hypothesis should be accepted already at a confidence level lower then $2.2e-16$. Hence, the theoretical bound *is clearly* violated by the original VFDT. The alternative hypothesis in the other three rows states that the number of incorrect decisions will stay within bound; this hypothesis is accepted.

In Table 3, we compare VFDT to `correctedVFDT` at a confidence level $1-\delta$ =99%. The results are similar to Table 2, except for the `correctedVFDT` with incorrect decision bound: the theoretic bound is violated (significantly, see last column), i.e. even a good method will ultimatively fail if the Hoeffding Inequality is invoked erroneously: both the corrected decision bound and a proper split function are necessary for good performance (see last row).

### 6.2 Experiments on a Real Dataset

We have shown that the `correctedVFDT` with *Accuracy Gain* and correct decision bound ($2\varepsilon$) leads to an essential reduction of incorrect split decisions and that the decisions are taken much sooner. We now investigate how these new components influence the classification performance and size of created models on a real dataset. We use the dataset "Adults" from the UCI repository [2].

Stream mining algorithms are sensitive to the order of data instances and to concept drift. To minimize the effect of concept drift in the dataset, we created 10 permutations of it and repeated our tests on each permutation, setting the grace period of each run to 1. Therefore, the results presented in this section are averages over ten runs. This also increases the stability of the measures and lowers the effect of random anomalies.

For the two algorithms, we used the parameter settings that lead to best performance. For VFDT, these were $1-\delta = 0.97$ and decision bound $\varepsilon = 0.05$, i.e. the invocation of the Hoeffding Inequality is incorrect. According to subsection

4.1, the true confidence is therefore much lower. For `correctedVFDT`, the correct decision bound $2\varepsilon$ was used, the confidence level was set to $1 - \delta = 0.6$. The second column of Table 4 shows the average accuracy over the 10 runs, the third columns shows the average number of nodes of the models built in all runs.

| Algorithm | Avg. Accuracy | Avg. # Nodes |
|---|---|---|
| VFDT | 81,992 | 863.7 |
| `correctedVFDT` | 80,784 | 547.2 |

**Table 4.** Performance of VFDT and `correctedVFDT` of it on the "Adult dataset". The columns "Avg. Accuracy" and "Avg. # Nodes" denote the accuracy and the number of nodes of the decision trees, as averaged over ten runs.

According to the results in Table 4, VFDT reached a high accuracy, but it also created very large models with 863.7 nodes on average. That high amount of nodes not only consumes a lot of memory, but it also requires much computation time to create such models. Furthermore, such extensive models often tend to overfit the data distribution. In the second row of the table we see that `correctedVFDT` maintained almost the same accuracy, but needed only 63.36% of the nodes that were created by VFDT.

Our `correctedVFDT` does not only have the advantage of lower computation costs regarding time and memory usage, but also a split confidence that is interpretable. As we have shown in the previous subsection, the Hoeffding Bound of the VFDT cannot be trusted, for it does not bound the error the way it is expected. Consequently, setting the split confidence to 0.97 does not mean that the split decisions are correct with this level of confidence. In contrast to that, our method does not violate the requirements for using the Hoeffding Bound and thus, we can rely on the split decisions with the confidence that we have set.

For this particular amount of data and concept contained in this dataset (approximatively) optimal results have been achieved using the confidence of 0.6. This is much lower than 0.97 used with the VFDT, but this is only an illusory disproportion. In fact, the confidence guaranteed by the VFDT was much lower due to the violations of the requirements of the Hoeffding bound and it is probably not possible to estimate it. Usage of our method allows to interpret the results. We can see that it is necessary to give up the high confidence to achieve the best result on a so small dataset.

## 7 Conclusions

We have shown that the prerequisites for the use of the Hoeffding Inequality in stream classification are not satisfied by the VFDT core [1] and its successors. In a controlled experiment, we have demonstrated that the prerequisite violations do have an impact in classifier performance.

To alleviate this problem, we have first shown that the Hoeffding Inequality must be invoked differently, to cater for an input that may take negative values. We have adjusted the decision bound accordingly. We have further specified a family of split functions that satisfies the Inequality's prerequisites and incorporated them to our new core `correctedVFDT`. Our experiments on synthetic data show that `correctedVFDT` has significantly more correct split decisions and needs less instances to make a decision than the original VFDT. Our experiments on real data show that `correctedVFDT` produces smaller models, converges faster and maintains a similar level of accuracy. More importantly, the performance results of `correctedVFDT` are reliable, while those of the original VFDT are not guaranteed by the Hoeffding Inequality.

We are currently extending `correctedVFDT` to deal with concept drift. Further, we want to explicate the premises under which arithmetical averages and more elaborate computations on the arriving stream (as in [10] for McDiarmid's Inequality) satisfy the prerequisite of independence. In our future work we are also going to investigate the performance of Accuracy Gain, its robustness to noise and concept drift on further datasets.

## References


1. P. Domingos and G. Hulten. Mining High Speed Data Streams. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2000.
2. A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010.
3. J. Gama, R. Rocha, and P. Medas. Accurate decision trees for mining high-speed data streams. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, New York, NY, USA, 2003. ACM.
4. João Gama and Petr Kosina. Learning Decision Rules from Data Streams. In Toby Walsh, editor, *IJCAI*, pages 1255–1260. IJCAI/AAAI, 2011.
5. Trevor Hastie, Robert Tibshirani, Jerome Friedman, and Ebooks Corporation. *The Elements of Statistical Learning*, chapter 9.2.3, pages 324–329. Springer, Dordrecht, 2009.
6. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.
7. G. Hulten, L. Spencer, and P. Domingos. Mining Time-Changing Data Streams. *ACM SIGKDD*, 2001.
8. Petr Kosina and João Gama. Very Fast Decision Rules for multi-class problems. In Sascha Ossowski and Paola Lecca, editors, *SAC*, pages 795–800. ACM, 2012.
9. Bernhard Pfahringer, Geoffrey Holmes, and Richard Kirkby. New Options for Hoeffding Trees. In *Australian Conference on Artificial Intelligence*, pages 90–99, 2007.
10. Leszek Rutkowski, Lena Pietruczuk, Piotr Duda, and Maciej Jaworski. Decision Trees for Mining Data Streams Based on the McDiarmid's Bound. *IEEE Trans. on Knowledge and Data Engineering*, 2012. accepted in 2012.
11. Wenhua Xu, Zheng Qin, Hao Hu, and Nan Zhao. Mining Uncertain Data Streams Using Clustering Feature Decision Trees. In Jie Tang, Irwin King, Ling Chen, and Jianyong Wang, editors, *ADMA (2)*, volume 7121 of *Lecture Notes in Computer Science*, pages 195–208. Springer, 2011.