New splitting criteria for decision trees in stationary data streams

Maciej Jaworski, Piotr Duda, Leszek Rutkowski, *Fellow, IEEE*

***Abstract*—The most popular tools for stream data mining are based on decision trees. In previous 15 years all designed methods, headed by the VFDT algorithm, relayed on the Hoeffding's inequality and hundreds of researchers followed this scheme. Recently, we have demonstrated that although the Hoeffding decision trees are an effective tool for dealing with stream data, they are a purely heuristic procedure; for example classical decision trees like ID3 or CART cannot be adopted to data stream mining using the Hoeffding's inequality. Therefore, there is an urgent need to develop new algorithms which are both mathematically justified and characterized by good performance. In this paper we address this problem by developing a family of new splitting criteria for classification in stationary data streams and investigating their probabilistic properties. The new criteria, derived using appropriate statistical tools, are based on the misclassification error and the Gini index impurity measures. The general division of splitting criteria into two types is proposed. Attributes chosen based on type-I splitting criteria guarantee, with high probability, the highest expected value of split measure. Type-II criteria ensure that the chosen attribute is the same, with high probability, as it would be chosen based on the whole infinite data stream. Moreover, in this paper two hybrid splitting criteria are proposed which are the combinations of single criteria based on the misclassification error and Gini index.**

***Index Terms*—Data stream, decision trees, classification, impurity measure, splitting criterion**

I. Introduction

In this paper a problem of data stream classification is considered. A data stream [1], [2], [3], [4], [5], [6] is a potentially infinite sequence of data elements which arrive continuously to the system, often with very high rate. Due to these characteristics traditional data classification algorithms, designed for static data, cannot be directly applied in this case. In case of static data the access to data elements is constantly available, hence they may be processed by the algorithm as many times as needed. On the other side, each data element from the stream usually can be processed at most once because of memory or computational power limitations. Existing methods should be modified or new algorithms should be developed taking the mentioned limitations into account. Data stream classification algorithms should be resource-aware [7]. Another issue associated with data streams is the occurrence of concept drift [8], [9], [10], [11], [12]. The distribution of stream data elements values may evolve in time. Although this issue is not addressed in this paper it should be noted that the concept drift also poses many difficulties in designing appropriate classification algorithms. In literature there is a variety of methods designed for data classification. The most popular are artificial neural networks [13], k-nearest neighbors [14] and decision trees [15], [16], [17]. In this paper the application of decision trees for data stream mining is considered. Models learned using decision trees have many advantages. They are easily interpretable for users and demonstrate relatively low memory complexity comparing to other methods. This paper focuses on the most critical point of decision tree induction algorithms, i.e. the choice of a splitting attribute in a considered node. In static decision trees the 'best' attribute is chosen on the basis of the data sample S available in the considered node. The choice is made based on some split measure function. The value of this function is calculated for all attributes and the attribute which provides the highest value of split measure is chosen as the splitting one. In the most popular decision tree algorithms, like ID3 [15] or CART [16], the split measure function is proposed as a reduction of some impurity measure $g(S)$. Let us assume that data elements are characterized by D attributes and can belong to one of K classes. Let $n(S)$ denote the number of data elements in S and let $n^k(S)$ denote the number of data elements from the set S which belong to the k-th class. Then the most popular impurity measures, i.e. information entropy, Gini index and misclassification error, are expressed in the following way, respectively

$$g^E(S) = -\sum_{k=1}^{K} \frac{n^k(S)}{n(S)} \log_2 \left(\frac{n^k(S)}{n(S)} \right), \tag{1}$$

$$g^G(S) = 1 - \sum_{k=1}^{K} \left(\frac{n^k(S)}{n(S)} \right)^2, \tag{2}$$

$$g^M(S) = 1 - \frac{\max_k \{n^k(S)\}}{n(S)}. \tag{3}$$

In data stream scenario the problem of choosing the most relevant attribute is more complicated since data elements arrive continuously to the considered node. A special splitting criterion is required to determine if the current number of data elements $n(S)$ in the node is sufficient to make a proper split. In [18] the authors proposed the Very Fast Decision Tree

L. Rutkowski is with the Institute of Computational Intelligence, Czestochowa University of Technology, ul. Armii Krajowej 36, 42-200 Czestochowa, Poland, and also with Information Technology Institute, University of Social Sciences, 90-113 Łódź, Poland (e-mail: leszek.rutkowski@iisi.pcz.pl)

M. Jaworski and P. Duda are with the Institute of Computational Intelligence, Czestochowa University of Technology, ul. Armii Krajowej 36, 42-200 Czestochowa, Poland, (e-mail: maciej.jaworski@iisi.pcz.pl, piotr.duda@iisi.pcz.pl)

This work was supported by the Polish National Science Center under Grant No. 2014/15/B/ST7/05264

(VFDT) algorithm. The core of this method is the Hoeffding tree, in which the Hoeffding's inequality is used to obtain the bound for the difference between split measure values for two attributes. Let $\Delta g_{i_{max}}(S)$ and $\Delta g_{i_{max2}}(S)$ denote values of some split measure function for attributes with the highest and the second highest value of this split measure, respectively. The bound obtained in [18] was given in the following form

$$\Delta g_{i_{max}}(S) - \Delta g_{i_{max2}}(S) > \epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n(S)}}, \quad (4)$$

where R is the range of values for the applied impurity measure, e.g. $R = \log_2 K$ for information entropy. The authors of this bound claimed that if it is satisfied, then the i_{max}-th attribute is with probability at least $1-\delta$ the same as it would be chosen based on the whole infinite data stream. Unfortunately there are several mistakes made in this approach. First, the Hoeffding's bound is valid only for the sum of random variables, hence it cannot be applied to nonlinear functions such as information entropy or Gini index. Although it can be applied to the misclassification error (for the first time it is shown in this paper - see subsection II-A), the range R appearing in the bound ϵ should be twice larger than it is proposed in [18] since the difference between two values of split measure is investigated. Additionally, the authors of [18] made a simplifying assumption that the attributes other than the i_{max}-th and i_{max2}-th provide sufficiently smaller split measure values. In consequence the probability that these attribute provide the highest split measure value according to the whole data stream is negligible. Under this assumption the probability that the i_{max}-th attribute is the same as would be chosen using the whole data stream is at least equal to 1δ. However, the assumption from [18] is not always true and actually the mentioned probability can be bounded by $(1-\delta)^{D-1}$. The problem with mathematical inconsistency concerning bound (4) was pointed out by various authors [19], [20], [21], [22]. In [19] the McDiarmid's inequality was proposed instead of the Hoeffding's one as a statistical tool to obtain splitting criteria for information gain and Gini gain. In [20] new impurity measures were proposed, for which the Hoeffding's inequality can be properly applied, and a correction for bound (4) proposed originally in [18] was suggested. In [23] and [24] the authors proposed a method which combines the Taylor's theorem with the Gaussian approximation and applied it to information gain and Gini gain, respectively. A split measure function based on the impurity measure called the misclassification error was introduced in [25] for the problem of decision trees induction for data streams. The Gaussian approximation was used to obtain an appropriate splitting criterion in this case. Additionally, new hybrid splitting criterion combining advantages of misclassification error and Gini index was proposed as well. In the same paper also an analytical example of violating the splitting criterion based on bound (4) for simple dataset was presented. In [21], [22] the authors derived bounds for biases of estimators for both information gain and Gini gain. In [26] the authors applied the Hoeffding's inequality not directly to the values of split measure function but to its arguments expressed as appropriate fractions of data elements.

Despite of the aforementioned misstatements concerning bound (4), it should be noted that the Hoeffding's tree still can be considered as a heuristic tool for data stream mining. In fact, for many datasets it provides very satisfactory results of classification accuracy. For this reason the Hoeffding's tree and its derivatives stand as a basis for many new methods used in data stream classification [9], [27], [28], [29], [30], [31].

In view of the aforementioned mistakes concerning bound (4), new splitting criteria are required, based on solid mathematical foundations. In this work two impurity measures are considered, i.e. misclassification error (3) and Gini index (2). Based on these functions corresponding split measures can be defined. For convenience in further investigations only binary trees will be considered, i.e. it is assumed that the partition of set S with respect to the i-th attribute provides two disjoint complementary subsets: S_{L_i} and S_{R_i}.

It is important to notice that this paper not only brings significant novelty in the theoretical aspect of data stream classification using decision trees. It also proposes to generalize the heuristic bound given in the right hand side of formula (4) and to induce decision trees with the bound lower than the one proposed in original Hoeffding decision trees (see Remark 1 and formula (5)). It should be noted that in this paper, like in [18] where the Hoeffding trees were introduced, only stationary data streams are considered, i.e. the problem of concept drift is not taken into account. In brief, the main contribution of this paper can be summarized as follows:

- A novel framework for defining splitting criteria is proposed. More precisely splitting criteria used in the online decision trees are divided into two types, depending on the interpretation of result they provide. Type-I criterion guarantees that, with assumed probability, the attribute chosen based on it maximizes the expected value of split measure. On the other hand type-II criterion ensures that, with assumed probability, the chosen attribute is the same as it would be chosen based on the whole infinite data stream.
- A new type-I splitting criterion (see Corollary 1) is proposed for accuracy gain split measure function which is based on the rarely used impurity measure called misclassification error. The Hoeffding's inequality is used, in a proper manner, to obtain the desired criterion.
- A new type-I splitting criterion (see Corollary 2) is presented for split measure called Gini gain (which is based on the Gini index impurity measure). It is obtained using the McDiarmid's inequality and it significantly improves our previous result [19]. Such an improvement allows to reduce four times the number of data elements required to make a decision.
- A type-II splitting criterion (see Corollary 3) is obtained for the Gini gain split measure. It merges the previously mentioned criterion with the bound for Gini gain bias values derived in [22].
- Two hybrid splitting criteria are proposed (see Corollaries 4 and 5), which combine component (single) criteria obtained for different impurity measures, i.e. the misclas-

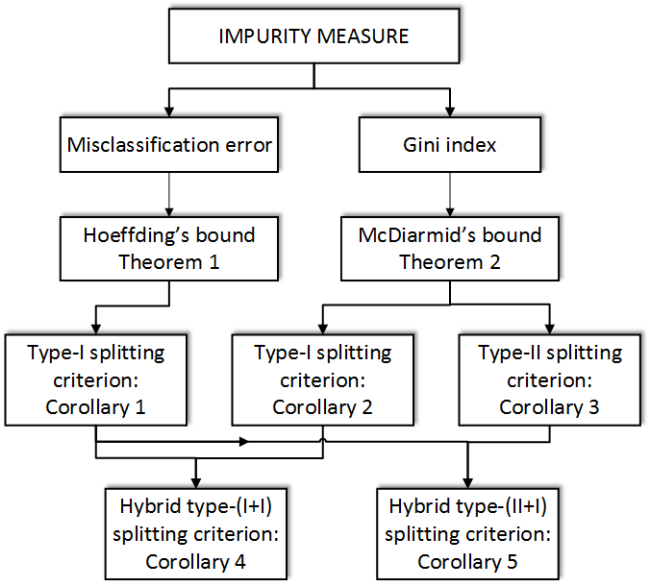

Fig. 1. Flowchart of new results (splitting criteria) presented in this paper.

sification error and the Gini index, and we claim that such a hybridization, based on theoretically justified splitting criteria, leads to the best performance of decision trees. The first one is a hybrid type-$(I + I)$ criterion which merges the type-I splitting criterion for misclassification error and type-I criterion for Gini index. The second one is a hybrid type-$(II + I)$ criterion. In this case the type-I criterion for Gini index is replaced by the type-II one.

- The classification accuracies of Online Decision Trees with various splitting criteria proposed in this paper are compared with each other. A comparison of proposed solutions with original Hoeffding's Trees described in [18] is presented as well.

The flowchart of all splitting criteria proposed in this paper is depicted in Fig. 1.

Remark 1: It is worth noticing that bound (4) proposed originally in [18], its correction in [20], as well as our bounds (16) and (30) proposed in Theorem 1 and Theorem 2 of this paper, respectively, can all be expressed in the following general form

$$\epsilon = C\sqrt{\frac{\ln(1/\delta)}{n(S)}}. \tag{5}$$

Constant C for bounds (4), (16) and (30) is equal to $R/\sqrt{2}$, $\sqrt{2}$ and $\sqrt{8}$, respectively. Whereas the last two values are theoretically justified by Theorems 1 and 2, the value corresponding to bound (4) does not have mathematical foundations (i.e. it is heuristic), as it was previously stated. Therefore, nothing stands in the way to construct a splitting criterion with a bound equal to, for example, half of the value given in (4), i.e. with $C = 0.5R/\sqrt{2}$. Decision trees containing such a splitting criterion are also compared with solutions proposed in this paper and the results are very surprising - the mentioned decision trees provide even better accuracies than Hoeffding decision trees (see section IV-E).

The rest of the paper is organized as follows. In section II new splitting criteria are presented in details. Each single splitting criterion is supported by an appropriate mathematical theorem. Additionally, two hybrid splitting criteria are proposed, which combine single criteria obtained for different split measures. In section III a general form of the Online Decision Tree is recalled. It is equivalent to the Hoeffding's Tree algorithm presented in [18], however, the unjustified splitting criterion is replaced by the proper ones. Additionally, a modified version of that algorithm suited for hybrid splitting criteria is also presented. In section IV the results of experimental simulations are depicted. Section V concludes the paper.

II. New splitting criteria

In this paper two types of splitting criteria are considered. Each type is characterized by a different interpretation of the result it provides. The distinction between two types results from the fact that the extrapolation of standard (batch) decision tree algorithms into the online scenario can be done in (at least) two ways. The attributes chosen to split a node have a slightly different interpretation in the case of each splitting criterion type. The considerations are valid only if the probability distribution of data elements does not change in time, i.e. the data stream is stationary. Let $\Delta g_i(S)$ be some split measure function calculated for the i-th attribute based on sample of data elements S. Then $E\left[\Delta g_i(S)\right]$ is its expected value. Moreover, let Δg_i denote the value of split measure which will be obtained for the whole infinite data stream (i.e. if the cardinality of set S is infinite). It is important to note that if the estimator $\Delta g_i(S)$ is biased, then $E\left[\Delta g_i(S)\right]$ differs from Δg_i. This difference determines that there are two types of splitting criteria:

I) type-I splitting criterion guarantees that the i_{max}-th attribute, chosen based on it, provides with probability at least $(1-\delta)^{D-1}$ the highest expected value of split measure function for $n(S)$ data elements among all attributes, i.e.

$$E\left[\Delta g_{i_{max}}(S)\right] = \max_{i\in\{1,\ldots,D\}}\{E\left[\Delta g_i(S)\right]\}. \tag{6}$$

II) type-II splitting criterion guarantees that the i_{max}-th attribute, chosen based on it, is the same, with probability at least $(1-\delta)^{D-1}$, as it would be chosen based on the whole infinite data stream, i.e.

$$\Delta g_{i_{max}} = \max_{i\in\{1,\ldots,D\}}\{\Delta g_i\}. \tag{7}$$

The distinction between the type-I and type-II criteria can be understood as follows. Let S_∞ be a data stream with an infinite number of data elements. If we could take all of them, calculate split measure function values for each attribute and choose the one with the highest value, then with probability at least $(1-\delta)^{D-1}$ it would be the attribute chosen based on a data sample using the type-II splitting criterion (if,

obviously, the criterion was satisfied). Now let us partition the stream S_∞ into an infinite number of subsets S_i, each having exactly n data elements in it. Let us calculate for each attribute an arithmetic average of split measure values for all subsets S_i and, as previously, choose the attribute with the highest obtained value. Then, this attribute is with probability at least $(1-\delta)^{D-1}$ the same as the attribute chosen based on the data sample of size n using the type-I splitting criterion (if, obviously, the criterion was satisfied). Since the estimator of split measure function calculated using a sample of n elements is biased (at least for the information gain, Gini gain and accuracy gain), then these two types of criteria can result in choosing different attributes. Actually, in the literature so far, including the basic paper [18], but also [19], [20], [23]- [25], the type-I splitting criteria were considered. Recently only in [21], [22] the authors introduced the bias term into the criteria and, although they did not call it in this way, for the first time proposed the type-II splitting criteria.

In [22] a bound for difference between information gain values for two attributes was presented. The bound, adapting it to notations used in this paper, is given in the following form

$$\epsilon = 2\ln\left(n(S)\right)\sqrt{\frac{2\ln\left(4/\delta\right)}{n(S)}} + \frac{4}{n(S)}. \tag{8}$$

Based on this bound a type-II splitting criterion for decision trees with information gain split measure can be created. After neglecting the bias term in (8) another form of bound can be proposed

$$\epsilon = 2\ln\left(n(S)\right)\sqrt{\frac{2\ln\left(4/\delta\right)}{n(S)}}, \tag{9}$$

which can be used to construct a type-I splitting criterion. However, as it will be presented in experimental section, bounds (8) and (9) are impractical in real applications. Due to the logarithmic term, the number of data elements needed to make a decision about the split is rather large. In consequence, the accuracy of induced decision tree grows very slow as the processing of data stream continues. Therefore, in this section new splitting criteria for split measures based on misclassification error (accuracy gain) and Gini index (Gini gain) are proposed. In the case of misclassification error only type-I splitting criterion is obtained, whereas for Gini index splitting criteria of both types are presented.

A. Misclassification error

Based on the misclassification error (3), similarly to the accuracy-based quality gain proposed in [20], a split measure called the accuracy gain can be proposed as a difference between misclassification error for set S and weighted average of misclassification error values for sets S_{L_i} and S_{R_i}

$$\Delta g_i^M(S) = g^M(S) - \frac{n(S_{L_i})}{n(S)}g^M(S_{L_i}) - \frac{n(S_{R_i})}{n(S)}g^M(S_{R_i}). \tag{10}$$

Combining (3) with (10) one obtains

$$\Delta g_i^M(S) = g^M(S) - 1 + \frac{\max_k\{n^k(S_{L_i})\} + \max_k\{n^k(S_{R_i})\}}{n(S)}. \tag{11}$$

In further considerations only differences between split measures for different attributes will be taken into account. Since the term $g^M(S)-1$ does not depend on attribute, it can be neglected. The truncated accuracy gain is considered, i.e.

$$g_i^M(S) = \frac{\max_k\{n^k(S_{L_i})\} + \max_k\{n^k(S_{R_i})\}}{n(S)}. \tag{12}$$

Obviously the following equality is true

$$\Delta g_i^M(S) - \Delta g_j^M(S) = g_i^M(S) - g_j^M(S). \tag{13}$$

The truncated accuracy gain (12) can be further simplified. The term $\max_k\{n^k(S_{L_i})\}$ denotes the number of elements reaching the 'left' leaf and belonging to the majority class in this leaf, if the split was made with respect to the i-th attribute. Observe that the simplest method of assigning a class to an unlabeled data element is to assign to it the majority class of the leaf. Therefore, the term $\max_k\{n^k(S_{L_i})\}$ is the number of correctly classified data elements of set S, which would be sorted down to the 'left' leaf. The same applies to the 'right' leaf. Hence the sum in the nominator of expression (12) is the total number of correctly classified data elements of set S, if the split was made with respect to the i-th attribute. As a result, function (12) is the accuracy obtained for elements of set S, if the split was made with respect to the i-th attribute. Hence, the choice of the attribute which maximizes the value of function (12) guarantees the maximum gain of the accuracy of the tree. Let $s_j,\ j=1,\ldots,n(S)$, denote the elements of the set S. Let us define a function $P_i(s)$ which is equal to 1 if the element s would be correctly classified (after splitting the considered node with respect to the i-th attribute) and 0 otherwise. It is obvious that $P_i(s)$ is a random variable from the binomial distribution with some unknown expected value μ_i and variance $\mu_i(1-\mu_i)$. Now function (12) can be expressed in the following way

$$g_i^M(S) = \frac{\sum_{j=1}^{n(S)} P_i(s_j)}{n(S)}. \tag{14}$$

Obviously, values of function (14) tend to μ_i as the size of the set S tends to infinity.

The accuracy gain (11) can be used to construct the splitting criterion for decision tree. Since the truncated accuracy gain (12) can be expressed as a sum of random variables (14), the Hoeffding inequality can be applied to derive an appropriate bound. Below the Hoeffding's theorem is recalled. In this paper the Hoeffding's inequality (see [32]) is cited as lemma 1.

Lemma 1. *Let X_i, $i=1,\ldots,n$, be independent random variables which satisfy $Pr(X_i \in [a_i;b_i]) = 1$, and let $R_i = b_i - a_i$. Then the following inequality holds*

$$Pr\left(\frac{\sum_{i=1}^{n} X_i}{n} - E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] > \epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(R_i)^2}\right) = \delta. \quad (15)$$

Now a theorem concerning the comparison of accuracy gain values for two different attributes can be introduced.

Theorem 1 (The Hoeffding's inequality for split measure called the accuracy gain (11) based on the misclassification error (3))**.** *Let S be the set of data elements (independent random variables) and let $\Delta g_i^M(S)$ denote the accuracy gain for set S and for the i-th attribute. If the following condition is satisfied*

$$\Delta g_i^M(S) - \Delta g_j^M(S) > \sqrt{\frac{2\ln(1/\delta)}{n(S)}}, \quad (16)$$

then with probability at least $(1-\delta)$ the following inequality also holds

$$E\left[\Delta g_i^M(S)\right] > E\left[\Delta g_j^M(S)\right]. \quad (17)$$

Proof: Since truncated accuracy gain values $g_i^M(S)$ and $g_j^M(S)$ are expressed as a sum of random variables (14), the Hoeffding's inequality can be applied to them

$$P\left(g_i^M(S) - g_j^M(S) - (E\left[g_i^M(S)\right] - E\left[g_j^M(S)\right]) < \epsilon\right) < \exp\left(\frac{-2n^2(S)\epsilon^2}{\sum_{p=1}^{n(S)}(R_p)^2}\right) = \delta. \quad (18)$$

where in this case $R_p = 2$, $p = 1, \ldots, n(S)$. Hence inequality (18) can be expressed as follows

$$P(g_i^M(S) - g_j^M(S) - (E\left[g_i^M(S)\right] - \left[g_j^M(S)\right]) < \epsilon) < \exp\left(\frac{-n(S)\epsilon^2}{2}\right) = \delta. \quad (19)$$

From inequality (19) the following statement can be derived: If

$$g_i^M(S) - g_j^M(S) > \epsilon = \sqrt{\frac{2\ln(1/\delta)}{n(S)}}, \quad (20)$$

then, with probability at least $1-\delta$ the following inequality is true

$$E\left[g_i^M(S)\right] - E\left[g_j^M(S)\right] > 0,. \quad (21)$$

According to (13), assumption (16) enforces inequality (20) to be true. According to (13) inequality (21) is equivalent to conclusion (17) of Theorem 1. ∎

Theorem 1 can be used to determine, with probability at least $1-\delta$, whether the relation between accuracy gain values of two attributes, the i-th and the j-th, calculated from data sample S is the same as the relation between their expected values. This theorem can be also used to determine whether the attribute with the highest value of accuracy gain calculated from data sample S provides the highest expectation value. This fact can be used to construct a type-I splitting criterion for decision trees.

Corollary 1 (Type-I splitting criterion based on misclassification error)**.** *Let i_{max} and i_{max2} denote the indices of the attributes with the highest and the second highest, respectively, values of accuracy gain calculated from data sample. If:*

$$\Delta g_{i_{max}}^M(S) - \Delta g_{i_{max2}}^M(S) > \sqrt{\frac{2\ln(1/\delta)}{n(S)}}, \quad (22)$$

then, according to Theorem 1, with probability at least $(1-\delta)^{D-1}$ also the following statement holds

$$i_{max} = \arg\max_{i\in\{1,\ldots,D\}}\{E\left[\Delta g_i^M(S)\right]\} \quad (23)$$

and the i_{max}-th attribute can be chosen to split the considered node.

B. Gini index

Analogously to the accuracy gain (11) one can define the split measure function based on the Gini index (2), called the Gini gain. It is a difference between the value of Gini index for set S and the weighted average of Gini index values for sets S_{L_i} and S_{R_i}

$$\Delta g_i^G(S) = g^G(S) - \frac{n(S_{L_i})}{n(S)}g^G(S_{L_i}) - \frac{n(S_{R_i})}{n(S)}g^G(S_{R_i}). \quad (24)$$

Combining (2) with (24) one obtains

$$\Delta g_i^G(S) = g^G(S) - 1 + \frac{1}{n(S)}\sum_{k=1}^{K}\left[\frac{\left(n^k(S_{L_i})\right)^2}{n(S_{L_i})} + \frac{\left(n^k(S_{R_i})\right)^2}{n(S_{R_i})}\right]. \quad (25)$$

Neglecting the term $g^G(S) - 1$, the truncated Gini gain can be obtained

$$g_i^G(S) = \frac{1}{n(S)}\sum_{k=1}^{K}\left[\frac{\left(n^k(S_{L_i})\right)^2}{n(S_{L_i})} + \frac{\left(n^k(S_{R_i})\right)^2}{n(S_{R_i})}\right]. \quad (26)$$

Analogously to (13) the difference between Gini gain values and the difference between truncated Gini gain for two attributes obey the following equality

$$\Delta g_i^G(S) - \Delta g_i^G(S) = g_i^G(S) - g_i^G(S). \quad (27)$$

In a way similar to the previous subsection the splitting criterion for decision tree based on the Gini index can be proposed. However, the truncated Gini gain (26) cannot be expressed as a sum of random variables. Therefore the Hoeffding's inequality cannot be applied to it. More general statistical tool is required, i.e. the McDiarmid's inequality which will be recalled below. The McDiarmid's inequality (see [33]) is cited here as lemma 2.

Lemma 2. *Let $X_1, \ldots, X_n$ be independent random variables and let $f : A_1 \times \cdots \times A_n \to R$ be a function satisfying the following conditions*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i} |f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,\hat{x}_i,\ldots,x_n)| \leq c_i,\ i = 1,\ldots,n. \tag{28}$$

Then the following inequality holds

$$Pr\left(f(X_i,\ldots,X_n) - E\left[f(X_i,\ldots,X_n)\right] > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(c_i)^2}\right) = \delta. \tag{29}$$

Now a theorem concerning the comparison of Gini gain values for two different attributes can be introduced.

Theorem 2 (The McDiarmid's bound for split measure called Gini gain (25) based on the Gini index (2))**.** *Let S be the set of data elements (independent random variables) and let $\Delta g_i^G(S)$ be the Gini gain value for set S and for the i-th attribute. If the following condition is satisfied*

$$\Delta g_i^G(S) - \Delta g_j^G(S) > \sqrt{\frac{8\ln(1/\delta)}{n(S)}}, \tag{30}$$

then, with probability at least $1-\delta$, the following inequality holds

$$E\left[\Delta g_i^G(S)\right] > E\left[\Delta g_j^G(S)\right]. \tag{31}$$

Proof: The proof is similar to the one presented in [19] or [22]. The set S consists of $n(S)$ data elements $s_1,\ldots,s_{n(S)}$. Let us denote by $\hat{S}_h$ the set of data elements $s_1,\ldots,\hat{s}_h,\ldots,s_{n(S)}$, i.e. sets S and $\hat{S}_h$ differ only in the h-th data element. First of all the bound for the difference $g_i^G(S) - g_i^G(\hat{S}_h)$ will be established. The component quantities used to compute the truncated Gini gain are $n(S_{L_i})$, $n(S_{R_i})$, $n^k(S_{L_i})$, $n^k(S_{R_i})$, $k = 1,\ldots,K$. There are mainly two possible cases when replacing data element s_h in set S by data element $\hat{s}_h$ in set $\hat{S}_h$

i) $n(\hat{S}_{h,L_i}) = n(S_{L_i}) - 1$,
$n(\hat{S}_{h,R_i}) = n(S_{R_i}) + 1$,
$n^p(\hat{S}_{h,L_i}) = n^p(S_{L_i}) - 1$,
$n^q(\hat{S}_{h,R_i}) = n^q(S_{R_i}) + 1$,
$n^k(\hat{S}_{h,L_i}) = n^k(S_{L_i})\ k \neq p$,
$n^k(\hat{S}_{h,R_i}) = n^k(S_{R_i})\ k \neq q$.

ii) $n(\hat{S}_{h,L_i}) = n(S_{L_i})$,
$n(\hat{S}_{h,R_i}) = n(S_{R_i})$,
$n^p(\hat{S}_{h,L_i}) = n^p(S_{L_i}) - 1$,
$n^q(\hat{S}_{h,L_i}) = n^q(S_{L_i}) + 1$,
$n^k(\hat{S}_{h,L_i}) = n^k(S_{L_i})\ k \neq p, q$,
$n^k(\hat{S}_{h,R_i}) = n^k(S_{R_i})\ k \in \{1,\ldots,K\}$.

There are two more cases obtained after exchanging L_i and R_i with each other in above equalities. However, they obviously provide the same final results as the cases presented above. Simple but tedious algebra leads to the following

$$\sup_{s_1,\ldots,s_{n(S)},\hat{s}_h} |g_i^G(S) - g_i^G(\hat{S}_h)| \leq \frac{2}{n(S)}. \tag{32}$$

From (32) the following inequality can be derived

$$\sup_{s_1,\ldots,s_{n(S)},\hat{s}_h} |g_i^G(S) - g_j^G(S) - (g_i^G(\hat{S}_h) - g_j^G(\hat{S}_h))| \leq \frac{4}{n(S)}. \tag{33}$$

Therefore, the assumption (28) of the McDiarmid's theorem is satisfied for difference $g_i^G(S) - g_j^G(S)$ with $c_h = \frac{4}{n(S)}$, $h = 1,\ldots,n(S)$. Then, according to the McDiarmid's theorem, the following inequality holds

$$Pr\left(g_i^G(S) - g_j^G(S) - E\left[g_i^G(S) - g_j^G(S)\right] > \epsilon\right) \leq \exp\left(\frac{n(S)\epsilon^2}{8}\right) = \delta. \tag{34}$$

From inequality (34) the following statement can be derived:
If

$$g_i^G(S) - g_j^G(S) > \epsilon = \sqrt{\frac{8\ln(1/\delta)}{n(S)}}, \tag{35}$$

then, with probability at least $1-\delta$, the following inequality is true

$$E\left[g_i^G(S)\right] - E\left[g_i^G(S)\right] > 0. \tag{36}$$

According to (27) inequality (35) is equivalent to assumption (30) of Theorem 2. According to (27) inequality (36) leads to conclusion (31) of Theorem 2.

■

Theorem 2 determines the bound which should be satisfied to predict, with probability at least $1-\delta$, the proper relation between expected values $E\left[\Delta g_i^G(S)\right]$ and $E\left[\Delta g_j^G(S)\right]$. This theorem 2 can be furthermore used to determine whether the attribute with the highest value of Gini gain calculated from data sample S provides also the highest expected value. This fact can be used to establish a splitting criterion for decision trees.

Corollary 2 (Type-I splitting criterion based on Gini index)**.** *Let i_{max} and i_{max2} denote the indices of attributes with the highest and the second highest, respectively, values of Gini gain calculated from data sample. If:*

$$\Delta g_{i_{max}}^G(S) - \Delta g_{i_{max2}}^G(S) > \sqrt{\frac{8\ln(1/\delta)}{n(S)}}, \tag{37}$$

then, according to Theorem 2, with probability at least $(1-\delta)^{D-1}$ also the following statement holds

$$i_{max} = \arg\max_{i\in\{1,\ldots,D\}}\{E\left[\Delta g_i^G(S)\right]\} \tag{38}$$

and the i_{max}-th attribute can be chosen to split the considered node.

Theorem 2 concerns the relation between expected values $E\left[\Delta g_i^G(S)\right]$ and $E\left[\Delta g_j^G(S)\right]$. However, this relation doesn't

give a correct answer about the relation between the values of Gini gain for the whole data stream, i.e. when the number of data elements in S is infinite. In this case the information about the bias is required. In [21] the authors estimated the bias of truncated Gini gain $g_i^G(S)$

$$|E\left[g_i^G(S)\right] - g_i^G| \leq \frac{4}{\sqrt{n(S)}}, \tag{39}$$

where g_i^G is the value of truncated Gini gain obtained for the whole infinite data stream. According to (27) and (39) the bias of difference between Gini gain values is given by

$$|E\left[\Delta g_i^G(S) - \Delta g_j^G(S)\right] - \left(\Delta g_i^G - \Delta g_j^G\right)| \leq \frac{8}{\sqrt{n(S)}}. \tag{40}$$

Theorem 2 together with inequality (40) can be used to determine whether the attribute with the highest value of accuracy gain calculated from data sample S provides the highest value of Gini gain calculated for the whole infinite data stream. This fact can be used to construct a splitting criterion for decision trees.

Corollary 3 (Type-II splitting criterion based on Gini index)**.** *Let i_{max} and i_{max2} denote the indices of attributes with the highest and the second highest, respectively, values of Gini gain calculated from data sample. If:*

$$\Delta g_{i_{max}}^G(S) - \Delta g_{i_{max2}}^G(S) > \sqrt{\frac{8\ln(1/\delta)}{n(S)}} + \frac{8}{\sqrt{n(S)}}, \tag{41}$$

then, according to Theorem 2 and inequality (40), with probability at least $(1-\delta)^{D-1}$ also the following statement holds

$$i_{max} = \arg\max_{i\in\{1,\ldots,D\}}\{\Delta g_i^G\} \tag{42}$$

and the i_{max}-th attribute can be chosen to split the considered node.

C. Hybrid splitting criteria

The splitting criteria proposed in subsections II-A and II-B are complementary in some sense. These criteria will be referred as single splitting criteria as opposed to the hybrid criteria proposed in this subsection. As it will be pointed out in the experimental section of this paper, decision tree with splitting criterion based on the misclassification error provides the fast growth of the tree at the beginning stages of its development. Then, for large number of data elements $n(S)$, it remains stable at some unsatisfactory level. On the other hand the decision tree based on the Gini index grows slowly at the beginning. The size of the tree is correlated with its classification accuracy. The accuracy increases as the new data elements from the stream are processed and for some large enough number $n(S)$ the classification accuracy of the tree with Gini index starts to exceed the accuracy obtained for the tree based on misclassification error. The splitting criteria proposed in previous subsections can be merged together into hybrid criteria, which reveal advantages of both components. Two hybrid splitting criteria are presented below: the first one joining inequalities (22) and (37) and the second one joining inequalities (22) and (41).

Corollary 4 (Type-$(I+I)$ hybrid splitting criterion based on Gini index and misclassification error)**.** *Let $i_{G,max}$ and $i_{G,max2}$ denote the indices of attributes with the highest and the second highest, respectively, values of Gini gain calculated from data sample. If:*

$$\Delta g_{i_{G,max}}^G(S) - \Delta g_{i_{G,max2}}^G(S) > \sqrt{\frac{8\ln(1/\delta)}{n(S)}}, \tag{43}$$

then, according to Theorem 2, with probability at least $(1-\delta)^{D-1}$, also the following statement holds

$$i_{G,max} = \arg\max_{i\in\{1,\ldots,D\}}\{E\left[\Delta g_i^G(S)\right]\} \tag{44}$$

and the $i_{G,max}$-th attribute can be chosen to split the considered node. If the split was not made, then another condition is additionally checked. Let $i_{M,max}$ and $i_{M,max2}$ denote the indices of attributes with the highest and the second highest, respectively, values of accuracy gain calculated from data sample. If:

$$\Delta g_{i_{M,max}}^M(S) - \Delta g_{i_{M,max2}}^M(S) > \sqrt{\frac{2\ln(1/\delta)}{n(S)}}, \tag{45}$$

then, according to Theorem 1, with probability at least $(1-\delta)^{D-1}$, also the following statement holds

$$i_{M,max} = \arg\max_{i\in\{1,\ldots,D\}}\{E\left[\Delta g_i^M(S)\right]\} \tag{46}$$

and the $i_{M,max}$-th attribute can be chosen to split the considered node.

Corollary 5 (Type-$(II+I)$ hybrid splitting criterion based on Gini index and misclassification error)**.** *Let $i_{G,max}$ and $i_{G,max2}$ denote the indices of attributes with the highest and the second highest, respectively, values of Gini gain calculated from data sample. If:*

$$\Delta g_{i_{G,max}}^G(S) - \Delta g_{i_{G,max2}}^G(S) > \sqrt{\frac{8\ln(1/\delta)}{n(S)}} + \frac{8}{\sqrt{n(S)}}, \tag{47}$$

then, according to Theorem 2 and inequality (40), with probability at least $(1-\delta)^{D-1}$, also the following statement holds

$$i_{G,max} = \arg\max_{i\in\{1,\ldots,D\}}\{\Delta g_i^G\} \tag{48}$$

and the $i_{G,max}$-th attribute can be chosen to split the considered node. If the split was not made, then another condition is additionally checked. Let $i_{M,max}$ and $i_{M,max2}$ denote the indices of attributes with the highest and the second highest, respectively, values of accuracy gain calculated from data sample. If:

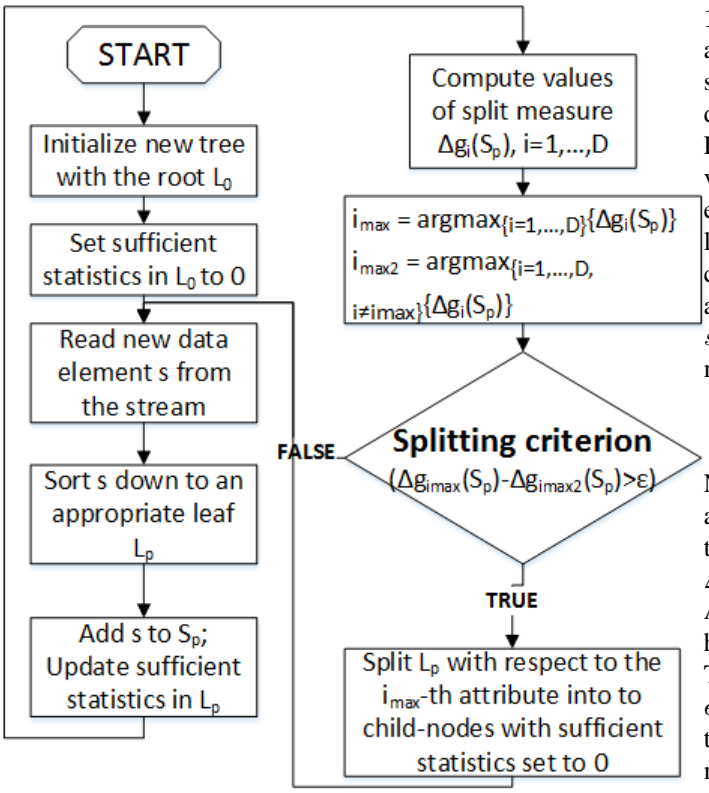

Fig. 2. Block diagram of the ODT algorithm with standard (single) splitting critera.

$$\Delta g^M_{i_{M,max}}(S) - \Delta g^M_{i_{M,max2}}(S) > \sqrt{\frac{2\ln(1/\delta)}{n(S)}}, \tag{49}$$

then, according to Theorem 1, with probability at least $(1-\delta)^{D-1}$*, also the following statement holds*

$$i_{M,max} = \arg\max_{i\in\{1,\ldots,D\}} \{E\left[\Delta g^M_i(S)\right]\} \tag{50}$$

and the $i_{M,max}$*-th attribute can be chosen to split the considered node.*

III. A General Form of the Online Decision Tree Algorithm

All the algorithms considered in this paper are based on the idea of the Hoeffding Tree algorithm presented in [18]. Since the Hoeffding's inequality is not the only statistical tool used to derive splitting criteria in this work (the McDiarmid's inequality is used as well), the name 'Hoeffding's tree' will be replaced simply by the 'Online Decision Tree (ODT)' further in the text. The only difference between all versions of the algorithm considered in this paper lies in the applied splitting criterion. The block scheme of the ODT algorithm with standard single splitting criteria, like (22), (41) or (37), is summarized in Fig. 2.

Let $n^k_{ij}(S)$ denote the number of data elements (in set S) from the k-th class, with the j-th value of the i-th attribute. The set of numbers $n^k_{ij}(S)$, $i = 1,\ldots,D$, $j = 1,\ldots,v_i$, $k = 1,\ldots,K$ is also called the sufficient statistics of set S. The algorithm starts with one single node - the root. The sufficient statistics in the root are all set to zero. Then, the tree is developed using subsequent data elements from the stream. Each data element s is sorted down the tree, according to the values of attributes and the current structure of the tree. An element s finally reaches a leaf L_p. The sufficient statistics in leaf L_p are updated. Let S_p denote the set of data elements collected so far in the leaf L_p, $j_{i,s}$ denote the value of the i-th attribute of data element s and k_s denote the class of element s. The update of sufficient statistics is made in the following manner

$$\forall_{i\in\{1,\ldots,D\}}\ n^{k_s}_{ij_{i,s}}(S_p) = n^{k_s}_{ij_{i,s}}(S_p) + 1. \tag{51}$$

Next, values of the applied split measure function $\Delta g_i(S_p)$ are calculated for each attribute, $i = 1,\ldots,D$. Obviously, if the accuracy gain is used as a split measure, then $\Delta g_i(S_p) \equiv \Delta g^M_i(S_p)$. In the case of Gini gain $\Delta g_i(S_p) \equiv \Delta g^G_i(S_p)$. Attributes with the highest (the i_{max}-th) and the second highest (the i_{max2}-th) values of split measure are chosen. Then, the splitting criterion is checked. The form of the bound ϵ corresponds to the applied splitting criterion. Depending on the impurity measure it can be type-I criterion (22) if the misclassification error is used or type-I criterion (37) or type-II criterion (41) in the case of Gini index. If the result of the test is positive, the leaf L_p becomes a node and is split into two children nodes (leaves). The sufficient statistics in both children nodes are initially set to zero. Then the whole procedure is performed for the next data element taken from the stream.

The algorithm has a slightly different form if hybrid splitting criteria are used. Corresponding block scheme is presented in Fig. 3. In this case both split measures should be calculated for each attribute, i.e. $\Delta g^G_i(S_p)$ and $\Delta g^M_i(S_p)$. For each of the two split measures attributes with the highest and the second highest values are chosen, i.e. $i_{G,max}$ and $i_{G,max2}$ for Gini index and $i_{M,max}$ and $i_{M,max2}$ for misclassification error, respectively. Then, the first part of hybrid splitting criterion is checked (denoted as 'Hybrid criterion, p.1' in Fig. 3). This part operates on the values of Gini gain. Type-I criterion (37) or type-II criterion (41) can be used. If the criterion is met, the leaf L_p is split into child nodes with respect to the $i_{G,max}$-th attribute. Otherwise, the second part of hybrid splitting criterion is checked (denoted as 'Hybrid criterion, p.2' in Fig. 3). It is type-I criterion (22), i.e. the values of accuracy gain are compared. If it is satisfied, then the considered leaf is split with respect to the $i_{M,max}$-th attribute.

IV. Simulation Results

In this section the performance of the ODT with proposed splitting criteria is discussed. Although this paper is mainly of theoretical importance, it is worth considering how the replacement of the splitting criterion proposed in [18] by new splitting criteria proposed in this work affects the classification accuracy of the decision tree. All simulations were conducted using our software implemented in C# language and it is available at www.iisi.pcz.pl/~pduda. For convenience some

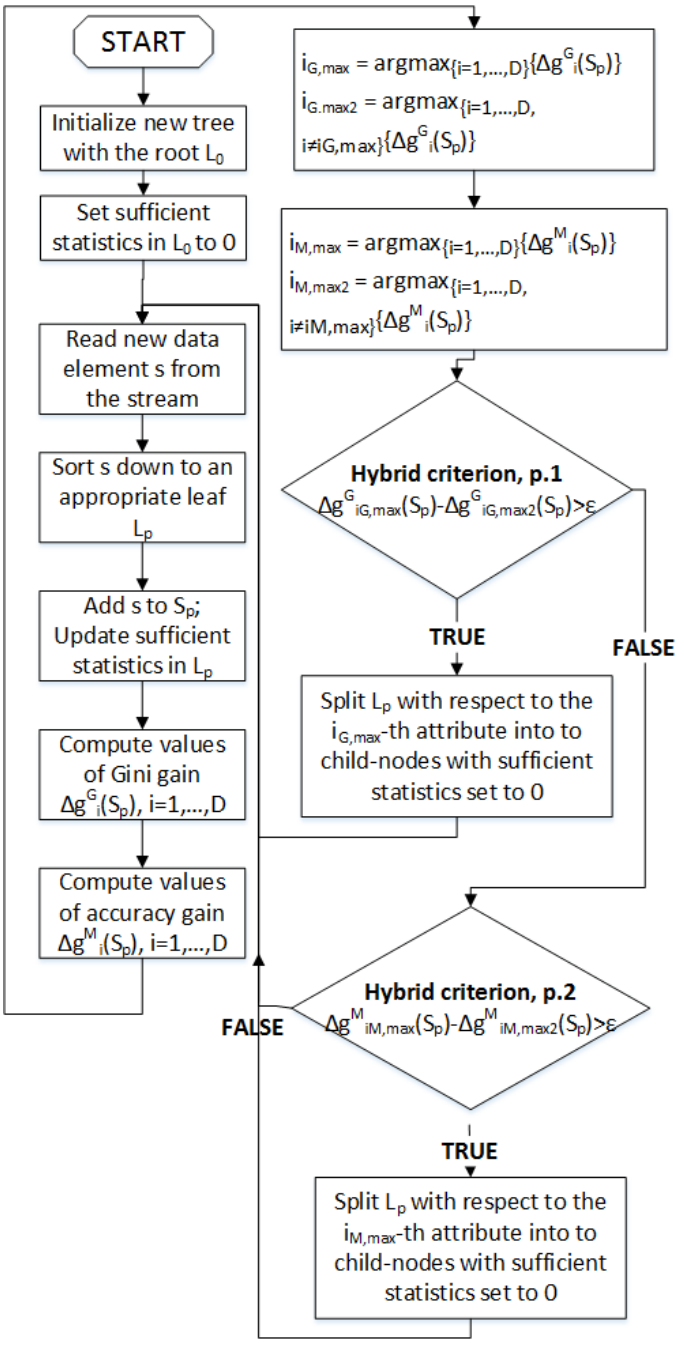

Fig. 3. Block diagram of the ODT algorithm with hybrid splitting critera.

abbreviations for ODTs with various splitting criteria will be now introduced:

- ODTm - Online Decision Tree based on the misclassification error with type-I splitting criterion (22);
- ODTG1 - Online Decision Tree based on the Gini index with type-I splitting criterion (37);
- ODTG2 - Online Decision Tree based on the Gini index with type-II splitting criterion (41);
- ODTEnt1 - Online Decision Tree based on the information entropy with type-I splitting criterion created using bound (9);
- ODTEnt2 - Online Decision Tree based on the information entropy with type-II splitting criterion created using bound (8);
- ODTh1 - Online Decision Tree based on the Gini index and the misclassification error with type-$(I+I)$ hybrid splitting criterion merging (37) and (22);
- ODTh2 - Online Decision Tree based on the Gini index and the misclassification error with type-$(II+I)$ hybrid splitting criterion merging (41) and (22);
- HDT - Online Decision Tree based on the Gini index with splitting criterion proposed in [18] (originally called the Hoeffding's Decision Tree);
- $1/2$HDT - Online Decision Tree based on the Gini index with splitting criterion containing the bound ϵ equal to the half of the value of the bound obtained in [18], i.e. $\epsilon = 0.5\sqrt{\frac{R^2 \ln(1/\delta)}{2n(S)}}$, where $R = 1$ for Gini index.

A. Data preparation

In this paper synthetic data were used to investigate the performance of the considered methods. Data were generated on a basis of synthetic decision trees. These synthetic trees were constructed in the same way as described in [18]. At each level of the tree, after the first d_{min} levels, each node is replaced by a leaf with probability ω. The higher value of parameter ω implicates the lower complexity of the tree. The maximum depth of the synthetic tree is d_{max} (at this level all nodes are replaced by leaves). After the structure of the whole tree is obtained, splitting attributes are randomly assigned to the nodes. At each path from the root to any leaf the assigned attributes must not repeat. To each leaf a class is randomly assigned. Different synthetic trees represent different data concepts. The data concept is a particular distribution of attributes values and classes. In brief, it determines the correlations between the attributes and classes. For the purpose of the simulations eight synthetic trees were generated (all of them with $D = 30$ binary attributes, $d_{min} = 3$ and $d_{max} = 18$). Four different values of ω were considered: 0.1, 0.15, 0.2 and 0.25. For each of these values two synthetic trees were generated. These eight synthetic trees provide eight different data concepts. For each concept a testing dataset consisting of 100000 elements was generated using the corresponding synthetic tree. Hence there are eight testing datasets, one for each data concept. Each testing dataset is used to evaluate the accuracy of the corresponding decision tree at every stage of its development. In the simulations presented below, for any training dataset size n and any value of parameter δ one obtains eight different decision trees and, as a result, eight values of classification accuracy (one for each synthetic data concept). The final result of accuracy for particular n and δ is calculated as the arithmetic average over all eight values.

B. Single splitting criteria

In this subsection the ODTm, ODTG1 and ODTG2 are compared. First, the dependence between the classification accuracy and parameter δ is investigated. The experiment was carried out for two different sizes of training dataset: $n(S) = 10^6$ and $n(S) = 10^8$. Five different values of δ were used, i.e. 0.00001, 0.0001, 0.001, 0.01 and 0.1. The obtained

TABLE I
THE MEAN AVERAGE AND STANDARD DEVIATION VALUES OF ONLINE DECISION TREES ACCURACY OBTAINED USING THREE SINGLE SPLITTING CRITERIA FOR VARIOUS VALUES OF δ AND $n = 10^6$ DATA ELEMENTS.

	ODTG1		ODTG2		ODTm	
δ	mean	std. dev.	mean	std. dev.	mean	std. dev.
0.00001	0.666	0.033	0.622	0.033	0.718	0.014
0.0001	0.666	0.033	0.622	0.033	0.718	0.014
0.001	0.670	0.033	0.637	0.035	0.721	0.014
0.01	0.672	0.032	0.651	0.031	0.722	0.014
0.1	0.722	0.023	0.656	0.034	0.724	0.013

TABLE II
THE MEAN AVERAGE AND STANDARD DEVIATION VALUES OF ONLINE DECISION TREES ACCURACY OBTAINED USING THREE SINGLE SPLITTING CRITERIA FOR VARIOUS VALUES OF δ AND $n = 10^8$ DATA ELEMENTS.

	ODTG1		ODTG2		ODTm	
δ	mean	std. dev.	mean	std. dev.	mean	std. dev.
0.00001	0.794	0.021	0.765	0.027	0.725	0.013
0.0001	0.796	0.021	0.767	0.026	0.725	0.013
0.001	0.799	0.021	0.769	0.026	0.725	0.013
0.01	0.808	0.022	0.777	0.024	0.725	0.013
0.1	0.824	0.021	0.781	0.023	0.725	0.013

average accuracies with standard deviations are depicted in Tab. I and II, respectively.

It can be seen that for both considered values of $n(S)$ the accuracy of each decision tree is nearly constant for δ in the range $[0.00001; 0.1]$. A slightly different (higher) accuracies in several cases were obtained for $\delta = 0.1$. It seems that the considered decision trees are insensitive on the value of δ if it is taken from the mentioned interval. Therefore, for further analysis only two values of δ are considered: 0.001 and 0.1. Although the comparison of accuracies of decision trees with various splitting criteria will be conducted comprehensively in next experiment for a wide range of possible numbers of processed data elements, one can perform an initial analysis of this issue based on Tables I and II. As can be seen, for $n(S) = 10^6$ the ODTm provides better accuracy for all considered values of δ. The differences were significant, what was confirmed by the Friedman's test on the level of significance 1%. Additionally Conover post-hoc test for pairwise comparisons demonstrated that the results obtained by the ODTm vary significantly from the results of the other algorithms. However, when $n(S) = 10^8$ data elements is taken into account, then decision trees with Gini index used as a split measure dominate. Similarly to the previous case, the significance difference between values was indicated by the Friedman's test and the post-hoc analysis confirm the advantage of the ODTG1 algorithm.

The comparison of accuracy dependence on the number of data elements for various decision trees is presented in Fig 4 and 5, respectively. Additionally to decision trees with previously considered splitting criteria, the ODTEnt1 is included into the analysis. The ODTEnt2 is not taken into account since it provides almost the same result as the ODTEnt1.

The decision tree with the misclassification error used as an impurity measure provides higher accuracy at the beginning stages of the tree induction. However, after around $n(S) = 10^6$ data elements processed, the accuracy does not increase

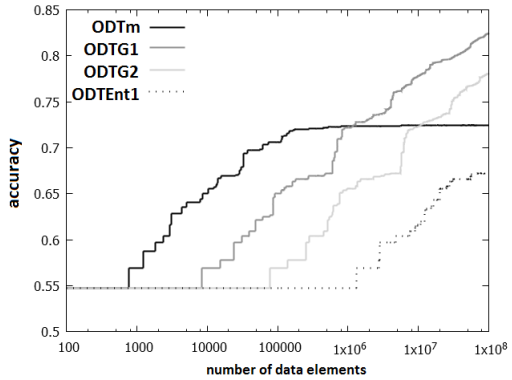

Fig. 4. Dependence between the average accuracy and number of training data elements for the ODTs with various single splitting criteria for $\delta = 0.1$.

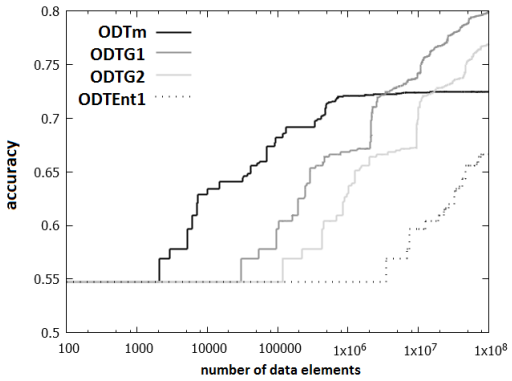

Fig. 5. Dependence between the average accuracy and number of training data elements for the ODTs with various single splitting criteria for $\delta = 0.001$.

significantly. On the other hand decision trees with Gini index start to increase the accuracy later and for large number of data elements they outperform the ODTm. Among the two decision trees with splitting criteria based on the Gini index the ODTG1 provides higher accuracy than the ODTG2. The plots corresponding to the ODTG2 have nearly identical shape as the plots for the ODTG1, however, they are moved to the 'right', i.e. to the higher values of $n(S)$. This means that the ODTG2 induces, in majority of cases, the same trees as the ODTG1 does, however, each split in the ODTG2 occurs later, i.e. after processing more data elements. It is understandable since both decision trees use the same impurity measure and the splitting criterion for the ODTG2 requires more data elements to make a decision about the split. The decision tree with information entropy used as an impurity measure occurs the worst in this comparison. Hence it will not be taken into account in further experiments.

C. Hybrid splitting criteria

Figures 4 and 5 clearly illustrate the reason why it is worth considering the fusion of splitting criteria based on misclassification error and Gini index into one hybrid splitting criterion. Each component criterion demonstrates its positives at different stages of tree development. One can expect that the hybrid splitting criterion reveals the advantages of both

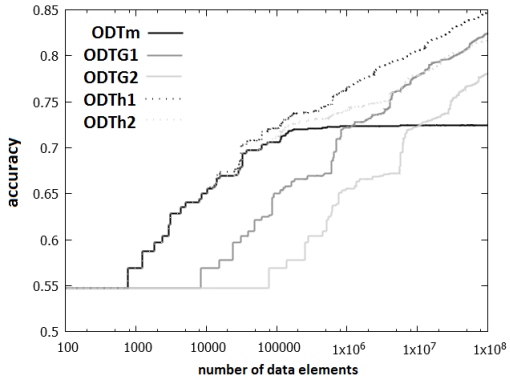

Fig. 6. Dependence between the average accuracy and number of training data elements for the ODTs with various single and hybrid splitting criteria for $\delta = 0.1$.

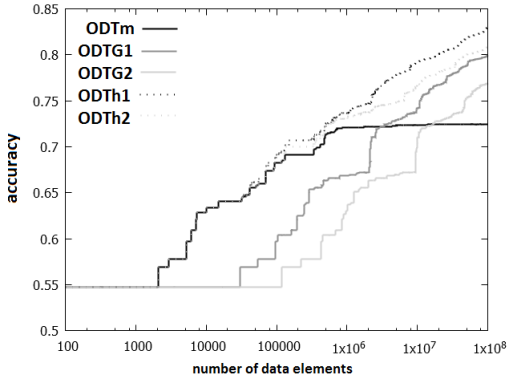

Fig. 7. Dependence between the average accuracy and number of training data elements for the ODTs with various single and hybrid splitting criteria for $\delta = 0.001$.

components. To prove this claim, the decision trees with hybrid splitting criteria were compared with decision trees investigated in previous subsection. The results for $\delta = 0.1$ and $\delta = 0.001$ are depicted in Fig. 6 and 7, respectively. Indeed, the ODTh1 demonstrates higher classification accuracy than both the ODTG1 and the ODTm for any number of data elements $n(S)$. Analogously, the same is true when comparing the ODTh2 with the ODTG2 and the ODTm. The accuracy results presented in Fig. 6 and 7 are obtained as an arithmetic average over the eight data concepts. In Fig. 8 the situation for four single data concepts is presented for values of ω, used to construct the corresponding synthetic tree, equal to 0.1, 0.15, 0.2 and 0.25. The parameter δ was set to 0.001. As in the case of average accuracy, in all four presented data concepts the accuracy of decision tree with hybrid splitting criterion outperforms the accuracy of decision trees with corresponding single criteria. The ODTh1 provides higher accuracy than the ODTh2. It is understandable since the ODTG1 outperforms the ODTG2, as it was demonstrated in the previous subsection. Looking for the reason why the decision trees with hybrid splitting criteria classify better than their counterparts with single criteria it is worth investigating how the number of leaves in decision tree grows as the number of processed data elements increases. This is visualized in Fig. 9. Comparing the

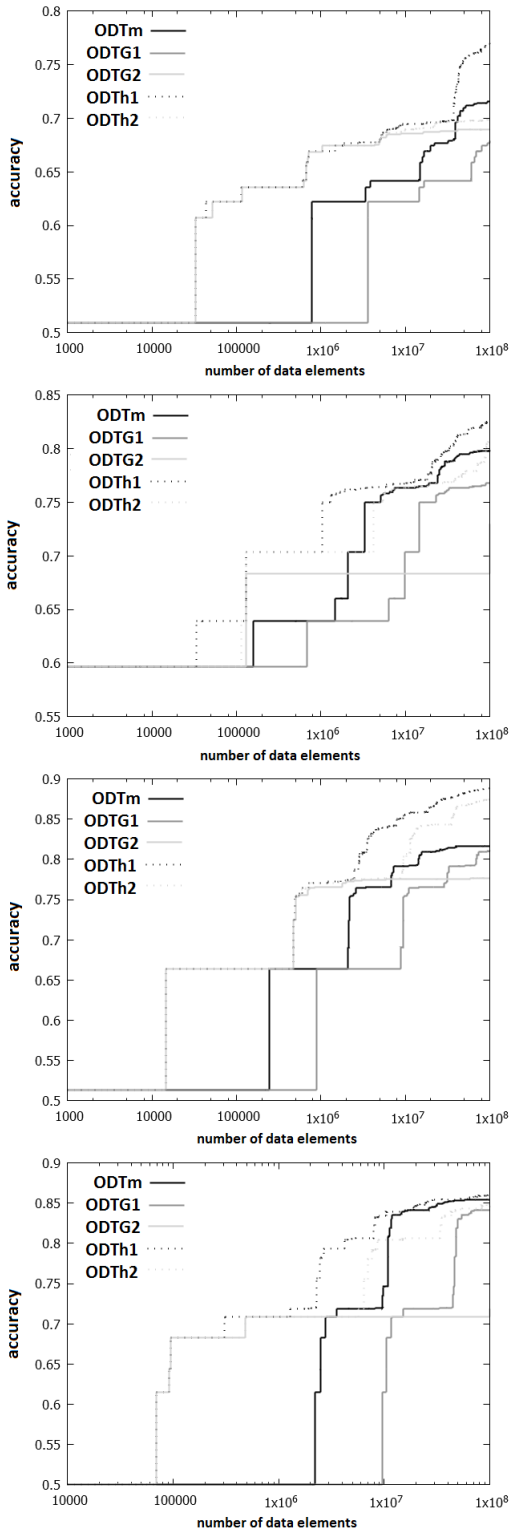

Fig. 8. Accuracies in the function of training dataset size for the ODTs with various single and hybrid splitting criteria for four data concepts with $\omega = 0.1, 0.15, 0.2, 0.25$ and for $\delta = 0.001$.

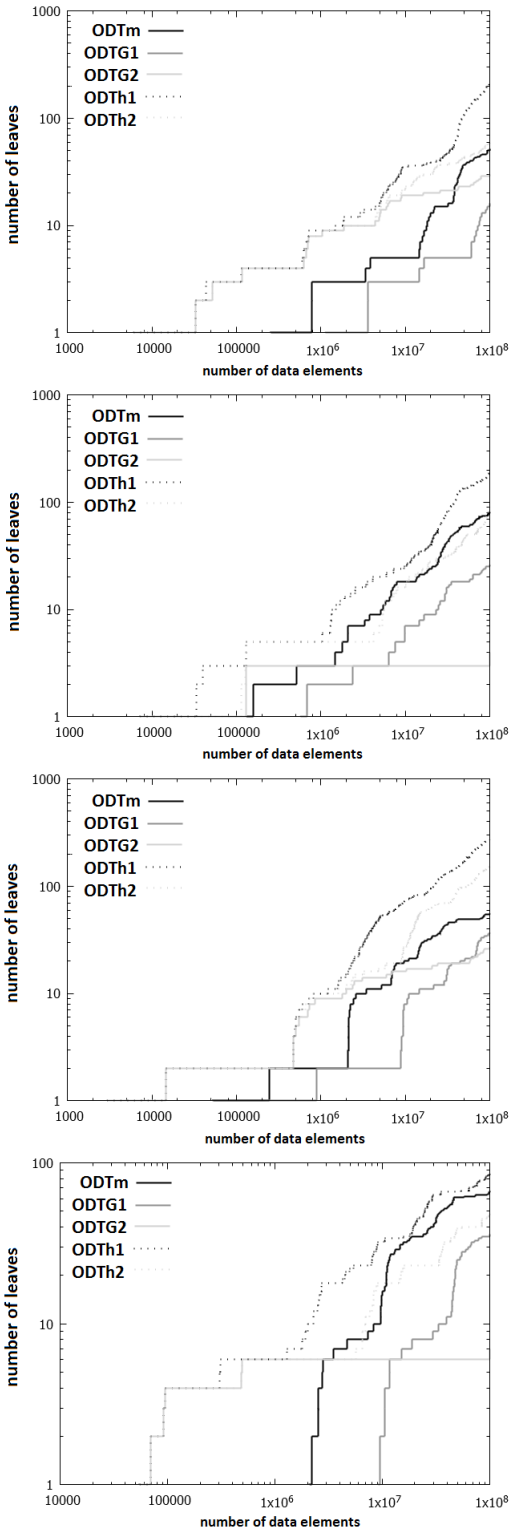

Fig. 9. Number of leaves in the function of training dataset size for the ODTs with various single and hybrid splitting criteria for four data concepts with $\omega = 0.1, 0.15, 0.2, 0.25$ and for $\delta = 0.001$.

corresponding plots for Fig. 8 and 9 it is easily seen that there exists a strong correlation between the classification accuracy of decision tree and the number of its leaves. The ODTs with hybrid splitting criteria provide many more splits than trees with single splitting criteria and for this reason they achieve higher classification accuracies. It should be noted that larger size of induced decision tree requires more memory to store it. It may be an important fact to take into account in data stream mining scenario. If there is not enough memory in the computing device it may occur that only a part of decision tree with hybrid splitting criterion may be induced. If on the other hand the same amount of memory would be enough to store the decision tree with a single splitting criterion it might occur that in such a case, for example, the ODTG1 would provide higher accuracy than the ODTh1. Nevertheless, if the amount of memory is enough for both trees, then the decision tree with a hybrid splitting criterion outperforms the tree with a single criterion.

D. Performance on real data sets

In this subsection two real data sets were analyzed, taken from the UCI machine learning repository [34]. The two chosen datasets consists of relatively large number of data elements, hence they can imitate the data stream. One of the considered datasets is the 'KDD CUP 99' with 4898431 data elements. Data are described by 41 attributes, 7 of which are nominal and 34 are numerical. Each data element belongs to one of the two classes - 'network attack' and 'normal network connection'. The data distribution in this set is skewed: the 'network attack' class constitutes about 80% of the whole dataset. The data set was divided into two equinumerous parts, first one used for learning and the second one for testing. Such procedure was performed four times, providing in total four training datasets and their complementary testing datasets. The accuracy evaluation was performed before every split of the node which occurred in the considered tree. The accuracy for one particular number of data elements is calculated as the average over four datasets described above. The second real set of data used in the following simulations is the 'Covertype' dataset, consisting of 581012 data elements. Data are characterized by 10 numerical and 44 nominal attributes. Each data element belongs to one of seven classes, representing forest cover type designation. The testing datasets were taken randomly from the training dataset in the same way as for the set 'KDD CUP 99'. For each dataset the simulations were conducted for two different values of parameter δ: 0.001 and 0.1. Results are presented in Fig. 10 and 11 for the 'KDD CUP 99' dataset and in Fig. 12 and 13 for the 'Covertype' dataset. For both used datasets it turned out that in the case of decision tree with hybrid splitting criterion, i.e. the ODTh1, no splits were made based on the Gini gain split measure. Hence, the results obtained for the ODTm and for the ODTh1 overlap.

E. Comparison with Hoeffding Decision Trees

Although, as it was previously stated, the HDT is a heuristic data mining tool with mathematically unjustified splitting criterion, it is worth comparing it with decision trees proposed

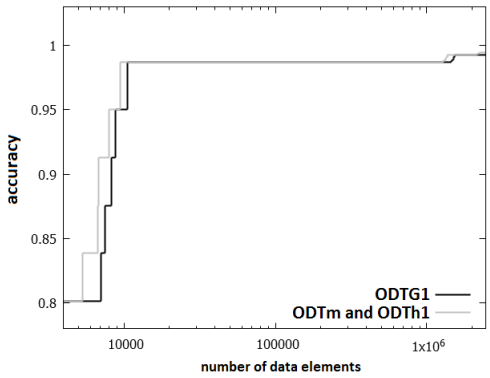

Fig. 10. Dependence between the average accuracy and the number of training data elements obtained for decision trees ODTG1, ODTm and ODTh1 with $\delta = 0.001$ for the 'KDD CUP 99' dataset.

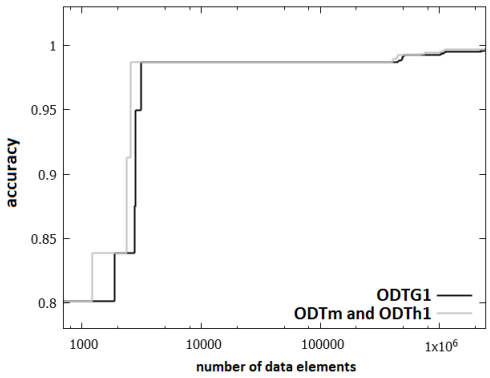

Fig. 11. Dependence between the average accuracy and the number of training data elements obtained for decision trees ODTG1, ODTm and ODTh1 with $\delta = 0.1$ for the 'KDD CUP 99' dataset.

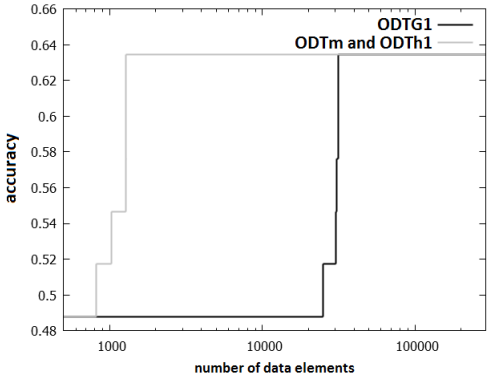

Fig. 12. Dependence between the average accuracy and the number of training data elements obtained for decision trees ODTG1, ODTm and ODTh1 with $\delta = 0.001$ for the 'Covertype' dataset.

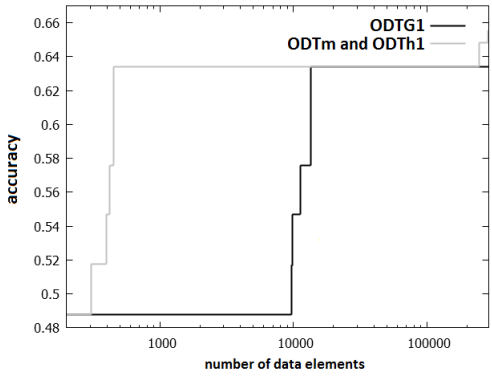

Fig. 13. Dependence between the average accuracy and the number of training data elements obtained for decision trees ODTG1, ODTm and ODTh1 with $\delta = 0.1$ for the 'Covertype' dataset.

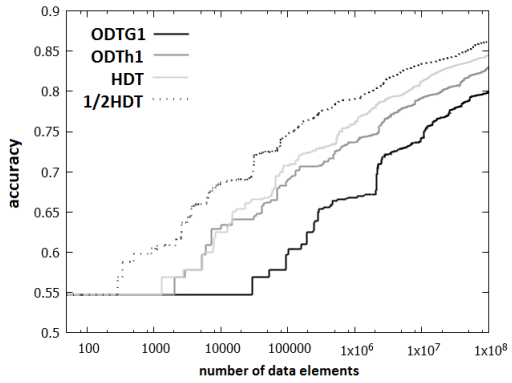

Fig. 14. Dependence between the average accuracy and the number of training data elements for the ODTG1, the ODTh1, the HDT and the 1/2HDT for $\delta = 0.001$.

in this paper. Since the ODTG1 outperforms the ODTG2 and the ODTh1 outperforms the ODTh2, only the ODTG1 and the ODTh1 will be taken into account in the following comparison. Additionally, the 1/2HDT is analyzed. The average accuracy results in the function of training dataset size obtained for $\delta = 0.001$ are presented in Fig. 14. The ODTh1 is only slightly less accurate than the popular HDT. The highest average accuracy is provided by the 1/2HDT. The reason is that the value of bound in splitting criterion for the 1/2HDT is twice lower than the bound corresponding to the HDT. This means that the 1/2HDT requires four times less data elements in order to make a decision about the split than it is needed in the case of the HDT. The 1/2HDT grows faster and, as it was previously stated, it is correlated with higher accuracy.

V. Conclusions

The main aim of this paper was to provide a better understanding of the process of decision trees induction in data stream scenario, particularly focusing on the mathematical foundations of splitting criteria in decision tree nodes. The two types of splitting criteria were distinguished. Type-I splitting criteria ensure that the splitting attribute chosen based on it provides, with high probability $(1-\delta)^{D-1}$, the highest expected value of the split measure. Type-II splitting criteria

ensure that the chosen attribute is the same as it would be if the whole data stream was available. In this paper new single splitting criteria were proposed, i.e. the type-I splitting criteria based on the misclassification error and the Gini index and the type-II splitting criterion based on the Gini index. Additionally, two hybrid splitting criteria were proposed. These criteria merge the type-I criterion based on the misclassification error with one of the criteria based on the Gini index. Finally, the type-$(I+I)$ and the type-$(II+I)$ hybrid criteria were obtained. The Online Decision Trees (ODT) with various splitting criteria were compared in the experimental simulations. The hybrid splitting criteria ensure higher classification accuracies than their single counterparts. Additionally, the proposed decision trees were compared with the Hoeffding Decision Tree as well as with decision tree with splitting criterion for which the bound is equal to the half of bound used in the Hoeffding tree. The last one provided the highest accuracy among all decision trees considered in this paper. We encourage researchers dealing with stream data mining to perform simulations with other values of constant C in formula (5).

Acknowledgments

The authors would like to thank the reviewers for helpful comments.

REFERENCES

[1] C. Aggarwal, *Data Streams: Models and Algorithms*. New York: Springer, LLC, 2007.

[2] A. Bifet, *Adaptive stream mining : pattern learning and mining from evolving data streams*, ser. Frontiers in artificial intelligence and applications. Amsterdam, Berlin: IOS Press, 2010.

[3] J. Gama, "A survey on learning from data strems: current and future trends," *Progress in Artificial Intelligence*, vol. 1, no. 1, pp. 45–55, April 2012.

[4] R. Polikar, L. Udpa, S. Udpa, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Transactions on System, Man and Cybernetics (C), Special Issue on Knowledge Management*, vol. 31, no. 4, pp. 497–508, 2001.

[5] H. He, S. Chen, K. Li, and X. Xu, "Incremental learning from stream data," *IEEE Transactions on Neural Networks*, pp. 1901–1914, 2011.

[6] K. Dyer, R. Capo, and R. Polikar, "Compose: A semisupervised learning framework for initially labeled nonstationary streaming data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 12–26, Jan 2014.

[7] M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," *Sigmod Record*, vol. 34, no. 2, pp. 18–26, June 2005.

[8] A. Tsymbal, "The problem of concept drift: definitions and related work," Computer Science Department, Trinity College Dublin, Ireland, Tech. Rep. TCD-CS-2004-15, April 2004.

[9] D. Brzezinski and J. Stefanowski, "Reacting to different types of concept drift: The accuracy updated ensemble algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 81–94, Jan 2014.

[10] J. Gomes, M. Gaber, P. Sousa, and E. Menasalvas, "Mining recurring concepts in a dynamic feature space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 95–110, Jan 2014.

[11] I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 27–39, Jan 2014.

[12] R. Bose, W. van der Aalst, I. Zliobaite, and M. Pechenizkiy, "Dealing with concept drifts in process mining," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 154–171, Jan 2014.

[13] A. Martin and P. Bartlett, *Neural network learning: Theoretical Foundations*. Cambridge University Press, 2009.

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[15] J. Quinlan, *Learning efficient classification procedures and their application to chess end games*. San Francisco, CA: Morgan Kaufmann, 1983, pp. 463–482.

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: Chapman and Hall, 1993.

[17] J. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann, 1993.

[18] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proc. 6th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 2000, pp. 71–80.

[19] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, "Decision trees for mining data streams based on the McDiarmid's bound," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1272–1279, 2013.

[20] P. Matuszyk, G. Krempl, and M. Spiliopoulou, "Correcting the usage of the hoeffding inequality in stream mining," in *Advances in Intelligent Data Analysis XII*, ser. Lecture Notes in Computer Science, A. Tucker, F. Höppner, A. Siebes, and S. Swift, Eds. Springer Berlin Heidelberg, 2013, vol. 8207, pp. 298–309.

[21] R. D. Rosa and N. Cesa-Bianchi, "Splitting with confidence in decision trees with application to stream mining," in *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015*, 2015, pp. 1–8.

[22] R. D. Rosa, "Confidence decision trees via online and active learning for streaming (big) data," *arXiv:1604.03278*, 2016.

[23] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "Decision trees for mining data streams based on the Gaussian approximation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 108–119, 2014.

[24] ——, "The CART decision tree for mining data streams," *Information Sciences*, vol. 266, pp. 1–15, 2014.

[25] ——, "A new method for data stream mining based on the misclassification error," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1048–1059, 2015.

[26] P. Duda, M. Jaworski, L. Pietruczuk, and L. Rutkowski, "A novel application of Hoeffding's inequality to decision trees construction for data streams," in *2014 International Joint Conference on Neural Networks (IJCNN)*, July 2014, pp. 3324–3330.

[27] A. Bifet and R. Kirkby, "Data stream mining a practical approach," University of WAIKATO, Tech. Rep., 2009.

[28] J. Gama, R. Fernandes, and R. Rocha, "Decision trees for mining data streams," *Intelligent Data Analysis*, vol. 10, no. 1, pp. 23–45, March 2006.

[29] B. Pfahringer, G. Holmes, and R. Kirkby, "New options for Hoeffding trees," in *AI 2007*, ser. LNAI, M.A.Orgun and J. Thornton, Eds., no. 4830. Springer, 2007, pp. 90–99.

[30] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proc. 7th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 2001, pp. 97–106.

[31] R. Kirkby, "Improving hoeffding trees," Ph.D. dissertation, University of Waikato, 2007.

[32] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

[33] C. McDiarmid, *On the method of bounded differences*, ser. Surveys in Combinatorics. London: London Mathematical Society Lecture Note Series 141, Cambridge University Press, 1989, no. 141, ch. 0, pp. 148–188.

[34] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.