

# Adaptive Parameter-free Learning from Evolving Data Streams

Albert Bifet and Ricard Gavaldà

Universitat Politècnica de Catalunya, Barcelona, Spain  
{abifet, gavalda}@lsi.upc.edu

**Abstract.** We propose and illustrate a method for developing algorithms that can adaptively learn from data streams that change over time. As an example, we take Hoeffding Tree, an incremental decision tree inducer for data streams, and use as a basis it to build two new methods that can deal with distribution and concept drift: a sliding window-based algorithm, Hoeffding Window Tree, and an adaptive method, Hoeffding Adaptive Tree. Our methods are based on using change detectors and estimator modules at the right places; we choose implementations with theoretical guarantees in order to extend such guarantees to the resulting adaptive learning algorithm. A main advantage of our methods is that they require no guess about how fast or how often the stream will change; other methods typically have several user-defined parameters to this effect. In our experiments, the new methods never do worse, and in some cases do much better, than CVFDT, a well-known method for tree induction on data streams with drift.

## 1 Introduction

Data streams pose several challenges on data mining algorithm design. Limited use of resources (time and memory) is one. The necessity of dealing with data whose nature or distribution changes over time is another fundamental one. Dealing with time-changing data requires in turn strategies for detecting and quantifying change, forgetting stale examples, and for model revision. Fairly generic strategies exist for detecting change and deciding when examples are no longer relevant. Model revision strategies, on the other hand, are in most cases method-specific.

Most strategies for dealing with time change contain hardwired constants, or else require input parameters, concerning the expected speed or frequency of the change; some examples are *a priori* definitions of sliding window lengths, values of decay or forgetting parameters, explicit bounds on maximum drift, etc. These choices represent preconceptions on how fast or how often the data are going to evolve and, of course, they may be completely wrong. Even more, no fixed choice may be right, since the stream may experience any combination of abrupt changes, gradual ones, and long stationary periods. More in general, an approach based on fixed parameters will be caught in the following tradeoff: the user would like to use large parameters to have more accurate statistics (hence, more precision) during periods of stability, but at the same time use small parameters to be able to quickly react to changes, when they occur.

Many ad-hoc methods have been used to deal with drift, often tied to particular algorithms. In this paper, we propose a more general approach based on using two primitive design elements: change detectors and estimators. The idea is to encapsulate

all the statistical calculations having to do with detecting change and keeping updated statistics from a stream an abstract data type that can then be used to replace, in a black-box way, the counters and accumulators that typically all machine learning and data mining algorithms use to make their decisions, including when change has occurred.

We believe that, compared to any previous approaches, our approach better isolates different concerns when designing new data mining algorithms, therefore reducing design time, increasing modularity, and facilitating analysis. Furthermore, since we crisply identify the nuclear problem in dealing with drift, and use a well-optimized algorithmic solution to tackle it, the resulting algorithms more accurate, adaptive, and time- and memory-efficient than other ad-hoc approaches. We have given evidence for this superiority in [3, 2, 4] and we demonstrate this idea again here.

We apply this idea to give two decision tree learning algorithms that can cope with concept and distribution drift on data streams: Hoeffding Window Trees in Section 4 and Hoeffding Adaptive Trees in Section 5. Decision trees are among the most common and well-studied classifier models. Classical methods such as C4.5 are not apt for data streams, as they assume all training data are available simultaneously in main memory, allowing for an unbounded number of passes, and certainly do not deal with data that changes over time. In the data stream context, a reference work on learning decision trees is the Hoeffding Tree or Very Fast Decision Tree method (VFDT) for fast, incremental learning [8]. The methods we propose are based on VFDT, enriched with the change detection and estimation building blocks mentioned above.

We try several such building blocks, although the best suited for our purposes is the `ADWIN` algorithm [3], described in Section 4.1. This algorithm is parameter-free in that it automatically and continuously detects the rate of change in the data streams rather than using apriori guesses, thus allowing the client algorithm to react adaptively to the data stream it is processing. Additionally, `ADWIN` has rigorous guarantees of performance (a theorem). We show that these guarantees can be transferred to decision tree learners as follows: if a change is followed by a long enough stable period, the classification error of the learner will tend, and the same rate, to the error rate of VFDT.

We test on Section 6 our methods with synthetic datasets, using the SEA concepts, introduced in [13], and two sets from the UCI repository, Adult and Poker-Hand. We compare our methods among themselves but also with CVFDT, another concept-adapting variant of VFDT proposed by Domingos, Spencer, and Hulten [11]. A one-line conclusion of our experiments would be that, because of its self-adapting property, we can present datasets where our algorithm performs much better than CVFDT and we never do much worse. Some comparison of time and memory usage of our methods and CVFDT is included.

A longer version of this paper [5] with additional theoretical results and experiments, is available from the first author webpage.

## 2 A Methodology for Adaptive Stream Mining

The starting point of our work is the following observation: In the data stream mining literature, most algorithms incorporate one or more of the following ingredients: windows to remember recent examples; methods for detecting distribution change in the

input; and methods for keeping updated estimations for some statistics of the input. We see them as the basis for solving the three central problems of

- what to remember or forget,
- when to do the model upgrade, and
- how to do the model upgrade.

Our claim is that by basing mining algorithms on well-designed, well-encapsulated modules for these tasks, one can often get more generic and more efficient solutions than by using ad-hoc techniques as required. Similarly, we will argue that our methods for inducing decision trees are simpler to describe, adapt better to the data, perform better or much better, and use less memory than the ad-hoc designed CVFDT algorithm, even though they are all derived from the same VFDT mining algorithm.

A similar approach was taken, for example, in [4] to give simple adaptive closed-tree mining algorithms. Using a general methodology to identify closed patterns based in Galois Lattice Theory, three closed tree mining algorithms were developed: an incremental one INCTREENAT, a sliding-window based one, WINTREENAT, and finally one that mines closed trees adaptively from data streams, ADATREENAT.

### 3 Incremental Decision Trees: Hoeffding Trees

Classical decision tree learners such as ID3, C4.5 [12], and CART [6] assume that all training examples can be stored simultaneously in main memory, and are thus severely limited in the number of examples they can learn from. In particular, they are not applicable to data streams, where potentially there is no bound on number of examples and these arrive sequentially.

Domingos and Hulten [8] developed Hoeffding trees, an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time.

Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quantifies the number of observations (in our case, examples) needed to estimate some statistics within a prescribed precision (in our case, the goodness of an attribute). VFDT (Very Fast Decision Trees) is the implementation of Hoeffding trees, with a few heuristics added, described in [8]; we basically identify both in this paper.

### 4 Decision Trees on Sliding Windows

We propose a general method for building incrementally a decision tree based on a keeping sliding window of the last instances on the stream. To specify one such method, we specify how to:

- place one or more change detectors at every node that will raise a hand whenever something worth attention happens at the node
- create, manage, switch and delete alternate trees

- maintain estimators of only relevant statistics at the nodes of the current sliding window

We call *Hoeffding Window Tree* any decision tree that uses Hoeffding bounds, maintains a sliding window of instances, and that can be included in this general framework. Figure 1 shows the pseudo-code of Hoeffding Window Tree.

```

HOEFFDING WINDOW TREE(Stream,  $\delta$ )
1  ▷ Let HT be a tree with a single leaf(root)
2  ▷ Init estimators  $A_{ijk}$  at root
3  for each example  $(x, y)$  in Stream
4    do HWTREEGROW( $(x, y)$ , HT,  $\delta$ )

HWTREEGROW( $(x, y)$ , HT,  $\delta$ )
1  ▷ Sort  $(x, y)$  to leaf  $l$  using HT
2  ▷ Update estimators  $A_{ijk}$ 
   at leaf  $l$  and nodes traversed in the sort
3  if this node has an alternate tree  $T_{alt}$ 
4    HWTREEGROW( $(x, y)$ ,  $T_{alt}$ ,  $\delta$ )
5  ▷ Compute  $G$  for each attribute
6  if  $G(\text{Best Attr.}) - G(\text{2nd best}) > \epsilon(\delta', \dots)$ a
7    then
8      ▷ Split leaf on best attribute
9  for each branch
10     do ▷ Start new leaf
11         and initialize estimators
12     if one change detector has detected change
13       then
14         ▷ Create an alternate subtree if there is none
15     if existing alternate tree is more accurate
16       then
17         ▷ replace current node with alternate tree

```

<sup>a</sup> Here  $\delta'$  should be the *Bonferroni correction* of  $\delta$  to account for the fact that many tests are performed and we want all of them to be simultaneously correct with probability  $1 - \delta$ . It is enough e.g. to divide  $\delta$  by the number of tests performed so far. The need for this correction is also acknowledged in [8], although in experiments the more convenient option of using a lower  $\delta$  was taken. We have followed the same option in our experiments for fair comparison.

**Fig. 1.** Hoeffding Window Tree algorithm

#### 4.1 HWT-ADWIN : Hoeffding Window Tree using ADWIN

Recently, we proposed an algorithm termed ADWIN [3] (for Adaptive Windowing) that is an estimator with memory and change detector. We use it to design HWT-ADWIN, a new Hoeffding Window Tree that uses ADWIN as a change detector. The main advantage of using a change detector as ADWIN is that as it has theoretical guarantees we can extend this guarantees to the learning algorithms. ADWIN keeps a variable-length window of recently seen items, with the property that the window has the maximal length statistically consistent with the hypothesis “there has been no change in the average value inside the window”.

ADWIN is parameter- and assumption-free in the sense that it automatically detects and adapts to the current rate of change. Its only parameter is a confidence bound  $\delta$ , indicating how confident we want to be in the algorithm’s output, inherent to all algorithms dealing with random processes.

Also important for our purposes, ADWIN does not maintain the window explicitly, but compresses it using a variant of the exponential histogram technique in [7]. This means that it keeps a window of length  $W$  using only  $O(\log W)$  memory and  $O(\log W)$  processing time per item, rather than the  $O(W)$  one expects from a naïve implementation.

#### 4.2 CVFDT

As an extension of VFDT to deal with concept change Hulten, Spencer, and Domingos presented Concept-adapting Very Fast Decision Trees CVFDT [11] algorithm. We review it here briefly and compare it to our method.

CVFDT works by keeping its model consistent with respect to a sliding window of data from the data stream, and creating and replacing alternate decision subtrees when it detects that the distribution of data is changing at a node. When new data arrives, CVFDT updates the sufficient statistics at its nodes by incrementing the counts  $n_{ijk}$  corresponding to the new examples and decrementing the counts  $n_{ijk}$  corresponding to the oldest example in the window, which is effectively forgotten. CVFDT is a Hoeffding Window Tree as it is included in the general method previously presented.

Two external differences among CVFDT and our method is that CVFDT has no theoretical guarantees (as far as we know), and that it uses a number of parameters, with default values that can be changed by the user - but which are fixed for a given execution. Besides the example window length, it needs:

1.  $T_0$ : after each  $T_0$  examples, CVFDT traverses all the decision tree, and checks at each node if the splitting attribute is still the best. If there is a better splitting attribute, it starts growing an alternate tree rooted at this node, and it splits on the currently best attribute according to the statistics in the node.
2.  $T_1$ : after an alternate tree is created, the following  $T_1$  examples are used to build the alternate tree.
3.  $T_2$ : after the arrival of  $T_1$  examples, the following  $T_2$  examples are used to test the accuracy of the alternate tree. If the alternate tree is more accurate than the current one, CVFDT replaces it with this alternate tree (we say that the alternate tree is promoted).

The default values are  $T_0 = 10,000$ ,  $T_1 = 9,000$ , and  $T_2 = 1,000$ . One can interpret these figures as the preconception that often about the last 50,000 examples are likely to be relevant, and that change is not likely to occur faster than every 10,000 examples. These preconceptions may or may not be right for a given data source.

The main internal differences of HWT-ADWIN respect CVFDT are:

- The alternates trees are created as soon as change is detected, without having to wait that a fixed number of examples arrives after the change. Furthermore, the more abrupt the change is, the faster a new alternate tree will be created.
- HWT-ADWIN replaces the old trees by the new alternates trees as soon as there is evidence that they are more accurate, rather than having to wait for another fixed number of examples.

These two effects can be summarized saying that HWT-ADWIN adapts to the scale of time change in the data, rather than having to rely on the *a priori* guesses by the user.

## 5 Hoeffding Adaptive Trees

In this section we present Hoeffding Adaptive Tree as a new method that evolving from Hoeffding Window Tree, adaptively learn from data streams that change over time without needing a fixed size of sliding window. The optimal size of the sliding window is a very difficult parameter to guess for users, since it depends on the rate of change of the distribution of the dataset.

In order to avoid to choose a size parameter, we propose a new method for managing statistics at the nodes. The general idea is simple: we place instances of estimators of frequency statistics at every node, that is, replacing each  $n_{ijk}$  counters in the Hoeffding Window Tree with an instance  $A_{ijk}$  of an estimator.

More precisely, we present three variants of a *Hoeffding Adaptive Tree* or HAT, depending on the estimator used:

- HAT-INC: it uses a linear incremental estimator
- HAT-EWMA: it uses an Exponential Weight Moving Average (EWMA)
- HAT-ADWIN : it uses an ADWIN estimator. As the ADWIN instances are also change detectors, they will give an alarm when a change in the attribute-class statistics at that node is detected, which indicates also a possible concept change.

The main advantages of this new method over a Hoeffding Window Tree are:

- All relevant statistics from the examples are kept in the nodes. There is no need of an optimal size of sliding window for all nodes. Each node can decide which of the last instances are currently relevant for it. There is no need for an additional window to store current examples. For medium window sizes, this factor substantially reduces our memory consumption with respect to a Hoeffding Window Tree.
- A Hoeffding Window Tree, as CVFDT for example, stores only a bounded part of the window in main memory. The rest (most of it, for large window sizes) is stored in disk. For example, CVFDT has one parameter that indicates the amount of main memory used to store the window (default is 10,000). Hoeffding Adaptive Trees keeps all its data in main memory.

## 5.1 Example of performance Guarantee

In this subsection we show a performance guarantee on the error rate of HAT-ADWIN on a simple situation. Roughly speaking, it states that after a distribution and concept change in the data stream, followed by a stable period, HAT-ADWIN will start, in reasonable time, growing a tree identical to the one that VFDT would grow if starting afresh from the new stable distribution. Statements for more complex scenarios are possible, including some with slow, gradual, changes, but require more space than available here.

**Theorem 1.** *Let  $D_0$  and  $D_1$  be two distributions on labelled examples. Let  $S$  be a data stream that contains examples following  $D_0$  for a time  $T$ , then suddenly changes to using  $D_1$ . Let  $t$  be the time that until VFDT running on a (stable) stream with distribution  $D_1$  takes to perform a split at the node. Assume also that VFDT on  $D_0$  and  $D_1$  builds trees that differ on the attribute tested at the root. Then with probability at least  $1 - \delta$ :*

- *By time  $t' = T + c \cdot V^2 \cdot t \log(tV)$ , HAT-ADWIN will create at the root an alternate tree labelled with the same attribute as VFDT( $D_1$ ). Here  $c \leq 20$  is an absolute constant, and  $V$  the number of values of the attributes.<sup>1</sup>*
- *this alternate tree will evolve from then on identically as does that of VFDT( $D_1$ ), and will eventually be promoted to be the current tree if and only if its error on  $D_1$  is smaller than that of the tree built by time  $T$ .*

If the two trees do not differ at the roots, the corresponding statement can be made for a pair of deeper nodes.

**Lemma 1.** *In the situation above, at every time  $t + T > T$ , with probability  $1 - \delta$  we have at every node and for every counter (instance of ADWIN)  $A_{i,j,k}$*

$$|A_{i,j,k} - P_{i,j,k}| \leq \sqrt{\frac{\ln(1/\delta') T}{t(t+T)}}$$

where  $P_{i,j,k}$  is the probability that an example arriving at the node has value  $j$  in its  $i$ th attribute and class  $k$ .

Observe that for fixed  $\delta'$  and  $T$  this bound tends to 0 as  $t$  grows.

To prove the theorem, use this lemma to prove high-confidence bounds on the estimation of  $G(a)$  for all attributes at the root, and show that the attribute *best* chosen by VFDT on  $D_1$  will also have maximal  $G(\text{best})$  at some point, so it will be placed at the root of an alternate tree. Since this new alternate tree will be grown exclusively with fresh examples from  $D_1$ , it will evolve as a tree grown by VFDT on  $D_1$ .

<sup>1</sup> This value of  $t'$  is a very large overestimate, as indicated by our experiments. We are working on an improved analysis, and hope to be able to reduce  $t'$  to  $T + c \cdot t$ , for  $c < 4$ .

**Table 1.** SEA on-line errors using discrete attributes with 10% noise

	CHANGE SPEED		
	1,000	10,000	100,000
HAT-INC	16.99%	16.08%	14.82%
HAT-EWMA	16.98%	15.83%	<b>14.64%</b>
HAT-ADWIN	16.86%	<b>15.39%</b>	14.73%
HAT-INC NB	16.88%	15.93%	14.86%
HAT-EWMA NB	<b>16.85%</b>	15.91%	14.73%
HAT-ADWIN NB	16.90%	15.76%	14.75%
CVFDT $ W  = 1,000$	19.47%	<i>15.71%</i>	15.81%
CVFDT $ W  = 10,000$	17.03%	17.12%	<i>14.80%</i>
CVFDT $ W  = 100,000$	<i>16.97%</i>	17.15%	17.09%

## 6 Experimental evaluation

We tested Hoeffding Adaptive Trees using synthetic and real datasets. In the experiments with synthetic datasets, we use the SEA Concepts [13]. In the experiments with real datasets we use two UCI datasets [1] Adult and Poker-Hand from the UCI repository of machine learning databases. In all experiments, we use the values  $\delta = 10^{-4}$ ,  $T_0 = 20,000$ ,  $T_1 = 9,000$ , and  $T_2 = 1,000$ , following the original CVFDT experiments [11].

In all tables, the result for the best classifier for a given experiment is marked in **boldface**, and the best choice for CVFDT window length is shown in *italics*.

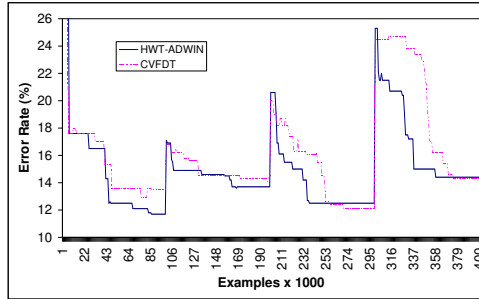
We included an improvement over CVFDT (which could be made on the original CVFDT as well). If the two best attributes at a node happen to have exactly the same gain, the tie may be never resolved and split does not occur. In our experiments this was often the case, so we added an additional split rule: when  $G(\text{best})$  exceeds by three times the current value of  $\epsilon(\delta, \dots)$ , a split is forced anyway.

We have tested the three versions of Hoeffding Adaptive Tree, HAT-INC, HAT-EWMA ( $\alpha = .01$ ), HAT-ADWIN, each with and without the addition of Naïve Bayes (NB) classifiers at the leaves. As a general comment on the results, the use of NB classifiers does not always improve the results, although it does make a good difference in some cases; this was observed in [9], where a more detailed analysis can be found.

First, we experiment using the SEA concepts, a dataset with abrupt concept drift, first introduced in [13]. This artificial dataset is generated using three attributes, where only the two first attributes are relevant. All three attributes have values between 0 and 10. We generate 400,000 random samples. We divide all the points in blocks with different concepts. In each block, we classify using  $f_1 + f_2 \leq \theta$ , where  $f_1$  and  $f_2$  represent the first two attributes and  $\theta$  is a threshold value. We use threshold values 9, 8, 7 and 9.5 for the data blocks. We inserted about 10% class noise into each block of data.

We test our methods using discrete and continuous attributes. The on-line errors results for discrete attributes are shown in Table 1. On-line errors are the errors measured





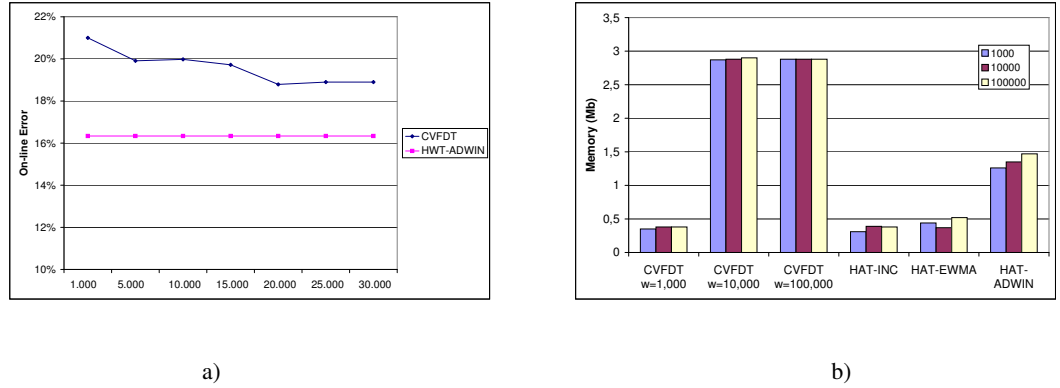
**Fig. 2.** Learning curve of SEA Concepts using continuous attributes

each time an example arrives with the current decision tree, before updating the statistics. Each column reflects a different speed of concept change. We observe that CVFDT best performance is not always with the same example window size, and that there is no optimal window size. The different versions of Hoeffding Adaptive Trees have a very similar performance, essentially identical to that of CVFDT with optimal window size for that speed of change. More graphically, Figure 2 shows its learning curve using continuous attributes for a speed of change of 100,000. Note that at the points where the concept drift appears HWT-ADWIN, decreases its error faster than CVFDT, due to the fact that it detects change faster.

We test Hoeffding Adaptive Trees on two real datasets in two different ways: with and without concept drift. We tried some of the largest UCI datasets [1], and report results on Adult and Poker-Hand. For the Covertype and Census-Income datasets, the results we obtained with our method were essentially the same as for CVFDT (ours did better by fractions of 1% only) – we do not claim that our method is always better than CVFDT, but this confirms our belief that it is never much worse.

An important problem with most of the real-world benchmark data sets is that there is little concept drift in them [14] or the amount of drift is unknown, so in many research works, concept drift is introduced artificially. We simulate concept drift by ordering the datasets by one of its attributes, the *education* attribute for Adult, and the first (unnamed) attribute for Poker-Hand. Note again that while using CVFDT one faces the question of which parameter values to use, our method just needs to be told “go” and will find the right values online.

The Adult dataset aims to predict whether a person makes over 50k a year, and it was created based on census data. Adult consists of 48,842 instances, 14 attributes (6 continuous and 8 nominal) and missing attribute values. The Poker-Hand dataset consists of 1,025,010 instances and 11 attributes. Each record of the Poker-Hand dataset is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes (suit and rank), for a total of 10 predictive



**Fig. 3.** a) On-line error on UCI Adult dataset, ordered by the *education* attribute and b) Memory used on SEA Concepts experiments.

**Table 2.** On-line classification errors for CVFDT and Hoeffding Adaptive Trees on Poker-Hand data set.

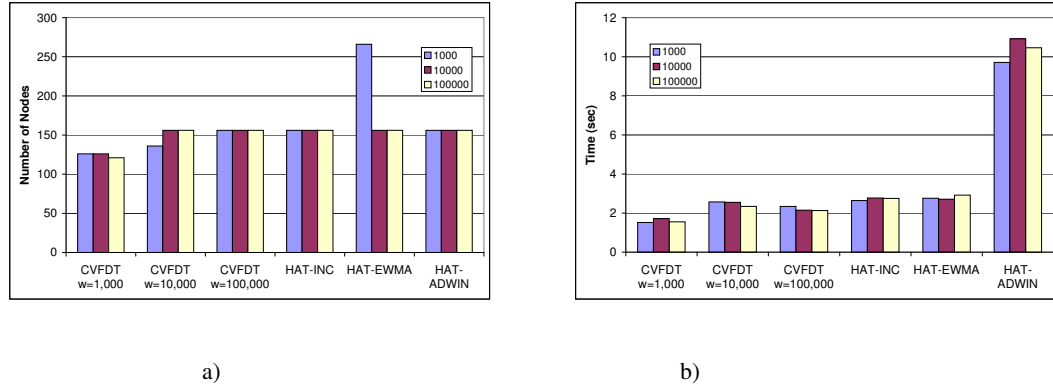
	NO DRIFT	ARTIFICIAL DRIFT
HAT-INC	38.32%	39.21%
HAT-EWMA	39.48%	40.26%
HAT-ADWIN	38.71%	41.85%
HAT-INC NB	41.77%	42.83%
HAT-EWMA NB	24.49%	<b>27.28%</b>
HAT-ADWIN NB	<b>16.91%</b>	33.53%
CVFDT $ W  = 1,000$	49.90%	49.94%
CVFDT $ W  = 10,000$	49.88%	49.88%
CVFDT $ W  = 100,000$	49.89%	52.13%

attributes. There is one Class attribute that describes the "Poker Hand". The order of cards is important, which is why there are 480 possible Royal Flush hands instead of 4.

Table 2 shows the results on Poker-Hand dataset. It can be seen that CVFDT remains at 50% error, while the different variants of Hoeffding Adaptive Trees are mostly below 40% and one reaches 17% error only. In Figure 3 we compare HWT-ADWIN error rate to CVFDT using different window sizes. We observe that CVFDT on-line error decreases when the example window size increases, and that HWT-ADWIN on-line error is lower for all window sizes.

## 7 Time and memory

In this section, we discuss briefly the time and memory performance of Hoeffding Adaptive Trees. All programs were implemented in C modifying and expanding the



**Fig. 4.** a) Number of Nodes used on SEA Concepts experiments and b) Time on SEA Concepts experiments

version of CVFDT available from the VFML [10] software web page. We have slightly modified the CVFDT implementation to follow strictly the CVFDT algorithm explained in the original paper by Hulthen, Spencer and Domingos [11]. The experiments were performed on a 2.0 GHz Intel Core Duo PC machine with 2 Gigabyte main memory, running Ubuntu 8.04.

Consider the experiments on SEA Concepts, with different speed of changes: 1, 000, 10, 000 and 100, 000. Figure 4 shows the memory used on these experiments. HAT-INC and HAT-EWMA, are the methods that use less memory. The reason for this fact is that they don't keep examples in memory as CVFDT, and that they don't store ADWIN data for all attributes, attribute values and classes, as HAT-ADWIN. We have used the default 10,000 for the amount of window examples kept in memory, so the memory used by CVFDT is essentially the same for  $W = 10,000$  and  $W = 100,000$ , and about 10 times larger than the memory used by HAT-INC memory.

Figure 4 shows the number of nodes used in the experiments of SEA Concepts. We see that the number of nodes is similar for all methods, confirming that the good results on memory of HAT-INC is not due to smaller size of trees.

Finally, with respect to time we see that CVFDT is still the fastest method, but HAT-INC and HAT-EWMA have a very similar performance to CVFDT, a remarkable fact given that they are monitoring all the change that may occur in any node of the main tree and all the alternate trees. HAT-ADWIN increases time by a factor of 4, so it is still usable if time or data speed is not the main concern.

## 8 Conclusions and Future Work

We have presented a general adaptive methodology for mining data streams with concept drift, and two decision tree algorithms. We have proposed three variants of Hoeffding Adaptive Tree algorithm, a decision tree miner for data streams that adapts

to concept drift without using a fixed sized window. Contrary to CVFDT, they have theoretical guarantees of performance, relative to those of VFDT.

In our experiments, Hoeffding Adaptive Trees are always as accurate as CVFDT and, in some cases, they have substantially lower error. Their running time is similar in HAT-EWMA and HAT-INC and only slightly higher in HAT-ADWIN, and their memory consumption is remarkably smaller, often by an order of magnitude.

We can conclude that HAT-ADWIN is the most powerful method, but HAT-EWMA is a faster method that gives approximate results similar to HAT-ADWIN. An obvious future work is experimenting with the exponential smoothing factor  $\alpha$  of EWMA methods used in HAT-EWMA.

We would like to extend our methodology to ensemble methods such as boosting, bagging, and Hoeffding Option Trees.

## References

- [1] D. N. A. Asuncion. UCI machine learning repository, 2007.
- [2] A. Bifet and R. Gavaldà. Kalman filters and adaptive windows for learning in data streams. In *Discovery Science*, pages 29–40, 2006.
- [3] A. Bifet and R. Gavaldà. Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining*, 2007.
- [4] A. Bifet and R. Gavaldà. Mining adaptively frequent closed unlabeled rooted trees in data streams. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.
- [5] A. Bifet and R. Gavaldà. Adaptive parameter-free learning from evolving data streams. Technical report, LSI-09-9-R, Universitat Politècnica de Catalunya, Barcelona, Spain, 2009.
- [6] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1994.
- [7] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining stream statistics over sliding windows. *SIAM Journal on Computing*, 14(1):27–45, 2002.
- [8] P. Domingos and G. Hulten. Mining high-speed data streams. In *Knowledge Discovery and Data Mining*, pages 71–80, 2000.
- [9] G. Holmes, R. Kirkby, and B. Pfahringer. Stress-testing hoeffding trees. In *PKDD*, pages 495–502, 2005.
- [10] G. Hulten and P. Domingos. VFML – a toolkit for mining high-speed time-changing data streams. <http://www.cs.washington.edu/dm/vfml/>. 2003.
- [11] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *7th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 97–106, San Francisco, CA, 2001. ACM Press.
- [12] R. J. Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, January 1993.
- [13] W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382, New York, NY, USA, 2001. ACM Press.
- [14] A. Tsymbal. The problem of concept drift: Definitions and related work. Technical Report TCD-CS-2004-15, Department of Computer Science, University of Dublin, Trinity College, 2004.