Decision tree-based Feature Ranking in Concept Drifting Data Streams

Jean Antonio Pereira Karax
Specialization in Big Data & Analytics
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Brazil
jeankarax@gmail.com

Andreia Malucelli
PPGIa – Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Brazil
malu@ppgia.pucpr.br

Jean Paul Barddal
PPGIa – Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Brazil
jean.barddal@ppgia.pucpr.br

ABSTRACT

Data stream mining targets the learning of predictive models that evolve over time according to changes in arriving data. Throughout the years, several approaches have been tailored to create and continuously update predictive models from these streams, and from these, Hoeffding Trees became a popular choice for learning decision trees from data streams. In this paper, we aim at quantifying and expressing the importance of features in dynamic scenarios is of the utmost importance as they allow domain experts to back up, or invalidate, a predictive model. Therefore, we propose and assess a positional gain method tailored for for both individual and ensembles of Hoeffding Trees and how these behave in both synthetic and real-world scenarios.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; **Online learning settings**; **Feature selection**;

KEYWORDS

Data Stream Mining; Feature Ranking; Concept Drift

ACM Reference format:
Jean Antonio Pereira Karax, Andreia Malucelli, and Jean Paul Barddal. 2019. Decision tree-based Feature Ranking in Concept Drifting Data Streams. In *Proceedings of The 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, April 8–12, 2019 (SAC '19),* 3 pages.
https://doi.org/10.1145/3297280.3297551

1 INTRODUCTION

Most of computational systems nowadays must process data that arrives sequentially over time, in the form of a potentially unbounded data stream. Yet, it is widely observed that the dimensionality of streams are also growing. High-dimensional data brings forward serious problems to learning algorithms, as data sparsity negatively impacts the **(i)** accuracy of the predictive model as the learners are prone to overfitting, **(ii)** complexity of the predictive model, and **(iii)** processing times of the model as a consequence of **(ii)**. Therefore, it becomes of importance to score and rank features to identify which are important from the entire feature set.

In this paper, we propose the use of decision trees to score and rank features during the processing of data streams. In contrast to existing proposals for ranking features from data streams in classification tasks, decision trees allow the identification of high-order interactions between features. Our proposal is to traverse the internal structure of decision trees to rank features according to their position in the internal structure and their goodness-of-fit at the moment of split. Furthermore, the feature ranks can be extracted from either single trees or ensembles.

2 PROBLEM STATEMENT

Let S denote a data stream providing instances in the $(\vec{x}^t, y^t)$ form, where $\vec{x}$ is a d-dimensional vector of values belonging to a feature set X that is possibly numerical, categorical, ordinal, or most likely mixed; and $y^t \in Y$ its corresponding class label. Throughout this work, we denote the i-th feature of a data stream as $X_i \in X$.

The goal in data stream classification is fit and update a predictive model $f : X \rightarrow Y$ over time. Formally, a concept C is given by a probability distribution $P[X, Y]$ [9]. Since the data distribution is ephemeral, we denote $P_t[X, Y]$ to be the concept at a timestamp t, while $P_{[t,u]}[X, Y]$ is the concept at a time period $[t, u]$. A concept drift occurs between instants t and u if $P_t[X, Y] \neq P_u[X, Y]$ holds, and similarly for time periods $[t, u]$ and $[v, w]$ if $P_{[t,u]}[X, Y] \neq P_{[v,w]}[X, Y]$. In this paper, we target one specific type of concept drift, called *feature drift* [1], which occurs when a subset of features becomes, or ceases to be, relevant to the learning task.

3 RELATED WORK

Focusing on the classification task for data streams, we highlight the Hoeffding Tree [5]. The definition of which attribute $X_i \in X$ is used in a Hoeffding Tree split node is given by comparing all available features and choosing the best according to an heuristic function J, e.g. gini index, conditional entropy, information gain, and symmetrical uncertainty. As traditional decision trees, the Hoeffding Tree continuously selects features for its branches according to the arrival of data. The traditional Hoeffding Tree assumes that the data generator process is stationary, and thus, it only grows over time. Consequently, authors in [3] proposed the Hoeffding Adaptive Tree, which uses a drift detector inside decision nodes to monitor the internal error rates of the tree, thus, continuously selecting features and building its predictive model over time.

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297551

In contrast to feature selection, feature scoring - the focus of this paper - associates to each feature a score that determines how 'important' each feature is w.r.t. class prediction. Recently, authors in [2] proposed a dynamic weighting scheme for the problem of classification over data streams. In practice, the weighting process relies on a flat evaluation of features according to a feature importance function $I(\cdot)$.

4 PROPOSAL

In this section we propose a scheme to quantify the importance of features according to their use in Hoeffding Trees. Quantifying and expressing the importance of features in dynamic scenarios is of the utmost importance since they allow domain experts to back up, or invalidate, a predictive model.

The rationale behind our proposal is to take into account the (i) position in which a feature is used in the tree internal structure, and (ii) the goodness-of-fit of the same feature computed during the time of the split. This proposal is similar to the Mean Decrease Impurity (MDI), in which the final feature importance $I(X_i), \forall X_i \in X$ is calculated as the sum of the heuristic values $J(\cdot)$ computed at the time of the split proportionally to the number of samples it splits, as this sum iterates over all trees in the case of an ensemble. The MDI is given by Equation 1, where t_i is an arbitrary tree inside the ensemble of trees $T = \{t_1, t_2, \ldots, t_E\}$, N_t is the number of instances observed in a split node b, N is the total number of samples observed in the entire tree, $J(b)$ is the heuristic value (goodness-of-fit) computed during the split process of b, and $\Omega(b)$ is the function that returns the feature $X_i \in X$ used in b.

$$I(X_i) = MDI(X_i) = \frac{1}{|T|}\sum_{t_i}^{T}\sum_{b}^{t_i}\begin{cases}\frac{N_t}{N} \times J(b), & \text{if } \Omega(b) = X_i \\ 0, & \text{otherwise}\end{cases} \tag{1}$$

The main drawback of using MDI in Hoeffding Trees resides in the $\frac{N_t}{N}$ component, since N_t is the grace period parameter and is approximately the same value across all split nodes, while N continuously grows as new data becomes available. Consequently, this component will equally penalize all features regardless of their position inside the tree structures. This behavior is different from the one observed in traditional decision trees, where the proportion $\frac{N_t}{N}$ is larger on the superior nodes and decreases as we traverse the tree structure and reach the leaves. We propose a slightly different approach to quantify the importance of a feature X_i inside a tree-based model. We replace the $\frac{N_t}{N}$ component by another ratio that takes into account the position of a node b inside the tree structure w.r.t. the entire tree depth. We name it the 'Mean Positional Gain' (MPG), as the importance of a feature X_i is given by Equation 2, where $\gamma(t_i)$ is the function that returns the number of split nodes in a tree $t_i \in T$, $h(t_i)$ is the tree height, $h(b, t_i)$ is the tree depth of a split node b in t_i, and R is the maximum value of the heuristic J.

$$\begin{aligned} I(X_i) &= MPG(X_i) = \\ &= \frac{1}{|T|}\sum_{t_i}^{T}\frac{1}{\gamma(t_i)}\sum_{b}^{t_i}\begin{cases}\frac{h(t_i)-h(b,t_i)}{h(t_i)} \times \frac{J(b)}{R}, & \\ \qquad \text{if } \Omega(b) = X_i & \\ 0, \quad \text{otherwise}\end{cases}\end{aligned} \tag{2}$$

In practice, the term $\frac{h(t_i)-h(b,t_i)}{h(t_i)}$ guarantees that split nodes closer to the leaves of a tree possess smaller importance rates compared to those observed closer to the tree root. Another small difference introduced in Equation 2 compared to Equation 1 is that the heuristic function value $J(b)$ is normalized in the $[0; 1]$ interval.

5 ANALYSIS

In this section, we apply the proposed feature scoring scheme on Hoeffding Trees [5], Hoeffding Adaptive Trees [3], and the Adaptive Random Forest [6] on both synthetic and real-world data. All experiments were conducted in the Massive Online Analysis (MOA) framework [4]. The synthetic experiments contain 500,000 instances and 200 features, with 2 gradual and equally spaced drifts [1], i.e. occur at 166,666 and 333,333 instances, each with a width of 10,000 instances. The Binary Generator (BG1) [7] is used to synthesize a stream where only 3 out of the 200 available features are relevant for predicting the class y; whereas in the SEA generator [8] as only 2 out of the 200 numeric features are relevant. In the BG1 experiment, features *attrib0*, *attrib1*, and *attrib2*; *attrib10*, *attrib20*, and *attrib50*; and *attrib10*, *attrib25*, and *attrib30* are relevant during each of the concepts we have in the stream. Similarly, features *attrib64* and *attrib113*; *attrib6* and *attrib87*; and *attrib29 attrib131* are relevant during each of the 3 concepts for the SEA experiment.

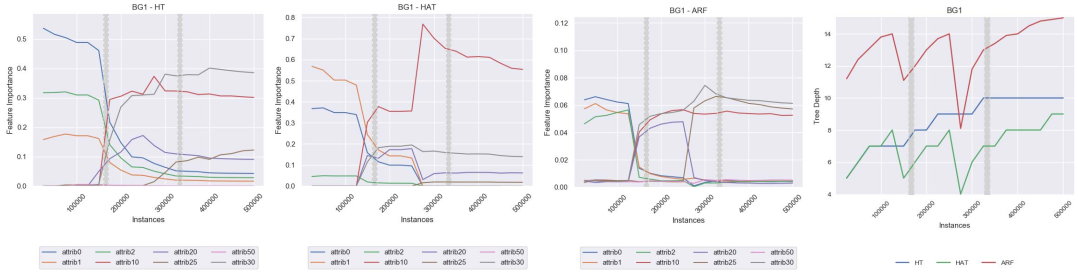
In Figure 1 we report the tree depth and feature importance scores obtained during the BG1 experiment. Focusing on the feature importances, we see in Figures 1 (b), (c), and (d) that the correct features are highlighted, with the exception of *attrib30.* This is explained by the fact that online decision trees are learned using samples of the data, and thus, if the sample that the tree uses to decide on which feature it should split on is skewed it may select an 'innapropriate' feature. Similarly, the results for the SEA experiment are given in Figure 2, where all learners are able to correctly highlighting the relevant features.

In this section, we report the feature importance scores obtained in the AIRLINES dataset which was obtained from the MOA dataset repository[1]. As we note on the Figure 3 (b), there are two features ("AirportFrom" and "AirportTo") that are highlighted. On Figure 3 (d), the same features are highlighted with more peaks of feature importance. On Figure 3 (c), only "AirportFrom" and "AirportTo" are highlighted. We note on Figure 3 (c) that "AirportTo" feature importance rate is higher when "AirportFrom" is lower. In practice, what happens here is that the Hoeffding Adaptive Tree is continuously signaling drifts, and thus, the tree branches are being reset and relearned (see Figure 3 (e) for tree depth).

6 CONCLUSION

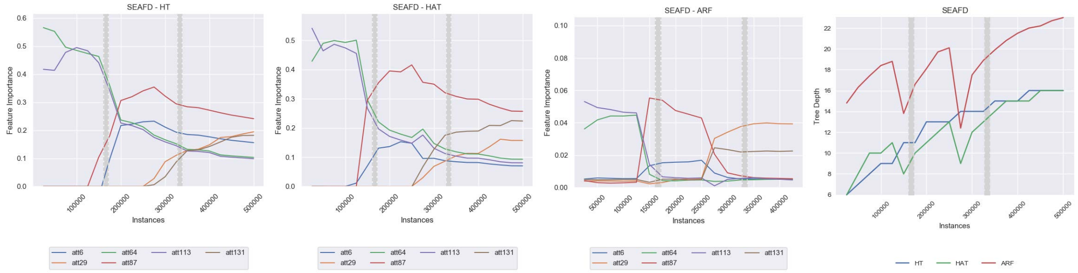
In this paper, we propose a feature ranking function to quantify the importance of features according to their usage in Hoeffding Tree-based classifiers. This scheme is able to depict high-order interactions between features, a gap observed in the existing works of the area [2]. As future works, we plan to use the scores obtained by our proposed schemes as the core of novel feature selection algorithms, and try to extend the work of [2], where the feature importance scores obtained here can be used inside other learning schemes, such as bayesian and instance-based learners.

[1] The MOA dataset repository can be found at https://moa.cms.waikato.ac.nz/datasets/



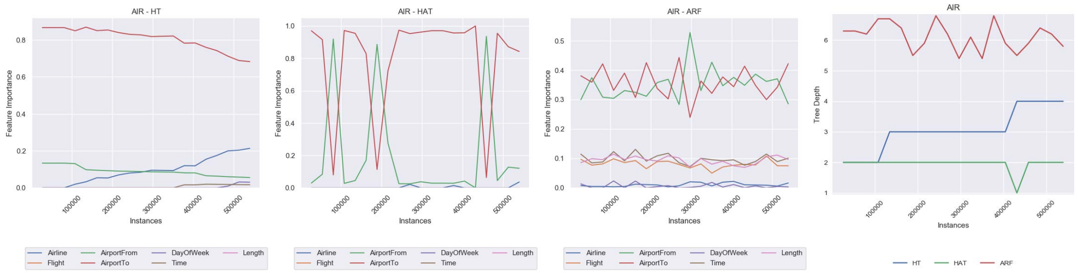
(a) HT **(b) HAT** **(c) ARF** **(d) Tree depth**

Figure 1: Results obtained in the BG1 experiment.



(a) HT **(b) HAT** **(c) ARF** **(d) Tree depth**

Figure 2: Results obtained in the SEA experiment.



(a) HT **(b) HAT** **(c) ARF** **(d) Tree depth**

Figure 3: Results obtained in the AIR experiment.

REFERENCES

[1] Jean Paul Barddal, Heitor Murilo Gomes, Fabricio Enembreck, and Bernhard Pfahringer. 2017. A survey on feature drift adaptation: Definition, benchmark, challenges and future directions. *Journal of Systems and Software* 127 (2017), 278 – 294. https://doi.org/10.1016/j.jss.2016.07.005

[2] Jean Paul Barddal, Heitor Murilo Gomes, Fabrício Enembreck, Bernhard Pfahringer, and Albert Bifet. 2016. On Dynamic Feature Weighting for Feature Drifting Data Streams. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II*. 129–144. https://doi.org/10.1007/978-3-319-46227-1_9

[3] Albert Bifet and Ricard Gavaldà. 2009. *Adaptive Learning from Evolving Data Streams*. Springer Berlin Heidelberg, Berlin, Heidelberg, 249–260. https://doi.org/10.1007/978-3-642-03915-7_22

[4] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. 2010. MOA: Massive Online Analysis. *The Journal of Machine Learning Research* 11 (2010), 1601–1604.

[5] Pedro Domingos and Geoff Hulten. 2000. Mining High-speed Data Streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*. ACM, New York, NY, USA, 71–80. https://doi.org/10.1145/347090.347107

[6] Heitor M. Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabricio Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. 2017. Adaptive random forests for evolving data stream classification. *Machine Learning* 106, 9 (01 Oct 2017), 1469–1495. https://doi.org/10.1007/s10994-017-5642-8

[7] Mark A. Hall. 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 359–366.

[8] W. Nick Street and Y. Kim. 2001. A streaming ensemble algorithm (SEA) for large-classification. In *Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM SIGKDD, 377–382.

[9] Geoffrey I Webb, Loong Kuan Lee, Bart Goethals, and Francois Petitjean. 2018. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery* (2018). https://doi.org/10.1007/s10618-018-0554-1