Stream Analysis of Incoming Events Using Various Data Analysis Methods

Matúš CIMERMAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matus.cimerman@gmail.com

Abstract. Data analysis is a non-trivial task and gets even harder when it comes to streaming data analysis. Several constraints need to be matched when analysing data streams, e.g. usage of limited time and memory or real-time latency. In this work, we focus on an ensemble of tools, aiming to ease data analysis process of streaming data for domain experts. Suppose domain expert doesn't have detailed knowledge of data mining methods and algorithms. We propose real-time visualization of results and resulting model emphasising occurred concept-drifts. Such visualization helps domain expert understand how model works and how it was affected by drifts in data stream. Using Hoeffding trees and classification task we evaluate both, method for classification quantitatively and visualization qualitatively.

1 Introduction

Real-time stream processing and analysis had rapid interest gain across industries. Properly selected and applied data analysis methods bring valuable insights. Strategic business and enterprise decisions are often driven by such insights. Traditional batch methods fail when used on real-time data, because their core requirement is to have all data available right now and moreover in memory.

Some systems require hard real-time response. Car breaks or airplane gyroscope must act in hard real-time which means exactly in the moment when required with negligible latency. Soft real-time systems are not being life critical if responds with small latency. As an example, social network Twitter where latency orders of magnitude in hundreds of milliseconds is acceptable until tweet is propagated to all followers.

Three well known issues associated with data processing are present: volume, velocity and veracity. Potentially infinite data streams obviously represent large volume data to be stored, processed and analyzed. Volume of data only grows with time. Another issue is velocity or rate at which data arrives and is needed to process. Sensor networks, or even sensors of one unit like an airplane, can produce hundreds of terabytes data in a short time [2]. Veracity is also an issue when dealing with data streams. When processing data for longer time window there is increasing chance concept drift occurs in data.

2 Related work

We define data stream as stream of potentially infinite sequence of elements [1]. Following constraints needs to be always satisfied while dealing with data streams [3][4]: *single-pass through data*, *real-time latency*, *usage of limited resources (memory, CPU)*, *concept-drift detection*.

Classification is a process of finding general model based on previous labeled observations. Constructed model is then used to classify new observations. Such classification methods are for example: Naïve Bayes, neural networks, k-nearest neighbors or decision trees [5]. Decision trees are often being used in practice thanks to their relative good accuracy and ease of model understanding [7][8][9][6]. Decision trees like ID3 and C4.5 aren't well suited for volumes of data. Several scalable decision trees methods have been introduced like SLIQ, RainForest or BOAT [6], but none of these can be properly used on data streams.

The most often used decision tree method, designed specifically for data streams, is Hoeffding tree (HT). HT is famous implementation of decision trees in the domain of data streams [9][6][5]. HT requires reading every new observation at most once. This characteristic of algorithm makes it suitable for data streams where memory and CPU bounds are critical. Hoeffding trees are in fact decision trees using Hoeffding bound. This bound is used to guarantee the fact, that small sample of observations are sufficient to select optimal decision attribute.

[*] Master degree study programme in field: Information Systems
Supervisor: Dr. Jakub Ševcech, Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

IIT.SRC 2017, Bratislava, April 27, 2017, pp. 1–4.

Suppose we have N independent observations of random variable $r \in \mathbb{R}$ where r is attribution selection metric (e.g. Gini index or information gain). If we calculate average of observed $\bar{r}$, Hoeffding bound then asserts that actual average r is at least $r - \epsilon$ with probability $1 - \delta$ [9]. Parameter δ is defined by user and

$$\epsilon = \sqrt{\frac{R^2 ln\left(1/_{\delta}\right)}{2n}}$$

HT algorithm uses Hoeffding bound to select the smallest number N – number of observations needed to be processed in tree node to select splitting attribute.

3 Interpretation and explanation of model

Interpretation and explanation of model and result of analytical task is critical component in data analysis process. For user, it's essential to understand what led to form a given model, so it is not just a black-box. Good example can be neural networks, usually hard to explain, and decision trees which are self-explanatory.

Integral part of model presentation is its relevant visual presentation for user understanding. Intention of such visualizations is to find balance between cognition and perception, so brain's full-potential is utilized. Appropriately explained model boosts user's trustworthiness in a model [11][12].

Recent research focused on sampling data represent data as temporal component. In this context, visualization is a summarization [11]. Challenge is to represent temporal component in 2D visualization. Another attribute, making visualization and model explanation ambitious task, is presentation of concept-drift. One approach could be to use parallel visualization for every change that occurs in data stream. This is suitable for short time window or data streams with carrying very little concept-drifts. Problem arises with growing amount of concept-drifts when visualization begins to be confusing leading to change blindness phenomena [11].

We identify problem of such visualization to deal with data streams where number of concepts are present as well as visualization more complex models such as decision trees. Also, these visualizations frequently lack of qualitative or expert evaluation, but rather subjective conclusions are made.

4 Our approach

In this work, we focus on classification task and model visualization particularly. We are using Hoeffding tree with ADWIN algorithm to detect concept-drift in stream. We emphasize to meet constraints when dealing with streaming data. Also, given classifier is immediately available for usage. Much attention is aimed on concept-drift detection. Resultant model, together with concept occurred in stream, is then presented to user with by visualization. Besides of that, goal of this method is that, mostly domain experts will understand resultant model without detailed knowledge of internal functions of method.

4.1 Data stream classification

We implement Hoeffding tree using Adaptive Windowing (ADWIN) algorithm [10] to detect changes in stream. Decision to experiment with ADWIN to detect concept-drift in stream was made based upon experiment where CVFDT algorithm wasn't sufficient. Algorithm ADWIN doesn't require any parameter adjusting like window size. Only one parameter is required, δ which is required by HT algorithm as well and it has the same meaning.

When change is detected, new alternating tree starts to grow in given node of tree. Alternating tree must read defined number of observations until it replaces original subtree with alternating tree, if quality limitation is matched. At the same time, can simultaneously existing more than one alternating tree. Situation when none of alternating trees replaces original tree might happen. Such situation occurs when: *not enough observations have been observed in given alternating tree*, *Hoeffding bound is not satisfied*, *quality of alternating tree is worse than original subtree*.

These events, especially evolution of influencing alternating trees together with their quality and Hoeffding bound we find useful to visualize.

4.2 Model visualization

Assuming domain experts doesn't have detailed knowledge of how internally classification method works, we aim to ease understanding of created model through our interactive real-time visualization. We've selected classification method of decision trees, because it's easy to understand even without previous awareness of decision trees [5]. We use an assumption that decision trees are self-explanatory or that they can be rapidly understood by the user.

Major motivation we seek in model visualization and explanation is that (to our best knowledge) very little attention is dedicated to this problem. Only two research works have been found which are focusing on visualization of changing data streams [11][14]. Major issue of these approaches is lack of real-time visualization.

We focus not only on visualization of resulting model at given time, but also on events of concept-drifts which influenced model at the moment. Concept is possible to re-run as an animation in historical overview look at the model. Furthermore,

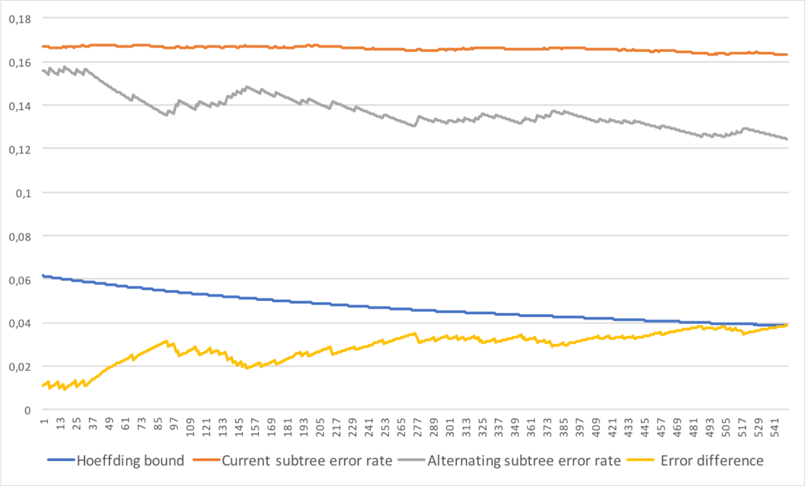

Figure 1: Progress of subtree replaced by its alternating subtree

we provide insights on individual alternating trees. These insights will contain such as Hoeffding bound evolution and quality of given alternating tree. Hoeffding bound and error rate is presented for every subtree and its alternating tree, if it exists. Another view will be composed from significant changes over time which affected the model. Thanks to these, user has quick overview of changes which influenced model. Also, significance of these changes can be observed. For every alternating tree, we will predict probability of replacement of origin. This property may help domain experts to better adapt on gradual, yet significant concept-drifts. On Figure 2: Simplified visualization of small model with 3 classes and 10 split nodes (2 split nodes are collapsed) you can see simplified visualization of very simple model. Visualization is interactive, you can collapse/expand nodes and rerun evolution of model for a given time.

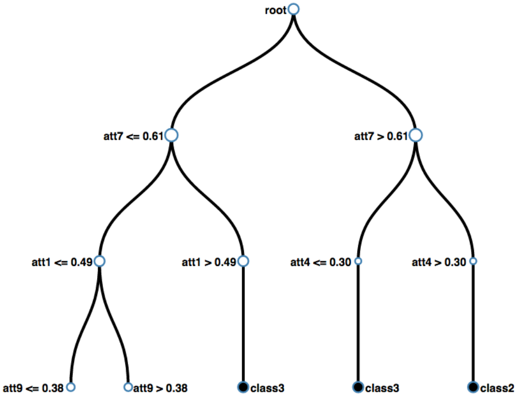

Figure 2: Simplified visualization of small model with 3 classes and 10 split nodes (2 split nodes are collapsed)

Problem of this visualization arise with increasing number of observations if numerous attribute values are present. This leads to many split nodes with questionable significance.

5 Evaluation

To evaluate relevance of proposed method we propose and perform few initial experiments. Since we are working with streaming data we will firstly evaluate integration of our method. More precisely, we will measure performance metrics measuring throughput. We are mostly concerned of fault tolerance, real-time latency and resource allocation depending on data stream volume.

Evaluation of classification method itself we will measure common metrics used for evaluation machine-learning models, such as: precision, recall and F1 measure. Since we're working with streams, also Kappa statistics are measured using Test-Then-Train and Prequential evaluation techniques.

We pay great attention to qualitative evaluation of visualization. First, we will evaluate application with three experts. These experts do have detailed knowledge of machine learning methods. Another experiment design is user study with domain experts without detailed knowledge of machine learning algorithms and methods. User study will be performed in known environment for every experiment participant. Participants will be performing following tasks: How many concept-drifts occurred. What concept-drifts affected current model? Is Hoeffding bound altering model? Given subtree and alternating tree, why wasn't original subtree replaced by an alternating tree? Can you find most significant concept-drift which affected resultant model? How would you describe model in one sentence? Most of these questions will be set at the scale from 1 to 5. User study has not been performed yet.

First experiments were performed to determine whether we are able successfully visualize all mentioned metrics. In the Figure 3: Progress of subtree replaced by its alternating subtree you can see number of alternating trees spawned in experiment using synthetic labeled data with three classes. As you can see, most of the alternating trees are creating in the early stages of the model. This is mainly due to imbalance of the tree. Later, very few alternating trees are created. This, as we found out, is caused by weak concept-drifts available in synthetic data stream. This we identify as an issue and future experiments will be performed on real-world data.

In Figure 3: Progress of subtree replaced by its alternating subtree you can see error rate and Hoeffding bound progress of given subtree along with its alternating subtree. As you can see, subtree is not replaced right away in the moment when alternating subtree is created. Subtree is being replaced in the moment when error rate difference of subtrees satisfied Hoeffding bound. This experiment was performed using MOA artificial stream generator generating 1 million events using configuration [13]: *generators.HyperplaneGenerator -c 4 -k 5 -t 0.10.* Results of experiments shows that accuracy of classification method is close to the one presented in [13] where 86.86% was achieved with similar approach in compare to our accuracy 88.99%. These results indicate applicability of selected classification method in future qualitative experiments where model explanation will be evaluated with domain experts.

6 Conclusions

In this work, we are adopting Hoeffding trees to perform classification task on data streams. Goal is to provide easy to use data analysis method for domain expert. Domain expert doesn't need to have detailed knowledge how this method works internally. We focus on interpretation, explanation and visualization of resultant model and results. For this task, we propose interactive web-based real-time visualization. Visualization provides overview how model evolved over time. We experimentally measured number of alternating trees and error rate of Concluding based upon these observation next steps. We need to enrich visualization with Hoeffding bound progress together with given quality metrics of subtrees. Real-time alerting and prediction of significant drifts we take into consideration. All future experiment needs to be performed on the real-world data containing reasonable concept-drifts. Finally, user study will be performed to qualitatively evaluate enlisted improvements.

Acknowledgement: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] TRAN, Dang-Hoan; GABER, Mohamed Medhat; SATTLER, Kai-Uwe. Change detection in streaming data in the era of big data: models and issues. ACM SIGKDD Explorations Newsletter, 2014, 16.1: 30-38.

[2] NEDELCU, Bogdan, et al. About big data and its challenges and benefits in manufacturing. Database Systems Journal, 2013, 4.3: 10-19.

[3] GABER, Mohamed Medhat; ZASLAVSKY, Arkady; KRISHNASWAMY, Shonali. Mining data streams: a review. ACM Sigmod Record, 2005, 34.2: 18-26.

[4] BABCOCK, Brian, et al. Models and issues in data stream systems. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002. p. 1-16.

[5] NGUYEN, Hai-Long; WOON, Yew-Kwong; NG, Wee-Keong. A survey on data stream clustering and classification. Knowledge and information systems, 2015, 45.3: 535-569.

[6] AGGARWAL, Charu C. A Survey of Stream Classification Algorithms. 2014.

[7] JIN, Ruoming; AGRAWAL, Gagan. Efficient decision tree construction on streaming data. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003. p. 571-576.

[8] HULTEN, Geoff; SPENCER, Laurie; DOMINGOS, Pedro. Mining time-changing data streams. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001. p. 97-106.

[9] DOMINGOS, Pedro; HULTEN, Geoff. Mining high-speed data streams. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000. p. 71-80.

[10] BIFET, Albert; GAVALDÀ, Ricard. Adaptive learning from evolving data streams. In: International Symposium on Intelligent Data Analysis. Springer Berlin Heidelberg, 2009. p. 249-260.

[11] DEMSAR, Jaka; BOSNIC, Zoran; KONONENKO, Igor. Visualization and concept drift detection using explanations of incremental models. Informatica, 2014, 38.4: 321.

[12] BARLOW, S. Todd; NEVILLE, Padraic. Case Study: Visualization for Decision Tree Analysis in Data Mining. In: infovis. 2001. p. 149-152.

[13] BRZEZIŃSKI, Dariusz. Block-based and online ensembles for concept-drifting data streams. 2015.

[14] YAO, Yuan; FENG, Lin; CHEN, Feng. Concept drift visualization. JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE, 2013, 10.10: 3021-3029.