Visualization of Explanations of Incremental Models

Jaka Demšar, Zoran Bosni, Igor Kononenko
University of Ljubljana, Faculty of Computer and Information Science
Vecna pot 113, SI-1000 Ljubljana, Slovenia
{jaka.demsar0@gmail.com} {zoran.bosnic@fri.uni-lj.si} {igor.kononenko@fri.uni-lj.si}

ABSTRACT: *The temporal dimension that is ever more prevalent in data makes the data stream mining (incremental learning) an important field of machine learning. In addition to accurate predictions, explanations of models and examples are a crucial component as they provide insight into model's decision and lessen its black box nature, thus increasing the user's trust. Proper visual representation of data is also very relevant to user's understanding — visualization is often utilised in machine learning since it shifts the balance between perception and cognition to take fuller advantage of the brain's abilities. In this paper we review visualisation in incremental setting and devise an improved version of an existing visualisation of explanations of incremental models. We discuss the detection of concept drift in data streams and experiment with a novel detection method that uses the stream of model's explanations to determine the places of change in the data domain.*

Key words: Visulaization, Data Representation, Data Models, Machine Learning

Received: 17 May 2019, Revised 3 August 2019, Accepted 18 August 2019

DOI: 10.6025/jic/2019/10/4/121-127

1. Introduction

Data streams are becoming ubiquitous. This is a consequence of the increasing number of automatic data feeds, sensoric networks and internet of things [1]. The defining characteristics of data streams are their transient dynamic nature and temporal component. In contrast with static datasets (used in batch learning), data streams (used in incremental learning) are large, changing, semi-structured and possibly unlimited. This poses a challenge for storage and processing as the data can be only read once. For incremental learning models, operations of model increment and decrement are vital. Concepts and patterns in data domain can change (*concept drift*) - we need to adapt to this phenomenon or the quality of our predictions deteriorates. According to *PAC* (*Probably approximately correct*) learning model, if the distribution, generating the instances is stationary, the error rate for sound machine learning algorithms will decline towards the Bayes error rate as the number of processed instances increases [9]. Consequently, when a statistically significant rise in error rate is detected, we can suggest that there has been a change in the generating distribution - concept drift.

The basis of *statistical process control* (*SPC*) [5] is detecting statistically significant error rate (using the central limit theorem) by monitoring the mean and standard deviation of a sequence of correct classification indicators.

Another method, *Page-Hinkley test* [10] was devised to detect the change of a Gaussian signal and is commonly used in signal processing.

Bare prediction quality is not a sufficient property of a good machine learning algorithm. *Explanation* (a form of data postprocessing) of individual predictions and model as a whole is needed to increase the user's trust in the decision and provide insight in the workings of the model, which increases the models credibility. *IME* (*Interactions-based Method for* Explanation) [13] with its efficient adaptation [12] is a model independent method of explanation, which also addresses interactions of features and therefore successfully tackles the problem of redundant and disjunctive concepts in data. The explanation of the prediction for each instance is defined as a vector of contributions of individual feature values. Positive contibution implies that the particular feature value positively influenced the prediction (and vice versa) while the absolute value of a contributon is proportional to the magnitude of influence on the decision, i.e. the importance of that feature value. The sum of all contributions is equal to the difference between the prediction using all feature values and a prediction using no features (prediction difference). The explanation of a single prediction can be expanded to the whole model [12] and also to incremental setting [3]. In the latter case, drift detection (SPC) and adaptation are used to compensate for concept drift. Explanation of a data stream is therefore itself a data stream.

Related to exaplantion is *data visualisation* - a versatile tool in machine learning that serves two purposes; sense-making (data analysis) and communication as it conveys abstract concepts in a form, understandable to humans (it shifts the balance between perception and cognition to take fuller advantage of the brain's abilities [4]). The majority of published visualizations depict data that has a temporal component [8]. In this context, visualization acts as a form of summarization, since the datasets can be extremely large. The challenge lies in representing the temporal component (including concept drift), especially if we are limited to two-dimensional non-interactive visualisations.

The main goal of this paper is improving the existing methodology for visualising explanations of incremental models [3]. The feature value contributions are represented with customised bar charts. Multiple such charts are required to explain the model at different points in time. They become very difficult to read as a whole because of the large number of visual elements that we have to compare (we sacrifice macro view completely in favour of micro view). To consolidate these images and address the *change blindness* phenomenon, charts are stacked into a single plot, where the age and size of the exaplanation are represented with transparency (older and "smaller" explanations fade out). The resulting visualisation is not tainted by first impressions (as it is only one image) and is adequately dense and graphically rich. However, the major flaw of this approach lies in the situations when columns, representing newer explanations override older ones and thus obfuscate the true flow of changing explanations, for example, when the concept drift precipitates the attribute value contributions to increase in size without changing the sign. Concepts can therefore become not only hidden; what's more, the visualization can be deceiving, which we consider to be worse than just being too sparse. Therefore, we need to clarify the presentation of the concept drift along with an accurate depiction of each explanation's contributions while maintaining the macro visual value, that enables us to detect patterns and get a sense of true concepts and flow of changes behind the model.

An additional goal was to devise a method of concept drift detection which monitors the stream of explanations and detects anomalies in it; the detected anomalies are interpreted as a concept drift. We test the improved visualization and the novel concept drift detection method on two datasets and evaluate the results.

2. Visualisation for Incremental Models

When visualising explanations of individual predictions, horizontal bar charts are a fitting method also in the incremental setting. Individual examples are always explained according to the current model which, in our case, can change. This is not an obstacle, since the snapshot of the model is in fact the model that classified the example.

This approach fails with explanations of incremental models as we need a new figure for each local explanation. To successfully represent the temporality of incremental models, we use two variations of a line plot where the x axis contains time stamps of examples and the splines plotted are various representations of contributions (y axis).

The first type of visualization (Figures 2 and 3) has one line plot for each attribute. Contributions of values of the individual attribute are represented with line styles. The mean positive and mean negative contribution of the attribute as a whole are

represented with two thick faded lines. Solid vertical lines indicate the spots where explanation of the model was triggered (and therefore become the joints for the plotted splines), while dashed vertical lines mark the places where the actual concept drift occurs in data. The second type is an aggregated version (Figure 3) where the mean positive and mean negative contributions of all attributes are visualized in one figure. In these two ways we condense the visualization of incremental models without a significant loss in information while still providing a quality insight into the model. Exact values of contributions along with timestamps of changes can be read out (micro view), while general patterns and trends can be recognised in the shapes of lines that are intuitive representations of flowing time (macro view). The resulting visualisations are dense with information, easily understandable (conventional plotting of independent variable, time, on x axis) and presented in gray-scale palette, making them more suitable for print.

3. Detecting Concept Drift using the Stream of Explanations

When explaining incremental models, the resulting explanations are, in themselves, a data stream. This gives us the option to process it with all the methods used in incremental learning. In our case, we'll devise a method to detect outliers in the stream of explanations and declare such points as places of concept drift. The reasoning behind this is the notion that if the model does not change, then also the explanation of the whole model will not change. When an outlier is detected, we consider this to be an indicator of a significant change in model and thus also in the underlying data. In addition to this, the method provides us with a stream of explanations that is continuous to a certain degree of granularity and so enables us to overview the concepts behind the data at more frequent intervals than the existing explanation methodology.

We use a standard incremental learning algorithm [5] (learn by incrementally updating the model, decrement the models if it becomes too big according to the parameter, rebuild the model if we detect change [6]) and introduce some additional parameters. *Granularity* determines how often the explanation of the current model will be triggered. The generated stream of explanations (vectors of feature value contributions) will be compared using cosine distance. For each new explanation, the average cosine distance from all other explanations that are in the current model, is calculated. These values are monitored using the Page Hinkley test. When the current average cosine distance from other explanations has risen significantly, we interpret that as a change in data domain - concept drift. The last examples are then used to rebuild the model, the Page Hinkley statistic and the local explanation storage are reset (to monitor the new model).

The cosine distance is chosen because, in the case of explanations, we consider the direction of the vector of contributions to be more important than its size, which is very influential in the traditional Minkowski distances. The page Hinkley test is used in favour of SPC because of its superior drift detection times [9] and the lack of need for a buffer - examples are already buffered according to the granularity. The method is therefore model independent.

4. Results

4.1 Testing Methodology and Datasets

We test the novel visualisation method and the concept drift detection method on two synthetic datasets, both containing multiple concepts with various degrees of drift between them. These datasets are also used in previous work [3], so a direct assessment of visualization quality and drift detection performance can be made. The naive Bayes classifier and the nearest neighbour classifier are used. Their usage yields very similar results in all tests, so only results obtained by testing with Naive Bayes are presented.

SEA concepts [11] is a data stream comprising 60000 instances with continuous numeric features $x_i \in [0, 10]$, where $i \in \{1, 2, 3\}$. x_1 and x_2 are relevant features that determine the target concept with $x_1 + x_2 \leq \beta$ where threshold $\beta \in \{7, 8, 9, 9.5\}$. Concepts change sequentially every 15000 examples. Although the changes between the generated concepts are abrupt, class noise is inserted into each block. The instances of second dataset, STAGGER [2], represent geometrical shapes which are in the feature space described by *size*, *color* and *shape*. The binary class variable is determined by one of the three target concepts ($small \wedge green$, $green \vee square$, $medium \vee large$). 4500 instances are divided into four blocks (concept-wise) with examples mixing near the change points according to a sigmoid function, so the dataset includes gradual concept drift.

4.2. Improved Visualizations

Concept drifts in STAGGER dataset are correctly detected and adapted to as reflected in Figure 3. The defined concepts can be

easily recognized from explanations triggered by the SPC algorithm - the change in explanation follows the change in concept. Windows generated by the vertical lines give us insight in local explanations of the model (where the concept is deemed to be constant). Disjunct concepts (2 and 3) and redundant feature values are all explained correctly (e.g. reduncacy of *shape* and disjunction of *size* values in concept 3). Figure 1 demonstrates how classifications of two instances with same feature values can be explained completely differently at different times - adapting to change is crucial in incremental setting. This is also evident in the aggregated visualization, which can be used to quickly determine the importance of each attribute.

For SEA dataset, explanations of instances are tightly corresponding to explanations of the model. As evident in Figure 2, the shape of contributions of features reflects the target concept; lower values increase the likelihood of positive classification and vice versa. Feature x_1 is correctly explained as irrelevant with its only contributions being the result of noise.

4.3. Concept Drift Detection

Evaluating the concept drift detection using the stream of explanations on the STAGGER dataset yielded positive results. As

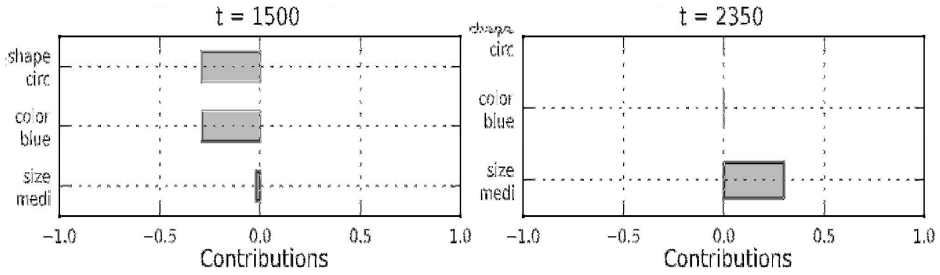

Figure 1. Explanations of a single prediction at different times

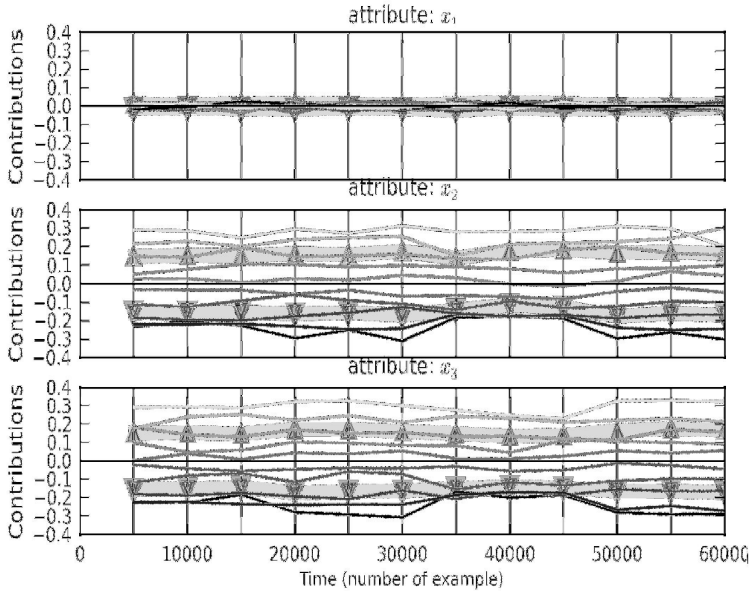

Figure 2. Periodically triggered explanations (SEA)

depicted in Figure 4, the method correctly detects concept drifts without false alarms and is in that regard similar to SPC method. The stream of explanations was similar to those obtained with other successful drift detection methods. Choices of larger granulations yielded similar results, but the change detection was obviously delayed. The concept drift was however never missed, provided that the granulation was smaller that the spacing between sequential changes in data. The delays of concept drift detection are correlated with the magnitude of change. For example, the last concept drift was detected with significant delay. In this regard, the proposed method is inferior to SPC algorithm - the concept drift detection in noticeably delayed and we're also dependant on two parameters – granulation and alert threshold, so the generality of the method is diminished.

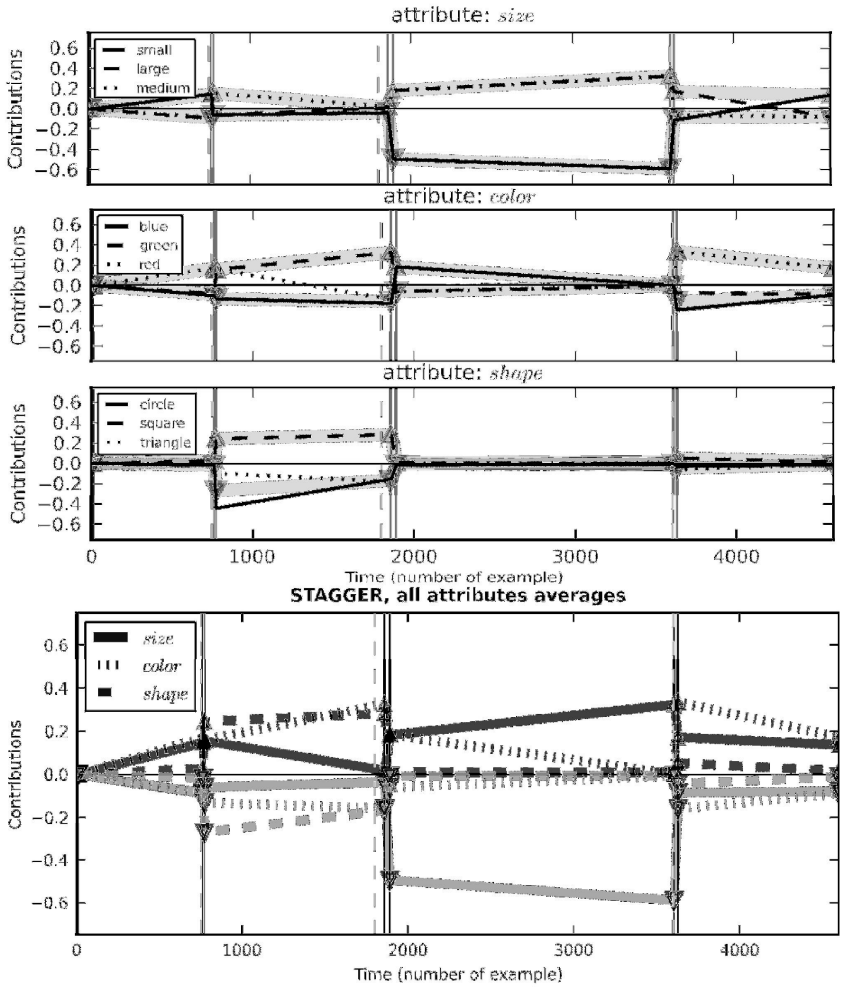

Figure 3. Explanations triggered at change detection (STAGGER)

When testing with SEA datasets, the concept drift was not correctly detected. Changing the granulation and Page Hinkley alert threshold parameter resulted in varying degrees of false alarms or non reaction to change (Figure 4). This behaviour can be attributed to a small magnitude of change that occurs in data - the difference between concepts in data is quite small and continuous. However, when explaining this (incorrectly adapted) model, we still recognise true underlying concepts. This can be attributed to automatically decrementing the model when it becomes too big. It is important to note that this does not perform well in general, if the prior knowledge is insufficient for us to correctly decide on the maximum model size.

We conclude that, in this form, the presented method is not a viable alternative to the existing concept drift detection methods. Its downsides include high level of parametrization which requires a significant amount of prior knowledge and can also become improper if the model changes drastically. Consequently, another assessment of data is needed - the required manual supervision and lack of adaptability in this regard can be very costly and against the requirements of a good incremental model. The concept drift detection is also not satisfactory - it is delayed in the best case or concepts can be missed or falsely alerted in the worst case. Another downside is the time complexity - the higher the granularity the more frequent explanations will be, which will provide us with a good stream of explanations but be very costly time-wise. The method is therefore not feasible in environments where quick incremental operations are vital. However, if we can afford such delays, we get a granular stream of explanations which gives us insight into the model for roughly any given time.

A note at the end: we should always remember that we are explaining the models and not the concepts behind the model. Only if the model performs well, we can claim that our explanations truly reflect the data domain [12]. This can be tricky in incremental learning, as at the time of a concept drift, the quality of the model deteriorates.

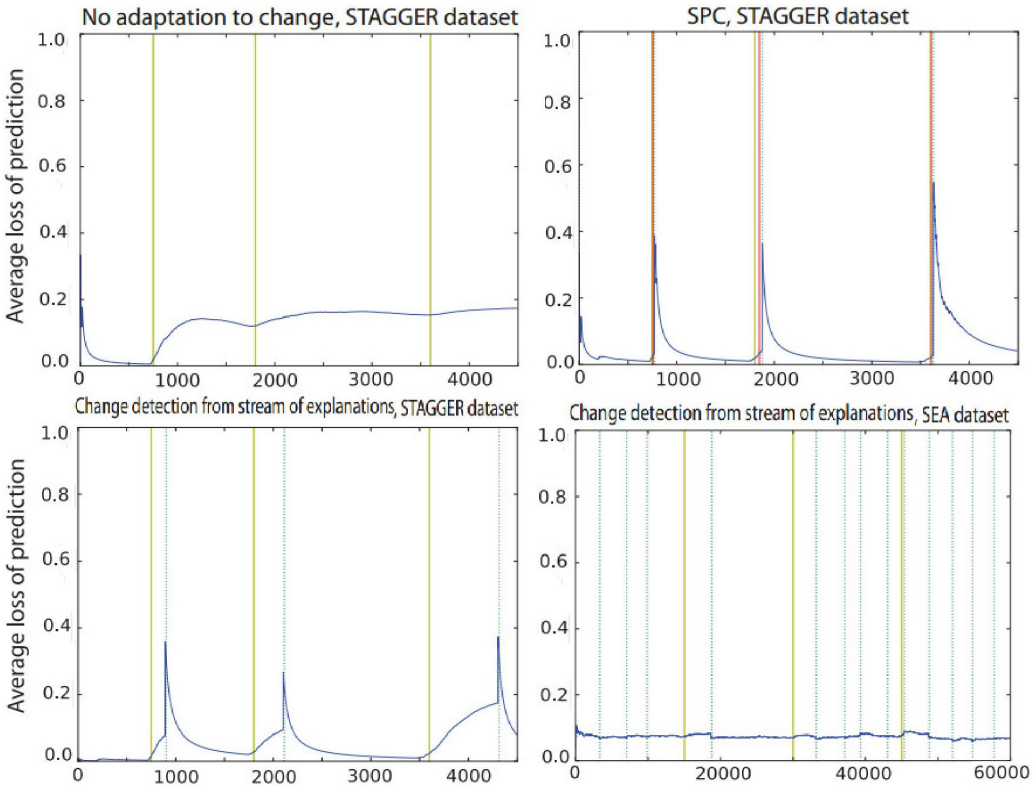

Figure 4. Performance of various change detection methods. Yellow line indicates true change in concepts, green line indicates change detection and adaptation)

5. Conclusion

The new visualization of explanation of incremental model is indeed an improvement compared to the old one. The overriding nature of the old visualisation was replaced with an easy to understand timeline, while the general concepts (macro view) can still be read out from the shape of the lines. Micro view is also improved as we can determine contributions of attribute values for any given time.

The detection of concept drift using the stream of explanations did not prove to be suitable for general use based on the initial experiments. It has shown to be hindered by delayed detection times, missed concept drift occurrences, false alarms, high level of parametrization and potential high time complexity. This provides motivation for further experiments in this field, especially because the stream of explanations provides good insight into the model with accordance to the chosen granulation.

The main goal of future research is finding a true adaptation of IME explanation methodology to incremental setting, i.e. efficient incremental updates of explanation at the arrival of each new example. Truly incremental explanation methodology would provide us with a stream of explanations of finest granularity. In addition to this, a number of new possibilities for visualisation would emerge, particularly those that rely on finely granular data, such as ThemeRiver [7].

References

[1] Aggarwal, C. C., Ashish, N., Sheth, A. P. (2013). The internet of things: A survey from the data-centric perspective. *In*: *Managing and Mining Sensor Data*. Springer.

[2] Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B., Braun, M. Moa: Massive online analysis.

[3] Jaka Demšar. (2012). Explanation of predictive models and individual predictions in incremental learning (In Slovene). B.S. Thesis, University of Ljubljana.

[4] Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 1st edition.

[5] Gama, J. (2010). *Knowledge Discovery from Data Streams.* Chapman & Hall/CRC, 1st edition.

[6] Haussler, D. (1995). Overview of the probably approximately correct (PAC) learning framework.

[7] Havre, S., Hetzler, B., Nowell, L. (2000). Themeriver: Visualizing theme changes over time. *In*: Proc. IEEE Symposium on Information Visualization.

[8] Ratanamahatana, C., Lin 0001, J., Gunopulos, D., Keogh, E. J., Vlachos, M., Das, G. (2005). Mining time series data. *In*: *The Data Mining and Knowledge Discovery Handbook*. Springer.

[9] Sebastião, R., Gama, J. (2009). A study on change detection methods. *In*: *P*rogress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence, EPIA.

[10] Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, 41, 100-115.

[11] Street, W. N., Kim, Y. S. (2001). A streaming ensemble algorithm for large-scale classification. *In*: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '01, New York, NY, USA.

[12] Strumbelj, E., Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11, 1–18.

[13] Strumbelj, E., Kononenko, I., Robnik Sikonja, M. (2009). Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10) 886–904, (October).