Information and Knowing When to Forget It

Rohit Sharma and Ognjen Arandjelović†
School of Computer Science
North Haugh
University of St Andrews
St Andrews KY16 9SX
Fife, Scotland
United Kingdom
†ognjen.arandjelovic@gmail.com

***Abstract*—In this paper we propose several novel approaches for incorporating forgetting mechanisms into sequential prediction based machine learning algorithms. The broad premise of our work, supported and motivated in part by recent findings stemming from neurology research on the development of human brains, is that knowledge acquisition and forgetting are complementary processes, and that learning can (perhaps unintuitively) benefit from the latter too. We demonstrate that if forgetting is implemented in a purposeful and date driven manner, there are a number of benefits which can be gained from discarding information. The framework we introduce is a general one and can be used with any baseline predictor of choice. Hence in this sense it is best described as a meta-algorithm. The method we described was developed through a series of steps which increase the adaptability of the model, while being data driven. We first discussed a weakly adaptive forgetting process which we termed passive forgetting. A fully adaptive framework, which we termed active forgetting was developed by enveloping a passive forgetting process with a monitoring, self-aware module which detects contextual changes and makes a statistically informed choice when the model parameters should be abruptly rather than gradually updated. The effectiveness of the proposed meta-framework was demonstrated on a real world data set concerned with a challenge of major practical importance: that of predicting currency exchange rates. Our approach was shown to be highly effective, reducing prediction errors by nearly 40%.**

I. Introduction

The tasks of information acquisition and of representation of knowledge are central to machine learning and related disciplines such as data mining and pattern recognition [1]. Therefore it should come as no surprise that an increase in data availability is all but universally seen as a positive thing: the more information an agent has at its disposal, the greater should its ability to form accurate models and make reliable predications be [2]. However, in the context of the practical constraints of the real world, this last conclusion is – perhaps unintuitively – not correct, and a learner can in fact benefit from a process which is diametrically opposite to that of information gathering: forgetting. As we will demonstrate, if forgetting is implemented in a purposeful manner, there are a number of benefits which can be gained from discarding information. Herein we specifically focus on temporal, sequentially ordered data, which is indeed highly relevant in a range of different application domains including financial markets [3], health care [4], human behaviour analysis [5], visual tracking [6], data mining from social media [7], [8], evolution of ideas [9], [10], and many others.

Useful insight can be gained by considering how human learning in its various forms benefits from discarding information. Indeed, there is an increasing amount of evidence which demonstrates the importance forgetting in human learning. For example, recent work in neurology suggests that the development of cognitive disorders lying on the autism spectrum [11] is associated with the malfunctioning of the synaptic pruning process [12], [13]. This leads to over-connectedness in certain areas of the brain and explains a range of impairments in social interaction and communication, such as compulsiveness and repetitive motor behaviour [14]. Synaptic pruning is just one type of forgetting whereby information stored in certain neural connections is 'purposefully' (in the context of a natural selection evolved brain) discarded.

Another interesting and insightful example of adaptive human forgetting concerns, amongst others, motoneural forgetting that astronauts experience in environments with different gravitational conditions [15]. Neuromuscular activation patterns used to handle objects, assess their mechanical properties (mass, moment of inertia around a certain axis etc) exhibit changes that can be interpreted as forgetting the 'training data' used to learn this tasks on Earth, and undergo a new learning experience. In contrast to the previous example, in this instance forgetting, in the sense of discarding largely inconsequential prior experiences happens abruptly – there is no slow (in the context of the required adaptations) transition from one to another gravitational environment. As we will shortly discuss in more detail, our framework was designed exactly with these considerations in mind: the need to adapt to a gradually changing environment, as well as to an abruptly altered one.

II. Proposed meta-framework

In Section II-A we start with a technical motivation of the proposed framework. In particular we illustrate why an explicit forgetting model is needed for nuanced, adaptive learning by showing that what may be incorrectly interpreted as implicit learning of forgetting by traditional methods, is in fact not addressing the challenge our work is aimed at solving.

A. Technical motivation

Consider a stream of temporally equispaced observations $\{x_t\} = \ldots, x_{-3}, x_{-2}, x_{-1}, x_0$ where t denotes the discrete timestamp associated with the observation x_t, and x_0 is the most recent observation. The challenge is that of predicting the next, yet unseen observation, that is x_1. Linear regression based prediction can be made using a window which encompasses the most recent l observations i.e. $x_{-l+1}, \ldots, x_0$ by linearly combining their values using the corresponding set of coefficients $c_{-l+1}, \ldots, c_0$:

$$x_1^* = \sum_{i=0}^{l-1} c_{-i} x_{-i}, \tag{1}$$

where x_1^* is the predicted value of x_1.

The optimum values of coefficients c_t can be learnt from historical, training data. In particular, if the entire historical record includes L observations, the optimum values can be learnt by minimizing the historical prediction error in terms of its Euclidean norm:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} (\|\mathbf{X}\mathbf{c} - \mathbf{Y}\|_2)^2 \tag{2}$$

where:

$$\mathbf{c} = [c_{-l+1}, c_{-l+2}, \ldots, c_0]^T, \tag{3}$$

$$\mathbf{X} = \begin{bmatrix} x_{-l} & x_{-l+1} & \ldots & x_{-1} \\ x_{-l-1} & x_{-l} & \ldots & x_{-2} \\ \ldots & \ldots & \ldots & \ldots \\ x_{-L-1} & x_{-L} & \ldots & x_{-L+l-1} \end{bmatrix}, \tag{4}$$

and

$$\mathbf{Y} = [x_0, x_{-1}, \ldots, x_{-L+l}]^T \tag{5}$$

The quadratic form of the optimization task in (2) leads to a closed form solution by simple differentiation with respect to $\mathbf{c}$ and the setting of the differential to zero, yielding:

$$\mathbf{c}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{6}$$

If this method is applied on the currency exchange rate data we describe in Section III, the coefficient values obtained are as shown in Figure 1. There are several things which can be readily noted. Firstly observe that the magnitude of the coefficients decays with the time difference between the current observation ('Day 0') and a past observation. This is to be expected: in simple terms, more recent data matters more. This supports the general premise of our work. The second interesting pattern which emerges is that of alternating signs of consecutive coefficients, suggesting that the recent *trend* (derivative) of the stream is important in the prediction. This too is to be expected from a well behaved function and Taylor series.

At this stage it is tempting to see the decaying coefficient magnitudes as addressing precisely the point which we set out to address with this work: data driven forgetting of historical information. However it is important to understand why this is not the case. What the result in Figure 1 shows is that progressively time distant observations matter less for prediction. However, there is no forgetting taking place here because the fitting process in (2) considers all historical errors as having equal significance. The fitting procedure considers distant historical errors to be as important as recent historical errors thereby demonstrating a lack of its ability to account for model changes across time. This is what we aim to achieve herein.

B. Previous work

Having motivated the challenge at the crux of the present work, in order to contextualize our contribution, we now overview some of the previous work on incorporating forgetting mechanisms which have been described in the machine literature to date. As noted earlier, work in this realm is scarce, and falls under the umbrellas of either preliminary and exploratory work, or simple and *ad hoc* methods.

Rubin and Wenzel [16] performed an empirical study to search for forgetting regularities over a larger number of data sets (210 to be precise) motivated by the premise of finding a universal forgetting function which they termed the retention function. It should be borne in mind that the data sets they adopted were all taken from the previous work in the area of psychology, thereby possibly introducing some domain based bias in the findings. Rubin and Wenzel described a set of 105 two parameter forgetting functions to which the data from the 210 sets was fitted. The functions considered included included the standard linear, hyperbolic, logarithmic, exponential, and power functions, and a number of others. Their preliminary findings identified four, namely the logarithmic, power, hyperbolic, and exponential as the most promising candidates for future work to focus on.

The work of Kahana and Adler [17] also considered the general problem of gradual forgetting and showed both from theory, using a minimal set of assumptions, as well as corroborated using simulations on synthetic data, that an exponential decay provides the most fundamental forgetting model.

A different line in inquiry was adopted by other authors who were interested in the phenomenon known as concept drift. Koychev [18] described a method based on *ad hoc*, manually predefined heuristic rules that abrupt forgetting can effect an improvement in predictive performance.

C. Learning to forget passively

A way of addressing the problem highlighted by our experiment in Section II-A was already alluded to in the discussion of the findings: the importance of model prediction errors on the training data corpus can be weighted using an exponential function whose free parameter – the forgetting rate – can also be learnt from data. Formalizing this model leads to the following extension of (2):

$$(\mathbf{c}^*, T^*) = \arg\min_{(\mathbf{c},T)} \frac{(\|\left(\mathbf{X}\mathbf{c} - \mathbf{Y}\right) \otimes \mathbf{W}\|_2)^2}{(\|\mathbf{W}\|_2)^2} \tag{7}$$

where the meaning of the symbols $\mathbf{c}$, $\mathbf{X}$, and $\mathbf{Y}$ remains the same as before, $\otimes$ denotes element-wise matrix multiplication,

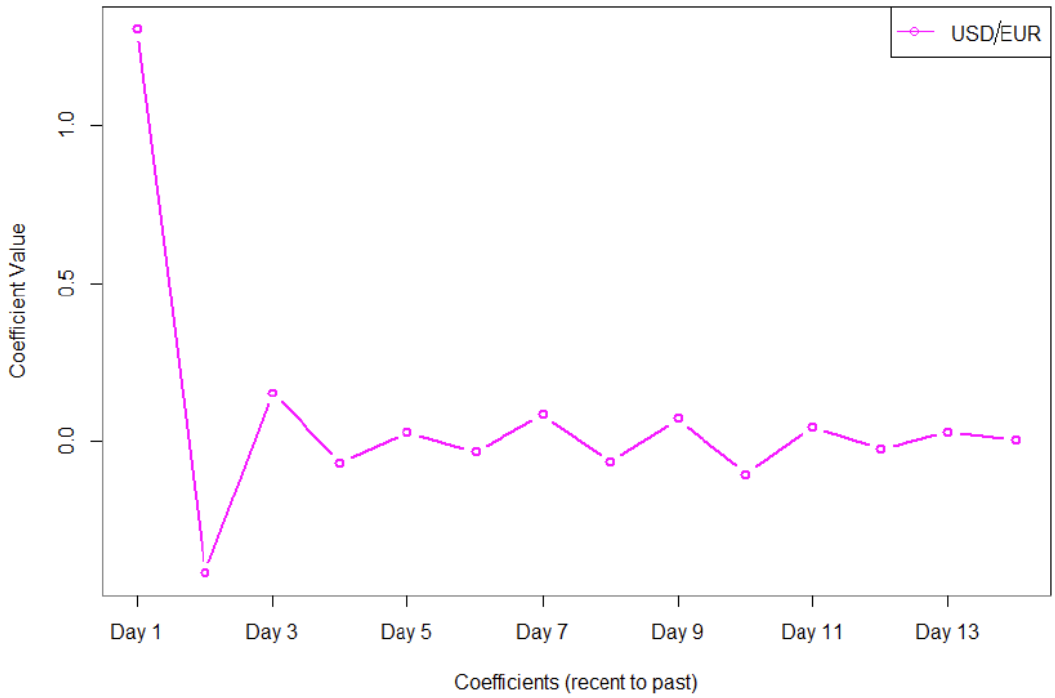

Fig. 1. Linear regression coefficients corresponding to the most recent observations, obtained by optimizing for their value using the entire training data set (3 years) of USD/EUR exchange values (daily averages). Observe (i) that the magnitude of the coefficients decays with the time difference between the current observation ('Day 0') and a past observation, and (ii) the alternating signs of consecutive coefficients, suggesting that it is the recent *trend* (derivative) of the stream that is important for prediction.

and:

$$\mathbf{W} = \begin{bmatrix} e^{-0/T} \\ e^{-1/T} \\ \vdots \\ e^{-l/T} \end{bmatrix}, \tag{8}$$

Unlike the minimization problem expressed in (2), the formulation proposed in (7) cannot be solved in closed form. Instead, we adopt an iterative procedure whereby alternating optimizations of the model parameters, namely the forgetting rate T and the vector of coefficients $\mathbf{c}$, is performed until convergence. More specifically, we start by an informed starting value of $\mathbf{c}_1^*$ (the subscript introduced here refers to the iteration number) estimated without forgetting, i.e. using (6), and then keeping the vector $\mathbf{c}_1^*$ fixed, estimate the best value of T_1^* for these coefficient values. Then, keeping the value of the forgetting rate T_1^* fixed, which reduces the optimization task of (7) to that of (2), we update the estimate of the regression coefficients to produce $\mathbf{c}_2^*$, and so on, until a certain threshold on the fitting quality is met. In particular in this work we terminate optimization when the updates effect an improvement in the relative error smaller than 0.001. For clarity, a flowchart with the key steps in the proposed method is included in Figure 2.

We term the type of forgetting just introduced *passive forgetting*. The reason for this is that while the method learns the average rate at which past observations cease to be important, it cannot account for possible abrupt changes in the phenomena producing observations. Real world examples include events such as market crashes, important political changes (e.g. the outcome of the recent referendum on the membership of the United Kingdom in the European Union – the so-called Brexit), etc. A more complex model is needed to account for this type of forgetting, one which monitors its own performance and detects a change in performance which justifies the forgetting of some data in its entirety.

D. Active forgetting & self-aware monitoring

To motivate the second facet of our algorithm, we estimated the value of the dominant regression coefficient (corresponding to the most recent historical observation) using only a sliding 3 month period across the training data corpus, using our USD/EUR currency exchange data set, introduced and explained in more detail in Section III. Its variation over time is shown in Figure 3. In addition to the expected stochastic variation of the coefficient, what can be readily observed are periods of dips and troughs. As noted in the previous section these are readily traced to major world events, such as those which can result from major changes on the stock market or on the global geopolitical scene. When such changes happen not a gradual but an abrupt way of forgetting historical training data is required. In our method this is achieved by having the predictor constant monitor its performance and when a decline in prediction accuracy is detected, a possible model change is considered. In particular, we assume that expected stochastic changes in the prediction error follow a Gaussian distribution. Then with each incoming observation we check if there is a sequence of the most recent historical observations such that the current model would be expected to encounter such an anomaly only once in 10 cases (a different criterion for possi-

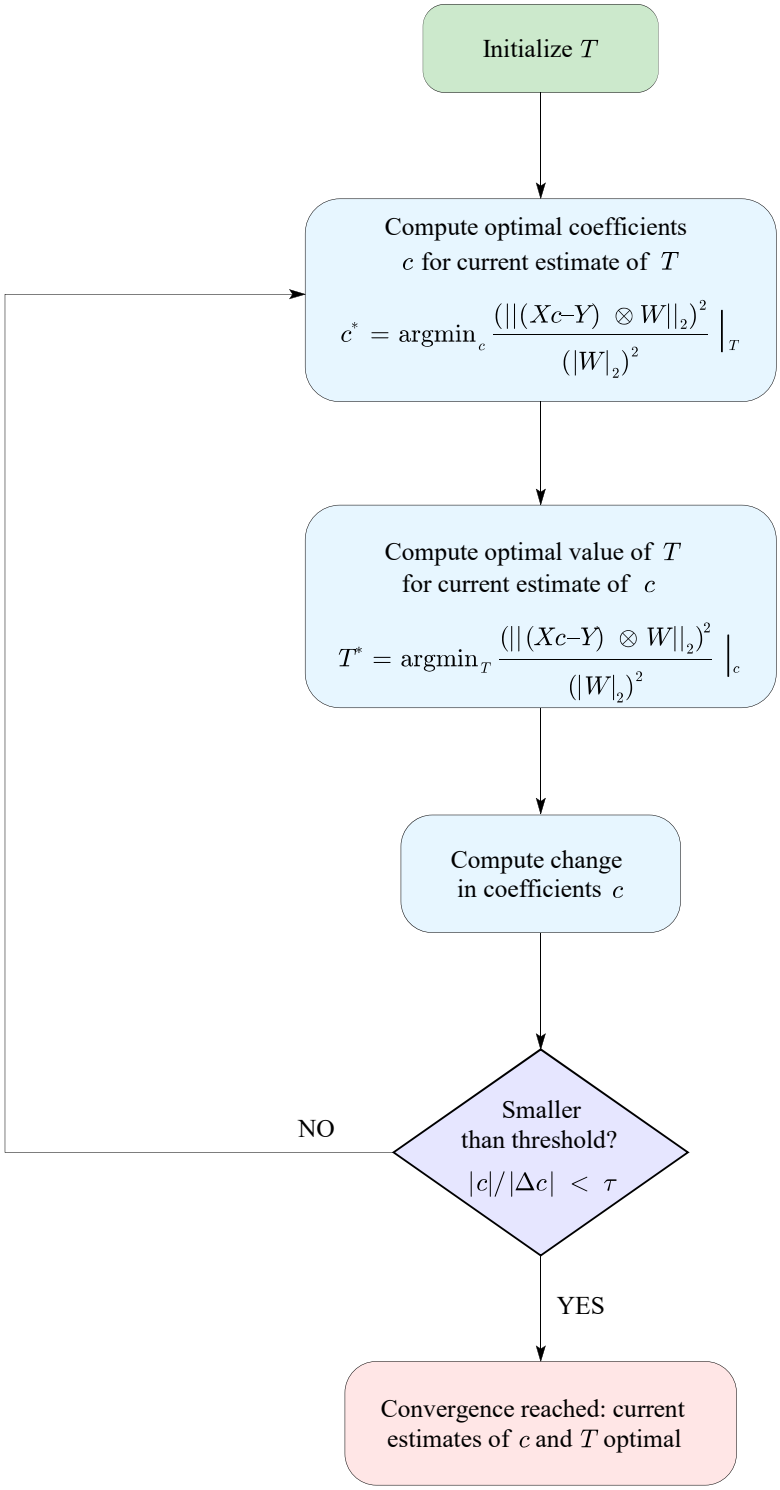

Fig. 2. A flowchart summarizing the proposed method of forgetting which we term *passive forgetting*. This approach successfully learns the average rate at which past observations cease to be important (the forgetting rate) but cannot account for possible abrupt changes in the phenomena producing observations.

ble model switching can be chosen based on the acceptability of risk in a specific application, as well as the understanding of the underlying phenomenon). When this happens a series of alternative hypotheses is postulated by fitting the model introduced in the previous section using varying amounts of most recent historical data. If the performance of one of these models on the outlying historical data is found to be superior to the current model, the superior model is adopted as the current model and the previous estimation and self-monitoring process continued as before.

III. Empirical evaluation

In this section we describe the experiments we conduced to evaluate and assess the performance of the proposed ideas, summarize the most important findings, and discuss their implications for practical application and future research. We being by introducing the data sets we adopted for the evaluation.

A. Data

To maximize the reliability of our results and minimize the possibility of accidental patterns emerging by chance, we sought data sets which are (i) appropriate for the task at hand, (ii) large (thus allowing a large number of predictions across the timespan of the data), (iii) practically relevant and challenging, (iv) publicly available, and (v) diverse in the nature of phenomena they correspond to. In line with these criteria we settled to two data sets, one which concerns currency exchange rates, and one which concerns near Earth surface atmospheric temperature measurements.

More specifically, the first data set used includes the mean daily exchange rate between the United States of America dollar (USD) and the official currency of the Eurozone – the Euro (EUR). We obtained the maximum amount of data freely available from OANDA web site: https://www.oanda.com/solutions-for-business/historical-rates/main.html.

The second data set we used in our experiments concerns the average daily temperature covering the period of the last 10 years. Much like the previous corpus it is available freely and can be obtained from the UK Meteorological Office web site: http://www.metoffice.gov.uk/climatechange/science/monitoring/ukcp09/available/daily.html.

B. Results

In order to assess the importance of both facets of our framework, namely its passive and active forgetting mechanisms, we performed evaluations first using passive forgetting only first, and then using the method as a whole. In all cases we examined the average Euclidean error (i.e. the L_2 norm) across the timespan of the data, relative to the error obtained by a non-forgetting baseline predictor.

Our results for the USD/EUR exchange rate and average daily temperature are shown, respectively, in Table I and Table II. Both qualitatively and quantitative the relative performances of passive and active forgetting algorithms are in agreement. Firstly, in all cases incorporating forgetting effected a performance improvement over the non-forgetting baseline. Passive forgetting reduced the average error rate by approximately 5% (to be precise, 4.33% and 6.43% respectively). The main result, however, is the remarkable error reduction achieved with the use of active forgetting. On the currency exchange corpus the prediction error was decreased by 36.31%, and on the temperature data set by 32.08%. All of the aforementioned findings confirm firstly the broad premise of the present work, demonstrating the importance of forgetting in sequential prediction machine learning algorithms, as well as the specific ideas underlying the newly proposed forgetting models.

TABLE I. Prediction performance improvement on the USD/EUR currency exchange rate data, achieved using (i) the passive forgetting framework introduced in Section II-C, and (ii) active forgetting of Section II-D i.e. the final meta-algorithm which envelopes the passive forgetting method with a self-monitoring agent able of making abrupt forgetting decisions. Both methods yield an improvement, the final meta-algorithm effecting a dramatic average reduction error of nearly 40%.

Method	Improvement over no forgetting
Passive forgetting	4.33%
Active forgetting	36.31%

TABLE II. Prediction performance improvement on the mean daily temperature rate data, achieved using (i) the passive forgetting framework introduced in Section II-C, and (ii) active forgetting of Section II-D i.e. the final meta-algorithm which envelopes the passive forgetting method with a self-monitoring agent able of making abrupt forgetting decisions. Both methods yield an improvement, the final meta-algorithm effecting a dramatic average reduction error of over 30%.

Method	Improvement over no forgetting
Passive forgetting	6.43%
Active forgetting	32.08%

As an additional illustration of the outstanding performance of our method in Figure 4 we also plotted the true USD/EUR exchange rate over the period of 3.5 years and superimposed the prediction of the proposed method. This plot shows that in addition to achieving a very low average error rate, our prediction tracks the ground truth remarkably closely, never erring significantly (the plot is shown in large magnification so that errors can at all be perceived by the naked eye). This has major practical implications: even if an algorithm achieves a low average error rate, even short lasting but significant (in magnitude) transient errors can lead to, say, major financial losses. The fact that our algorithms exhibits such consistently superior performance illustrates its robustness and appropriateness for deployment in the real world.

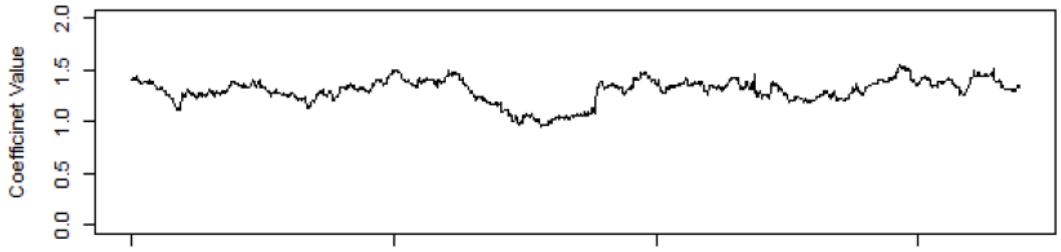

Fig. 3. Variation in the dominant regression coefficient (corresponding to the most recent historical observation) with time, when only a 3 month period of training data is used, using our USD/EUR currency exchange data set (see Section III). In addition to the expected stochastic variation of the coefficient, what can be observed are periods of dips and troughs which correspond to salient changes in the underlying phenomenon, such as those which can result from major changes on the stock market or on the global geopolitical scene.

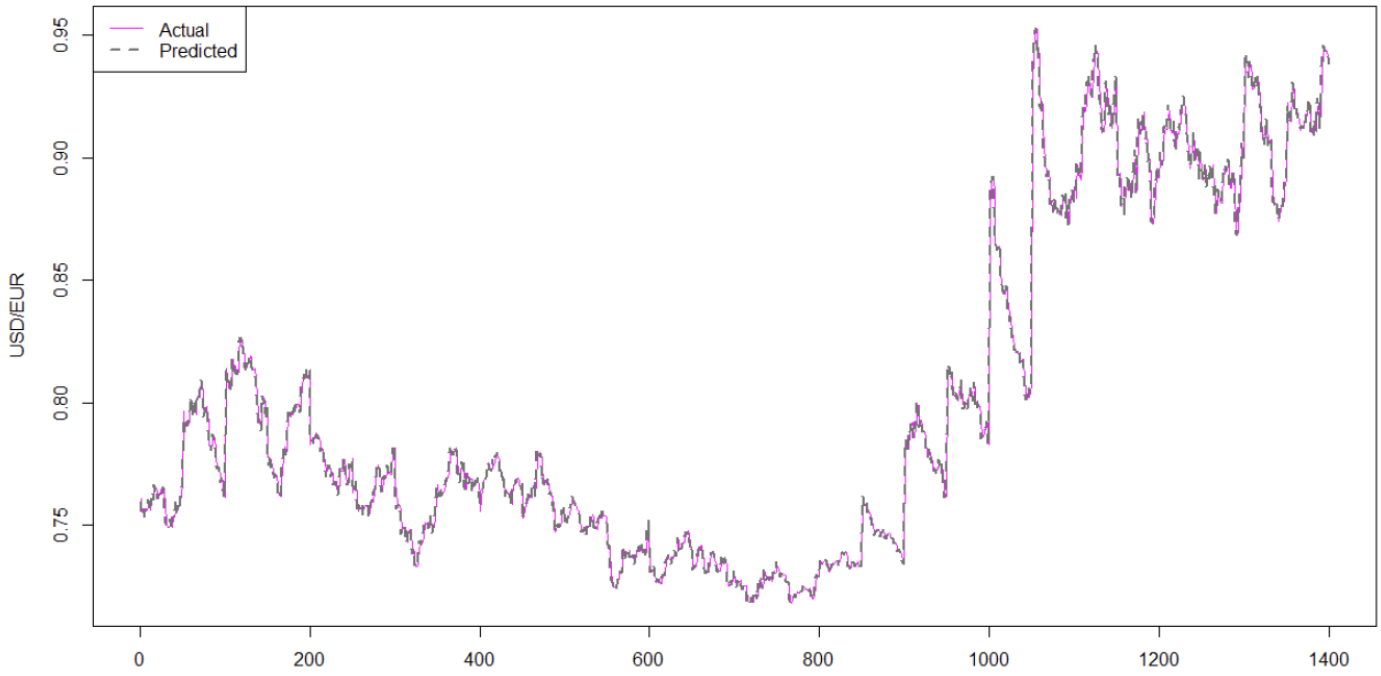

Fig. 4. Plot of the ground truth (solid pink) and the prediction of our active forgetting meta-algorithm (dotted blue), of 3.5 years of the mean daily USD/EUR exchange rate. Our prediction tracks the ground truth remarkably closely, achieving both a very low average error rate (also see Table I) as well as a very low maximum error across the timespan. In particular notice that no major errors can be noticed at any point in the plot. The plot is shown in large size so that errors can be perceived by the naked eye.

IV. Summary and Conclusions

There is a mounting body evidence that forgetting plays a crucial role in human learning. Yet, this aspect of adaptive learning has all but entirely failed to be recognized in the existing machine learning literature. The few attempts at incorporating forgetting into the learning process have treated the issue rather superficially, relying on only simplistic models, and *ad hoc*, non-adaptively set parameters.

In this paper we introduced a novel approach for incorporating forgetting into sequential machine learning algorithms. Our approach is general and can be used with any baseline predictor of choice. Hence in this sense it can be described as a meta-algorithm. The method we described was developed through a series of steps which increase the adaptability of the model, while being data driven. We first discussed a weakly adaptive forgetting process which we termed passive forgetting. A fully adaptive framework, which we termed active forgetting was developed by enveloping a passive forgetting process with a monitoring, self-aware module which detects contextual changes and makes a statistically informed choice when the model parameters should be abruptly rather than gradually updated. The effectiveness of the proposed meta-framework was demonstrated on two real world data sets concerned with challenges of major practical importance: those of predicting currency exchange rates and daily temperatures. On both tasks our approach was shown to be highly effective, dramatically reducing prediction errors.

Our future work will explore possible applications of the proposed framework to face recognition when the training and query data are separated by a large time gap, which is an outstanding practical and research challenge [19], to the prediction of disease progression patterns [20], [21], [22], and a number of other domains.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer-Verlag, 2007.

[2] E. Junqué de Fortuny, D. Martens, and F. Provost, "Predictive modeling with big data: is bigger really better?" *Big Data*, vol. 1, no. 4, pp. 215–226, 2013.

[3] T. G. Andersen and T. Bollerslev, "Intraday periodicity and volatility persistence in financial markets." *Journal of Empirical Finance*, vol. 4, no. 2, pp. 115–158, 1997.

[4] O. Arandjelović, "Discovering hospital admission patterns using models learnt from electronic hospital records." *Bioinformatics*, vol. 31, no. 24, pp. 3970–3976, 2015.

[5] ——, "Contextually learnt detection of unusual motion-based behaviour in crowded public spaces." *In Proc. International Symposium on Computer and Information Sciences*, pp. 403–410, 2011.

[6] R. Martin and O. Arandjelović, "Multiple-object tracking in cluttered and crowded public spaces." *In Proc. International Symposium on Visual Computing*, vol. 3, pp. 89–98, 2010.

[7] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli, "Data-mining Twitter and the autism spectrum disorder: a pilot study." *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pp. 349–356, 2014.

[8] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh, "Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis." *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pp. 1354–1361, 2015.

[9] V. Andrei and O. Arandjelović, "Identification of promising research directions using machine learning aided medical literature analysis." *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2471–2474, 2016.

[10] ——, "Complex temporal topic evolution modelling using the Kullback-Leibler divergence and the Bhattacharyya distance." *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 1, pp. 1–11, 2016.

[11] S. E. Levy, D. S. Mandell, and R. T. Schultz, "Autism." *Lancet*, vol. 374, no. 9701, pp. 1627–1638, 2009.

[12] A. H. Stephan, B. A. Barres, and B. Stevens, "The complement system: An unexpected role in synaptic pruning during development and disease." *Annual Review of Neuroscience*, vol. 35, pp. 1063–1070, 2012.

[13] G. Tang, K. Gudsnuk, S.-H. Kuo, M. L. Cotrina, G. Rosoklija, A. Sosunov, M. S. Sonders, E. Kanter, C. Castagna, A. Yamamoto, Z. Yue, O. Arancio, B. S. Peterson, F. Champagne, A. J. Dwork, J. Goldman, and D. Sulzer, "Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits." *Neuron*, vol. 83, no. 5, pp. 1131–1143, 2014.

[14] "Autism spectrum disorder fact sheet," *American Psychiatric Publishing*, 2013.

[15] A. B. Newberg, "Changes in the central nervous system and their clinical correlates during long-term spaceflight." *Aviation, Space, and Environmental Medicine*, 1994.

[16] D. C. Rubin and A. E. Wenzel, "One hundred years of forgetting: a quantitative description of retention." *Psychological Review*, vol. 103, no. 4, p. 734, 1996.

[17] M. J. Kahana and M. Adler, "Note on the power law of forgetting." University of Pennsylvania, unpublished note., Tech. Rep., 2002.

[18] I. Koychev, "Gradual forgetting for adaptation to concept drift." *In Proc. ECAI Workshop on Current Issues in Spatio-Temporal Reasonning*, 2000.

[19] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, "Infrared face recognition: a literature review." *In Proc. IEEE International Joint Conference on Neural Networks*, pp. 2791–2800, 2013.

[20] I. Vasiljeva and O. Arandjelović, "Prediction of future hospital admissions – what is the tradeoff between specificity and accuracy?" *In Proc. International Conference on Bioinformatics and Computational Biology*, pp. 3–8, 2016.

[21] ——, "Towards sophisticated learning from EHRs: increasing prediction specificity and accuracy using clinically meaningful risk criteria." *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2452–2455, 2016.

[22] ——, "Automatic knowledge extraction from EHRs." *In Proc. International Joint Conference on Artificial Intelligence Workshop on Knowledge Discovery in Healthcare Data*, 2016.