Predictive Analytics on Evolving Data Streams

Anticipating and adapting to changes in known and unknown contexts

Mykoa Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
The Netherlands

INVITED TALK EXTENDED ABSTRACT

Ever increasing volumes of sensor readings, transactional records, web data and event logs call for next generation of big data mining technology providing effective and efficient tools for making use of the streaming data. Predictive analytics on data streams is actively studied in research communities and used in the real-world applications that in turn put in the spotlight several important challenges to be addressed. In this talk I will focus on the challenges of dealing with *evolving* data streams. In dynamically changing and nonstationary environments, the data distribution can change over time. When such changes can be anticipated and modeled explicitly, we can design context-aware predictive models. When such changes in underlying data distribution over time are unexpected, we deal with the so-called problem of *concept drift*. I will highlight some of the recent developments in the *proactive* handling of concept drift and link them to research in context-aware predictive modeling. I will also share some of the insights we gained through the performed case studies in the domains of web analytics, stress analytics, and food sales analytics.

A. Concept drift in evolving data streams

Predictive analytics as a research field studies how to extract useful knowledge from various data sources to induce different kinds of predictive models, which are ubiquitous for data-driven optimization, decision support, and decision-making. Predictive analytics includes, but is not limited to prediction, e.g. user intent [6] or food sales [15] prediction; matching, e.g. driver route recognition [11]; classification, e.g. antibiotic resistance prediction [13]; monitoring, e.g. switching of the operating regime [12] or stress detection [2,8]; forecasting and nowcasting, e.g. website traffic or solar radiation [16]; causal effect prediction and modeling what-if scenarios, e.g. predicting the effect of a marketing action [14].

In the past a typical predictive analytics R&D project life-cycle would assume the main involvement of a data mining expert for data modeling, model induction, and its evaluation and then at the moment of model deployment into operational settings, for live testing and final fine-tuning of the model. The current typical operational settings of data-driven decision making are much more dynamic. In many modern applications of predictive analytics, data evolve over time and must be analyzed in near real time. Patterns and relations in such data often evolve over time, thus, predictive models induced from such data may become obsolete over time unless we inject some adaptive learning mechanisms into predictive modeling. Predictions are also inherently context sensitive requiring adaptation of predicting to the current context. The number of contextual factors that may affect the behavior of the modeled concept can be enormous and hard to model explicitly, especially because some of such contexts may be unobservable to the predictive models. Unexpected changes in underlying data distribution over time are referred to as concept drift [5]. In data mining, machine learning, pattern recognition, and signal processing the phenomenon has been studied under other names, including but not limited to covariate shift, dataset shift, and nonstationarity. In supervised learning, concept drift refers to changes in the conditional distribution of the target variable given the input features, while the distribution of the input may stay unchanged. Changes in underlying data may occur due to different kinds of reasons; cf. changing personal interests vs. changes in population vs. adversary activities vs. a complex nature of the changing environment. A typical example of the concept drift is a change in user's interests when following an online news stream. Even if the distribution of the incoming news documents remains the same, the conditional distribution of the interestingness of news for that user changes.

B. Proactive handling of concept drift with adaptive learning

Adaptive learning refers to updating predictive models online during their operation to react to concept drifts. Over the last decade research related to learning with concept drift has been increasingly growing and many drift-aware adaptive learning algorithms have been developed. A comprehensive summary of the state-of-the-art methodologies and techniques for handling concept drift is available in the recent survey [5]. Two main strategies include evolving learning models continuously, e.g. retraining them on the most recent data window of some fixed size, and using a statistical change detection test as a trigger mechanisms to initiate a model update. Both single models and ensemble approaches are used to cope with concept drift. Unlike single models, ensembles, can maintain some memory of different concepts. The predictions can be casted by considering the output(s) of the most relevant base model(s) from the pool of existing models.

Most of the work on concept drift assumes that the changes happen in a hidden (not observable) context. Hence, concept

This research is partly supported by STW CAPA project

drift is considered to be unpredictable, and its detection and handling is mostly reactive. However, there are various application settings in which concept drift is expected to reappear along the time line and across different objects in the modeled domain inviting us for developing more proactive concept drift handling mechanisms [17]. Indeed, seasonal effects for some object(s) would be common e.g. in food demand prediction [15] or in driver route recognition [11]. Explicitly modeling recurrence or periodicity of changes improves the speed and accuracy of online change detection [10]. Availability of external contextual information or extraction of hidden (exceptional) contexts [9] from the predictive features may help to better handle recurrent concept drift, e.g. with use of a meta-learning approach [4]. Temporal relationships mining can be used to identify related drifts, e.g. in the distributed or peer-to-peer settings in which concept drift in one peer may precede another drift in related peer(s) [1].

C. Outlook: from accuracy to transparency and trust

Domain experts play an important role in acceptance of predictive analytics. They often need to trust underlying techniques and be certain that they are really going to react to changes when they happen. They need to understand how these changes are detected and what adaptation would happen. Therefore, besides reliable detection of changes and timely adaptation to them, it is important to develop techniques for change description, localization and visualization (cf. concept drift in process mining [3] as a motivating example) providing insights to domain experts in how and why changes happened thus improving utility, usability and trust in adaptive learning and context-aware systems being developed for many of the big data applications.. A deeper integration of context-aware predictive modeling with concept drift handling approaches is one promising direction to do that.

***Keywords**: predictive analytics; eloving data streams; concept drift; adaptation; context-awareness*

ACKNOWLEDGMENT

I would like to thank many of the current and former collaborators, in particular J. Bakker, A. Bifet, J. Gama, J. Kiseleva, A. Maslov, A. Tsymbal, and I. Žliobaitė.

BIOGRAPHY

MYKOLA PECHENIZKIY is Associate Professor in Predictive Analytics at the Department of Computer Science, Eindhoven University of Technology (TU/e), the Netherlands. He received his PhD from the Computer Science and Information Systems department at the University of Jyvaskyla, Finland in 2005. Since June 2013 he is also Adjunct Professor (Dosentti) in Data Mining for Industrial Applications at the Department of the Mathematical Information Technology, University of Jyväskylä. His expertise and research interests are in knowledge discovery and predictive analytics from evolving data, and in their application to real-world problems in industry, commerce, medicine and education. He develops generic frameworks and effective approaches for designing adaptive, context-aware predictive analytics systems. He has actively collaborated on this with industry. He has co-authored over 100 peer-reviewed publications and co-organized several workshops, conferences and tutorials in these areas. He has co-edited the Handbook of Educational Data Mining and served as a guest editor of the special issues with SIGKDD Explorations, Evolving Systems, Data and Knowledge Engineering and Artificial Intelligence in Medicine journals. He is the current President of IEDMS.

REFERENCES

[1] H.H. Ang, V. Gopalkrishnan, I. Žliobaitė, M. Pechenizkiy, S.C.H. Hoi, (2013) "Predictive Handling of Asynchronous Concept Drifts in Distributed Environments," IEEE Transactions on Knowledge and Data Engineering 25(10), pp. 2343-2355.

[2] J. Bakker, L. Holenderski, R. Kocielnik, M. Pechenizkiy, N. Sidorova, "Stess@work: From measuring stress to its understanding, prediction and handling with personalized coaching," in Proc. of the 2nd ACM SIGHIT International health informatics symposium, 2012, pp. 673-678.

[3] R.P.J.C. Bose, W.M.P. van der Aalst, I. Žliobaitė, M. Pechenizkiy, "Dealing with Concept Drifts in Process Mining". IEEE Transactions on Neural Networks and Learning Systems 25(1), pp. 154-171.

[4] J. Gama and P. Kosina, "Learning about the learning process," in Proc. of the 10th Int. Conf. on Advances in intelligent data analysis, IDA2011, pp. 162-172, 2011, Springer.

[5] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. (2014). "A Survey on Concept Drift Adaptation". ACM Computing Surveys 46(4), article 44.

[6] J. Kiseleva, H.T. Lam, M. Pechenizkiy, T. Calders, "Predicting current user intent with contextual markov models" ICDMW, Workshop Proc. of the IEEE 13th Int. Conf. on Data Mining, 2013, pp. 391-398.

[7] J. Kiseleva, E. Crestan, R. Brigo, R. Dittel, "Modelling and Detecting Changes in User Satisfaction," in Proc. of the 23rd ACM Int. Conf. on Information and Knowledge Management, 2014, pp. 1449-1458, ACM.

[8] H. Kurniawan, A. Maslov, M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in Proc. of 26th IEEE Int. Conf. on Computer-Based Medical Systems (CBMS), 2013, pp. 209-214.

[9] J.M. Luna, M. Pechenizkiy, S. Ventura, (2015) "Mining Exceptional Relationships with Grammar-Guided Genetic Programming", Knowledge and Information Systems, in press.

[10] A. Maslov, M. Pechenizkiy, T. Karkkainen, and I. Žliobaitė, "How to improve the speed and accuracy of online change detection by modeling recurrent events," unpublished.

[11] O. Mazhelis, I. Žliobaitė, M. Pechenizkiy, "Context-aware personal route recognition," in Proc. of 14th Int. Conf. on Discovery Science, 2011, pp. 221-235, Springer.

[12] M. Pechenizkiy, J. Bakker, I. Žliobaitė, A. Ivannikov, T. Kärkkäinen, "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift," ACM SIGKDD Explorations 11 (2), pp. 109-116.

[13] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, (2008) "Dynamic integration of classifiers for handling concept drift," Information Fusion, 9(1), pp. 56-68.

[14] I. Žliobaitė, M. Pechenizkiy, "Learning with Actionable Attributes: Attention--Boundary Cases!" ICDMW, Workshop Proc. of the IEEE 10th Int. Conf. on Data Mining, pp. 1021-1028, IEEE.

[15] I. Žliobaitė, J. Bakker, M. Pechenizkiy, (2012) Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? Expert Systems with Applications 39 (1), pp. 806-815.

[16] I. Žliobaitė, J. Hollmén, H. Junninen, (2014). "Regression models tolerant to massively missing data: a case study in solar radiation nowcasting". Atmospheric Measurement Techniques 7, 4387-4399.

[17] I. Žliobaitė, M. Pechenizkiy, J. Gama, "An overview of concept drift applications," in Big Data Analysis: New Algorithms for a New Society, Springer, in press.