On Optimising Sample Selection in Credit Scoring with Active Learning

Georg Krempl[1] and Daniel Kottke[2]

[1] Knowledge Management & Discovery
Otto-von-Guericke University Magdeburg, Germany
georg.krempl@iti.cs.uni-magdeburg.de

[2] IES Group, Faculty of Computer Science,
University of Kassel, Germany
daniel.kottke@uni-kassel.de

Abstract. Constructing classifiers for credit scoring requires labelled training data, i.e. data of applicants with known true good or bad creditworthiness status. Ideally, such a training sample constitutes an unbiased through-the-door sample. Nevertheless, obtaining such a sample is costly, as it involves accepting customers who are associated with a risky score. While these risky customers would have been rejected otherwise, they are accepted for the purpose of constructing a classifier model. In order to avoid this, reject inference procedures have been proposed, which aim to infer the true status of rejected customers from data obtained on the accepted ones. However, for obtaining valid reject inference, literature (e.g., [Hand and Henley, 2004]) suggests to obtain a so-called calibration sample, which requires a subsample of the instances in the costly reject regions. A related problem in machine learning is selective sampling [Settles, 2012], where costly labels are solely queried for of the most informative instances. Various approaches have been proposed for this active learning problem, and investigating their use in credit scoring has been identified as a promising direction of further research recently (eg., in [Crone and Finlay, 2012]). Popular active learning approaches are uncertainty sampling and Query-by-Committee, although they have sometimes shown unreliable performance (see e.g. [Attenberg and Provost, 2011], and [Evans et al., 2013]). More recently, a probabilistic active learning approach [Krempl et al., 2015] has been proposed with reliable performance with several classifier technologies [Beyer et al., 2015].

Therefore, in this work we propose to obtain a calibration subsample of instances in the costly reject regions by selective sampling techniques. In particular, we study the use of probabilistic active machine learning techniques for optimising the selection of such a costly calibration sample. In an experimental evaluation, we compare different active selection techniques to a randomly selection strategy, and evaluate their impact on the performance of the obtained classifier.

References

[Attenberg and Provost, 2011] Attenberg, J. and Provost, F. (2011). Online active inference and labelling. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2011)*.

[Beyer et al., 2015] Beyer, C., Krempl, G., and Lemaire, V. (2015). How to select information that matters: A comparative study on active learning strategies for classification. In *Proc. of the 15^{th} Int. Conf. on Knowledge Technologies and Data-Driven Business (i-KNOW 2015)*, pages 2:1–2:8. ACM.

[Crone and Finlay, 2012] Crone, S. F. and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1):224–238.

[Evans et al., 2013] Evans, L. P., Adams, N. M., and Anagnostopoulos, C. (2013). When does active learning work? In Tucker, A., Höppner, F., Siebes, A., and Swift, S., editors, *Advances in Intelligent Data Analysis XII 12^{th} International Symposium, IDA 2013, London, UK, October 2013*, volume 8207 of *Lecture Notes in Computer Science*, pages 174–185. Springer.

[Hand and Henley, 2004] Hand, D. J. and Henley, W. E. (2004). Can reject inference ever work? In Thomas, L. C., Edelman, D. B., and Crook, J., editors, *Readings in Credit Scoring: Foundations, Developments, and Aims*, Oxford Finance Series. Oxford University Press.

[Krempl et al., 2015] Krempl, G., Kottke, D., and Lemaire, V. (2015). Optimised probabilistic active learning (OPAL) for fast, non-myopic, cost-sensitive active classification. *Machine Learning*, 100(2).

[Settles, 2012] Settles, B. (2012). *Active Learning*. Number 18 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.