Technology
Arts Sciences
TH Köln

Approximating the optimal threshold for an abstaining classifier based on a reward function with regression

Bachelor Thesis

Jonas Fassbender jonas@fassbender.dev
11117674

In the course of studies
Computer Science

For the degree of
Bachelor of Science

Technical University of Cologne
Faculty of Computer Science and Engineering

First supervisor:	Prof. Dr. Heinrich Klocke Technical University of Cologne
Second supervisor:	Prof. Dr. Fotios Giannakopoulos Technical University of Cologne

Overath, July 2019

1. Introduction

An abstaining classifier (see e.g. Vanderlooy et al., 2009)—also called a classifier with reject option (see e.g. Fischer et al., 2016)—is a kind of confidence predictor. It can refuse from making a prediction if its confidence in the prediction is not high enough. High enough, in this context, means that the confidence is greater than a certain—hopefully optimal—threshold. Optimality is dependent on a performance metric set beforehand.

This thesis introduces a new kind of method for approximating the optimal threshold based on a reward function—better known from reinforcement learning than from the supervised learning setting (see e.g. Sutton and Barto, 2018). The method treats the reward function as unknown, making it a very general approach and giving quite the amount of freedom in designing the reward function.

In supervised learning the concept that is closest to a reward function is a cost function and many abstract types of cost in supervised learning are known (see Turney, 2002).

Probably today's most used methods for obtaining the optimal threshold for reducing the expected cost of an abstaining classifier are based on the receiver operating characteristic (ROC) rule (see Tortorella, 2000; Pietraszek, 2005; Vanderlooy et al., 2009; Guan et al., 2018).

The method presented in this thesis is more flexible than the methods based on the ROC rule and can—depending on the context of the classification problem—produce results better interpretable (see Chapter 2). Also it is more natural with multi-class classification problems than the methods based on the ROC rule, all assuming binary classification problems, wherefore the classifiers generated by these methods must be transformed to multi-class classifiers for non-binary problems.

On the other hand the presented method can suffer from its very general approach and only produces approximations. This can result in non-optimal and unstable thresholds.

This thesis first presents a motivational example. In Chapter 3 the proposed method is presented. After that experiments on data sets from the UCI machine learning repository are discussed (see Dua and Graff, 2017). At last further research ideas are listed and a conclusion is drawn.

2. Motivational example

Abstaining classifiers—compared to typical classifiers, which classify every prediction—can be easily integrated into processes where they partially replace the decision making, since they can delegate the abstained predictions back to the underlying process. This is a valuable property if there does not exist a typical classifier good enough to fully replace the underlying process.

Many real world application domains for abstaining classifiers can express a cost function associated to the decisions about predicting and abstaining of the classifier—which then chooses the threshold with which it produces the least amount of cost, therefore minimizing the cost.

For example, the real world application domain could be a facial recognition system at a company which regulates which employee can enter a trust zone and which can not. In this example, the cost for miss-classifying an unauthorized person as authorized can have huge costs for the company while abstaining or classifying an authorized employee as unauthorized produces quite low costs—the authorized employee just has to start the manual process, which should be replaced by the facial recognition system.

On the other hand, for some real world application domains a reward function based on which the abstaining classifier chooses the threshold by maximizing the reward—rather than minimizing the cost—comes more natural.

Such a domain would be the finance industry, where we often can associate a certain amount of money an abstaining classifier can produce or safe by supporting the decision making of an underlying process.

An example for such a process would be a bank which wants to grand a consumer credit. The bank requests information about the consumer from a credit bureau in order to assess the consumer's credit default risk. Now the bank wants to predict the consumer's credit default risk based on information the bank has about the consumer. If the credit default risk is very high or very low the bank can save money not making a request to the credit bureau for this consumer. The optimal threshold for the abstaining classifier making the prediction about the credit default risk can easily be expressed by a reward function. Every correct decision saves the bank the money the request to the credit bureau costs. Every miss-classification costs the bank either the amount of money it would gain by granting the credit, or the money it loses by giving a credit to somebody that does not pay the rates. Abstention cost is the cost of making a request to the credit bureau.

Using a reward function—like in the example above—instead of a cost function has an advantage in readability. One can easily assess the gain of introducing the abstaining classifier to the process. Is the reward generated by the abstaining classifier higher than zero, the process is enhanced by the abstaining classifier. Otherwise the abstaining classifier would produce more cost than it would save and it is not valuable for the bank to introduce it to its process of assessing a consumer's credit default risk.

3. Proposed method based on a reward function

4. Experiments

5. Further research

6. Conclusion

Appendix

A. Plots

References

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Lydia Fischer, Barbara Hammer, and Heiko Wersing. Optimal local rejection for classifiers. *Neurocomputing*, 214:445 – 457, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2016.06.038. URL http://www.sciencedirect.com/science/article/pii/S0925231216306762.

Hongjiao Guan, Yingtao Zhang, Heng-Da Cheng, and Xianglong Tang. Abstaining classification when error costs are unequal and unknown. *CoRR*, abs/1806.03445, 2018. URL http://arxiv.org/abs/1806.03445.

Tadeusz Pietraszek. Optimizing abstaining classifiers using roc analysis. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 665–672, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102435. URL http://doi.acm.org/10.1145/1102351.1102435.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 2nd edition, 2018.

Francesco Tortorella. An optimal reject rule for binary classifiers. pages 611–620, 08 2000. doi: 10.1007/3-540-44522-6_63.

Peter D. Turney. Types of cost in inductive concept learning. *CoRR*, cs.LG/0212034, 2002. URL http://arxiv.org/abs/cs.LG/0212034.

Stijn Vanderlooy, Ida G. Sprinkhuizen-Kuyper, Evgueni N. Smirnov, and H. Jaap van den Herik. The roc isometrics approach to construct reliable classifiers. *Intell. Data Anal.*, 13(1):3–37, January 2009. ISSN 1088-467X. URL http://dl.acm.org/citation.cfm?id=1551758.1551760.

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum	Rechtsverbindliche Unterschrift