Confidence-Based Ensemble Modeling in Medical Data Mining

Lukas Kammerer
lukas.kammerer@fh-hagenberg.at
Heuristic and Evolutionary Algorithms Laboratory
University of Applied Sciences Upper Austria
Hagenberg, Austria
Institute for Formal Models and Verification
Johannes Kepler University
Linz, Austria

Michael Affenzeller
michael.affenzeller@fh-hagenberg.at
Heuristic and Evolutionary Algorithms Laboratory
University of Applied Sciences Upper Austria
Hagenberg, Austria
Institute for Formal Models and Verification
Johannes Kepler University
Linz, Austria

ABSTRACT

A recent approach for improving the accuracy of ensemble models is confidence-based modeling. Thereby, confidence measures, which indicate an ensemble prediction's reliability, are used for identifying unreliable predictions in order to improve a model's accuracy among reliable predictions. However, despite promising results in previous work, no comparable results for public benchmark data sets have been published yet.

This paper applies confidence-based modeling with GP-based symbolic binary classification ensembles on a set of medical benchmark problems to make statements about the concept's general applicability. Moreover, extensions for multiclass classification problems are proposed.

CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods**; *Genetic programming*; Modeling methodologies; • **Information systems** → Data mining;

KEYWORDS

Machine Learning, Genetic Programming, Confidence-based Ensemble Modeling, Medical Data Mining

ACM Reference Format:
Lukas Kammerer and Michael Affenzeller. 2018. Confidence-Based Ensemble Modeling in Medical Data Mining. In *GECCO '18 Companion: Genetic and Evolutionary Computation Conference Companion, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3205651.3205722

1 INTRODUCTION

The application of machine learning models on new data means in practice that some non-identifiable, wrong predictions are made. In most domains, these random errors are simply accepted as noise. However, this might not be an option in very critical domains like medicine when even no prediction at all is favored over an unreliable one, for example when supporting physicians in making diagnoses.

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3205722

1.1 Confidence-Based Ensemble Modeling

Such scenarios require additional information by the machine learning model, if the model is able to make a reliable prediction or if it cannot provide meaningful information. A measure for such an estimation of predictions' reliability in binary classification problems was introduced by Affenzeller et al. [2] for ensembles of GP-based symbolic classification models. They proposed *confidence measures* which indicate whether an ensemble's prediction is either trustworthy and correct or unreliable and error-prone. Their measures describe the clearness among the ensemble members' predictions – equal predictions from many different members result in high confidence values, ambiguity indicates low confidence values.

The basic idea of confidence was already used before in several different ways, for example to improve the generalization error in *boosting* [8] or to estimate the probability of a class estimation in random forest [4]. Affenzeller et al. [3] used their proposed confidence measures to reject unreliable predictions based on their confidence values and a predefined confidence threshold. If the confidence values is below this threshold, no output at all is made for an instance. Such confidence-based modeling improved the ensemble's average accuracy among remaining instances and therefore its trustworthiness for users of different domains like physicians.

1.2 Goals

While confidence-based modeling showed very promising results in the cited previous work, it focused on a few non-public binary classification problems and did not make statements about the proposed methods' general applicability. It was also noted that ensemble techniques and genetic programming are still rarely combined although they are a reasonable match [6] due to GP's stochasticity and the resulting, necessary diversity among ensemble members [5].

The research goal of this work is to show the general applicability of the previously proposed methodology: the basic accuracy in comparison with results in literature, the distribution of confidence values among correct and wrong predictions and the trade-off between accuracy gain and loss in the number of instances for which a prediction is still made.

This work applies confidence-based modeling with GP-based classification ensembles on well-known medical benchmark problems from the UCI machine learning repository: the *Cleveland Heart Disease*, *Pima Indians Diabetes*, *Hepatitis*, *SpectF Heart*, *Ljubljana Breast Cancer*, *Audiology*, *Lymphography* and *Primary Tumor* data set. Additionally, extensions for the application in multiclass classification problems are proposed.

2 EXPERIMENTS

The experiments are set up as the ones by Affenzeller et al. [2, 3], due to the similar domain. The ensembles consist of symbolic classification models as described by Winkler [9]. A single symbolic classification model is trained using GP with strict offspring selection (OSGP) and gender-specific selection [1]. This combination has been shown to be robust to non-optimal algorithm settings. The experiments were conducted with the *HeuristicLab*[1] framework.

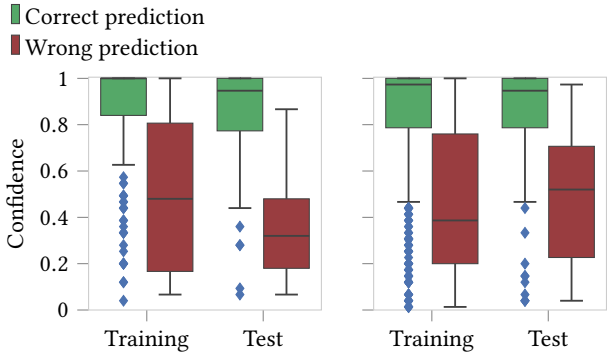
For each ensemble in binary classification tasks, 900 symbolic classification models are trained. Each ensemble is pruned to contain only the best 75 models regarding training performance. Since multiclass problems with nominal classes are not handled well in the used models, they are modeled with the One vs. All binarization scheme [7] in which one (binary) ensemble is trained for each class. Each such ensemble predicts, whether an instance belongs to its class or not. The final output is the class with the most votes from its assigned ensemble. Therefore, the confidence measures were extended to cover ambiguities, when multiple ensembles provide a similar number of votes for their class.

3 RESULTS

The conducted experiments show similarly promising results as in the cited work [2, 3]. Those results could be fully reproduced on most data sets: Starting from the general test accuracy of the GP-based symbolic classification ensembles to the distribution of confidence values among correct and wrong predictions and therefore the accuracy gains in confidence-based modeling. A drawback of the proposed methodology is the high computational effort due to the high number of generated models and the computationally expensive training algorithms.

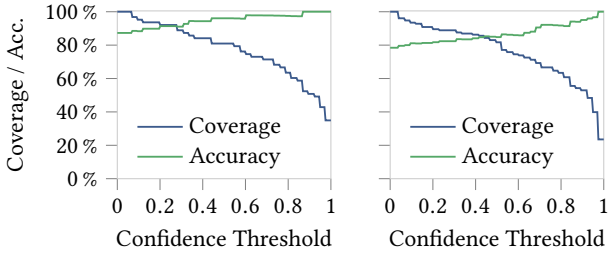
However, without any parameter tuning, the ensembles achieve for most of the data sets accuracy values which are comparable to the best ones found in literature. Also a strong connection between a prediction's correctness and its confidence value can be seen in most data sets. Figure 1 shows this relation for two exemplary data sets and highlights the interquartile range of the confidence value distributions. The hypothesis, that correct predictions are attended by higher confidence values than wrong ones, could be confirmed for all the given data sets.

[1] *https://heuristiclab.com*



(a) Cleveland Heart Disease **(b) Pima Indians Diabetes**

Figure 1: Confidence distribution among correct and incorrect predictions.



(a) Cleveland Heart Disease **(b) Pima Indians Diabetes**

Figure 2: Progression of coverage loss and accuracy gain with continually increasing confidence threshold.

This again leads to the beneficial effect of confidence-based modeling that can be also observed in the experiments. Confidence-based modeling is able to further improve the prediction quality notably with still reasonable coverage of given data in nearly all cases. Figure 2 illustrates the trade-off between instance coverage and accuracy with continually increasing confidence threshold. These accuracy improvements were achieved both with the existing confidence measures in binary classification problems, as well as with the proposed measures in multiclass problems. Notable is also the accuracy gain up to perfect test accuracy in most data sets. The theoretical possibility of perfect accuracy for unseen data is especially interesting for domains in which machine learning techniques cannot be applied currently due to high criticality of predictions – given that making no prediction for instances is allowed.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support by the Austrian Research Promotion Agency (FFG) within the COMET Project Heuristic Optimization in Production and Logistics (HOPL), #843532.

REFERENCES

[1] Michael Affenzeller, Stefan Wagner, Stephan Winkler, and Andreas Beham. 2009. *Genetic algorithms and genetic programming: modern concepts and practical applications.* CRC Press.
[2] Michael Affenzeller, Stephan Winkler, Stefan Forstenlechner, Gabriel Kronberger, Michael Kommenda, Stefan Wagner, and Herbert Stekel. 2012. Enhanced confidence interpretations of gp-based ensemble modeling results. In *Proceedings of the 24th European Modeling & Simulation Symposium.* 340–345.
[3] Michael Affenzeller, Stephan Winkler, Herbert Stekel, Stefan Forstenlechner, and Stefan Wagner. 2013. Improving the Accuracy of Cancer Prediction by Ensemble Confidence Evaluation.. In *EUROCAST (1).* 316–323.
[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[5] Thomas G Dietterich. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning* 40, 2 (2000), 139–157.
[6] Maarten Keijzer and Vladan Babovic. 2000. Genetic programming, ensemble methods and the bias/variance tradeoff-introductory investigations. In *EuroGP.* Springer, 76–90.
[7] Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *Journal of machine learning research* 5, Jan (2004), 101–141.
[8] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics* (1998), 1651–1686.
[9] Stephan M Winkler. 2008. *Evolutionary system identification: modern concepts and practical applications.* Trauner.