

Well-Calibrated Rule Extractors

COPA 2022

Ulf Johansson, Tuwe Lofström, Niclas Ståhl

Jönköping AI Lab
Jönköping University
ulf.johansson@ju.se

August, 2022



Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Introduction

- Predictive models are often required to provide **explanations** for predictions produced.
- Most research about explanation methods focuses on creating **local** explanations, i.e., explanations for single predictions, e.g., LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017)
- In many scenarios, where a more **global understanding** of the model and the underlying relationship is needed, these single prediction explanations will not suffice.



Introduction

- **Rule extraction** is a technique for approximating global models with interpretable models, e.g., decision trees or rule sets.
- Such interpretable approximations of the opaque model enable global explanations, allowing for inspection and analysis.
- Given an extracted interpretable model, it is straightforward to obtain detailed explanations for single predictions, i.e., local explanations.
- Depending on the exact situation, the extracted model may either be used to make the actual predictions, or simply to explain the predictions made by the opaque model.



Introduction

- In rule extraction, **fidelity** measures the extent to which an extracted model makes the same predictions as the opaque model.
- For classification, this simply means the proportion of instances where the opaque and transparent models agree.
- A low-fidelity transparent model will be misleading, producing predictions that differ substantially from the opaque model.
- Rule extraction techniques are designed to somehow optimize fidelity, but there are **no guarantees** that fidelity on training data will carry over to new unseen data.



Contribution

- In this paper we introduce and evaluate **rule extractors with well-calibrated fitness estimations**.
- In the specific setup suggested, **Venn-Abers** are used for calibrating extracted decision trees resulting in what we call **fitness estimation trees (FETs)**.
- The result is a very informative model where each leaf in the tree contains a **well-calibrated fidelity estimation interval**.



Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Rule extraction

- **Pedagogical** or **black-box** rule extraction employs a machine learning technique (that produces transparent models) to learn the input-output relationship of the opaque model.
 - It uses the original input patterns together with the predictions from the opaque model as targets.
 - It is model agnostic, in the sense that it may be used on any type of opaque model.
- **Open-box** rule extraction produces a transparent model based on the inner workings of the opaque model.
 - These techniques use e.g., the architecture of the opaque model, and are thus tailored to a specific type of models, most often a neural network.



Rule extraction

- Pedagogical rule extraction results in transparent models that approximate the opaque model, similarly to the way a model approximates a data set in inductive learning.
- Open-box techniques will produce exact, but possibly very complex, transparent representations.
- Thus, pedagogical rule extraction has the distinct advantage of being model agnostic, but provides no fidelity guarantees.
- Open-box methods, often per design obtains perfect fidelity, but are restricted to a certain type of opaque models.



Rule extraction

- All pedagogical rule extraction techniques somehow optimize fidelity, but similar to inductive models generated to optimize predictive performance, there are no guarantees that fidelity on training data will carry over to new unseen data
- A single measure of model fidelity on a test set only indicates the **average** infidelity rate, but does not give any indication of whether a particular instance can be expected to be predicted identically to the opaque model or not.
- We want to add **well-calibrated fitness estimates** on the **instance level** to pedagogical rule extraction.
- The key idea is to regard the extracted model as a **probabilistic predictor**, but remembering that the estimates are for **fidelity**, not accuracy.



Probabilistic prediction

- A probabilistic predictor outputs both the predicted class label and a **probability distribution** over the labels.
- Calibration: The estimate should be close to the true probability.

$$p(c_j | p^{c_j}) = p^{c_j}, \quad (1)$$

where p^{c_j} is the probability estimate for class j .

- For rule extraction, we want **fitness** estimates to be well-calibrated.
- Since the interpretable model is trained using the opaque models predictions as targets, the probability estimates represent fidelity estimates.



Venn-Abers predictors

- **Venn predictors** are multi-probabilistic predictors with proven validity properties (Vovk, Gammerman, and Shafer 2005).
- **Venn-Abers predictors** (Vovk and Petej 2012) operates on scoring classifiers, i.e, they are restricted to two-class problems.
- Since Venn-Abers predictors are Venn predictors, they inherit the validity guarantees.
- Venn-Abers predictors use isotonic regression (Zadrozny and Elkan 2001) for the fitting.



Venn-Abers predictors

A multiprobabilistic prediction from an inductive Venn-Abers predictor is produced as follows:

- 1 Let $\{z_1, \dots, z_{l+q}\}$ be a training set where each instance $z_i = (x_i, y_i)$ consists of two parts; an *object* x_i and a *label* y_i .
- 2 Let the training set be divided into a proper training set Z_T with q instances and a calibration set $\{z_1, \dots, z_l\}$.
- 3 Train a scoring classifier using the proper training set Z_T to produce the prediction scores s_0 for $\{z_1, \dots, z_l, (x_{l+1}, 0)\}$ and s_1 for $\{z_1, \dots, z_l, (x_{l+1}, 1)\}$.
- 4 Let g_0 be the isotonic calibrator for $\{(s_0(x_1), y_1), \dots, (s_0(x_l), y_l), (s_0(x_{l+1}), 0))\}$ and g_1 be the isotonic calibrator for $\{(s_1(x_1), y_1), \dots, (s_1(x_l), y_l), (s_1(x_{l+1}), 1))\}$.
- 5 Let the probability interval for $y_{l+1} = 1$ be $[g_0(s_0(x_{l+1})), g_1(s_1(x_{l+1}))]$.

Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Drug discovery case

- A **DNN** with 5 hidden layers and a total of 91 456 free parameters is trained to predict the inhibition of the **Cytochrome P450 2C19 enzyme**.
- When trained and evaluated, the DNN achieves an accuracy of **76.4 %**.
- The data set consists of 12 665 instances (molecules) represented by 10 commonly used and human-understandable descriptors.

Feature name	Feature description
Weight	Molecular weight in Dalton.
2*LogP	Partition Coefficient, which describes how easily each molecule is dissolved in water.
HDonors	Number of hydrogen donors.
HAcceptors	Number of hydrogen acceptors.
AromaticRings	Number of aromatic rings.
2*TPSA	The topological polar surface area, which is the surface sum over all polar parts of the molecule.
RotatableBonds	Number of bonds which allow free rotation around themselves.
HeavyAtomCount	Number of non-hydrogen atoms.
FractionCSP3	The fraction of C atoms that are SP3 hybridized.
RingCount	Number of rings.



Method - benchmark data sets

- Single- and multi-layer MLPs were used as opaque models
- Since Venn-Abers needs a separate labeled data set for the calibration, two different MLPs were trained; one using all training instances and one dividing the training instances into a proper training set (2/3) and a calibration set (1/3).

The setups:

- **ANNa**: MLPs trained using all training data.
- **ANNt**: MLPs trained using 2/3 of the training data.
- **Uncal**: Pedagogic rule extraction using decision trees.
- **VA**: Pedagogic rule extraction using decision trees and Venn-Abers calibration.



Experimental details - benchmark data sets

- All experimentation was carried out using scikit learn, keras and tensorflow.
- The activation functions in the hidden and output layers were ReLU and sigmoid, respectively.
- The number of hidden units h was chosen as $h = \lfloor \frac{2}{3}a \rfloor$ where a is the number of attributes.
- The loss function was set to cross entropy, and Adam was used as the optimizer.
- Standard decision trees were used as rule extractors. All parameter values were left at default, with the exception that the minimum number of training instances in each leaf was set to 5.
- For the actual evaluation, 10x10-fold cross validation was used, so all results are averaged over the 100 folds.



Evaluation metrics - benchmark data sets

- Accuracy and area under the ROC-curve (AUC) are used to measure the predictive performance.
- Calibration quality is evaluated using
 - log loss
 - Brier loss
 - expected calibration error (ECE)
- When comparing Venn-Abers calibrations to other techniques, the output probability intervals (p_0, p_1) , must be aggregated into a single probability estimate.
- In this study, we use a regularized value:

$$p = \frac{p_1}{1 - p_0 + p_1}$$

(2)


Benchmark data sets

In the benchmark experiments, 25 publicly available data sets are used.

Data set	#inst	#attrib	Source	Data set	#inst	#attrib	Source
colic	328	23	UCI	kc2	522	22	Promise
creditA	690	16	UCI	kc3	325	39	Promise
diabetes	768	9	UCI	liver	345	7	UCI
german	1000	21	UCI	pc1req	320	9	Promise
haberman	306	4	UCI	pc4	1458	38	Promise
heartC	303	13	UCI	sonar	208	61	UCI
heartH	270	12	UCI	spect	218	22	UCI
heartS	270	14	UCI	spectf	348	45	UCI
hepati	155	20	UCI	transfusion	748	5	UCI
iono	351	35	UCI	ttt	958	10	UCI
je4042	274	9	Promise	vote	435	17	UCI
je4243	363	8	Promise	wbc	699	10	UCI
kc1	2109	22	Promise				

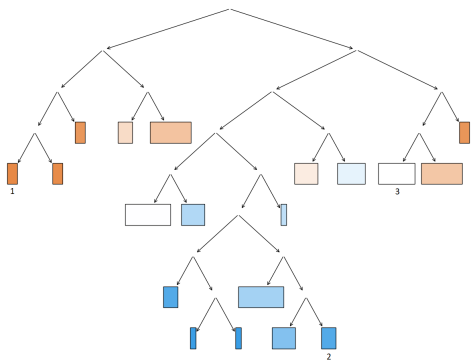


Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Drug discovery case

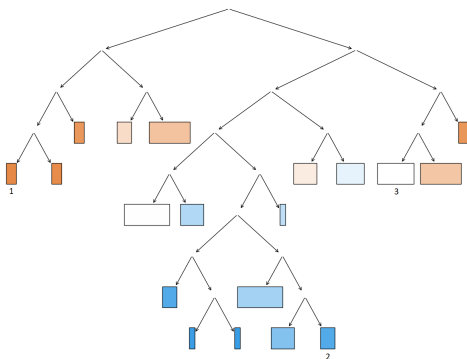


The FET conveys three things:

- The class predicted by the opaque model in different parts of the input space, represented by the colors blue and orange.
- The fidelity to the opaque model in different parts of the input space, represented by the color intensity of the leaves.
- How certain it is about its own fidelity estimation, represented by the width of the leaves.



Drug discovery case



Global explanations:

- The left part of the FET, indicates that the DNN will predict NO CYP2C19 inhibition
- In the middle part, the DNN most often predicts CYP2C19 inhibition
- In the right part, the FET shows that it is uncertain about the DNN predictions



Drug discovery case

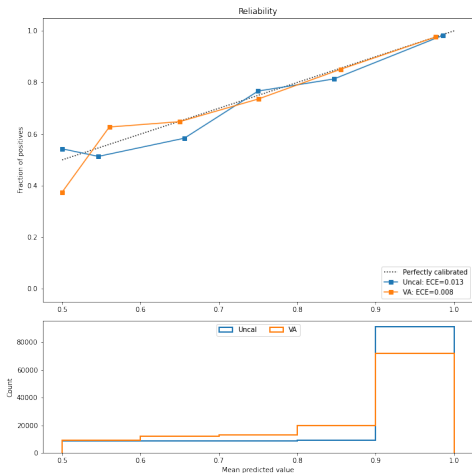
- 1) $\text{LogP} \leq 1.2$
→ **No CYP2C19 inhibition** [0.961, 1.0]
- 2) $\text{LogP} > 4.1$
& $\text{FractionCSP3} \leq 0.34$
& $\text{AromaticRings} \leq 6$
& $\text{RingCount} \leq 3$
& $\text{RotatableBonds} \geq 5$
→ **CYP2C19 inhibition** [0.929, 0.995]
- 3) $\text{LogP} > 2.5$
& $0.47 < \text{FractionCSP3} \leq 0.56$
→ **Indecisive** [0.451, 0.656]

Each leaf is a local explanation:

- ① If the logP is low, then the underlying model will predict that there will be no CYP2C19 inhibition.
- ② Rule indicating that the DNN will predict CYP2C19 inhibition.
- ③ A region where the FET is uncertain about the prediction of the underlying model. This does not mean that the DNN will fail for inputs in this region, just that no easily obtained explanations exist.



Drug discovery case



- The extracted FET is rather well-calibrated, but slightly overconfident
- Venn-Abers lowers the extreme estimates from the FET, resulting in better calibration
- For this data set, we would know that the fidelity estimates are very good, i.e., the FET could be used to explain the DNN

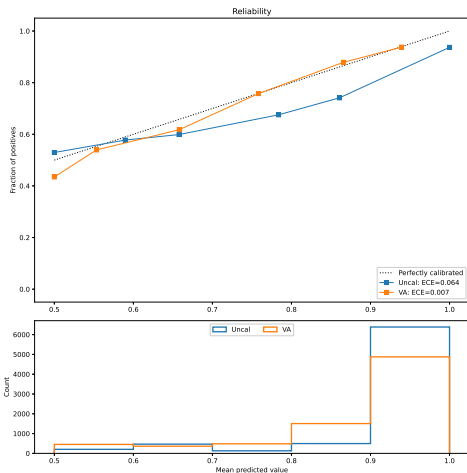


Predictive performance - benchmark data sets

	Accuracy				Fidelity		Size	
	ANNa	ANNt	Uncal	VA	Uncal	VA	Uncal	VA
colic	.804	.789	.779	.804	.837	.824	49.2	33.1
creditA	.850	.848	.842	.844	.882	.874	65.8	46.1
diabetes	.765	.760	.748	.747	.886	.874	66.7	47.7
german	.649	.650	.647	.671	.823	.827	149.4	106.5
haberman	.713	.719	.714	.720	.981	.983	6.4	4.0
heartC	.819	.815	.780	.777	.857	.829	38.1	27.1
heartH	.828	.829	.784	.774	.866	.865	32.1	22.9
heartS	.832	.828	.774	.779	.851	.841	31.7	22.2
hepati	.848	.843	.783	.800	.835	.859	17.4	11.8
iono	.917	.915	.871	.872	.857	.870	28.7	20.4
je4042	.714	.711	.719	.702	.900	.894	25.8	16.3
je4243	.626	.625	.618	.612	.912	.885	32.7	24.6
kc1	.762	.759	.753	.750	.935	.931	55.0	39.6
kc2	.793	.791	.797	.793	.942	.937	16.7	11.9
kc3	.871	.867	.874	.870	.942	.948	18.4	12.8
liver	.686	.640	.610	.597	.772	.793	53.3	34.1
pc1req	.683	.654	.691	.639	.853	.822	16.7	12.0
pc4	.904	.902	.879	.880	.908	.919	87.6	60.2
sonar	.841	.816	.717	.697	.715	.733	29.8	19.5
spect	.883	.881	.865	.884	.948	.975	19.1	9.9
spectf	.791	.788	.749	.782	.803	.829	33.3	22.3
transfusion	.749	.752	.746	.745	.975	.974	14.0	9.8
ttt	.981	.960	.913	.909	.912	.903	84.9	68.5
vote	.860	.856	.862	.846	.910	.902	53.2	36.7
wbc	.971	.970	.954	.952	.971	.968	19.9	15.4
Mean	.806	.799	.779	.778	.883	.882	41.8	29.4
Mean rank	1.16	1.84	1.40	1.60	1.40	1.60	2.00	1.00

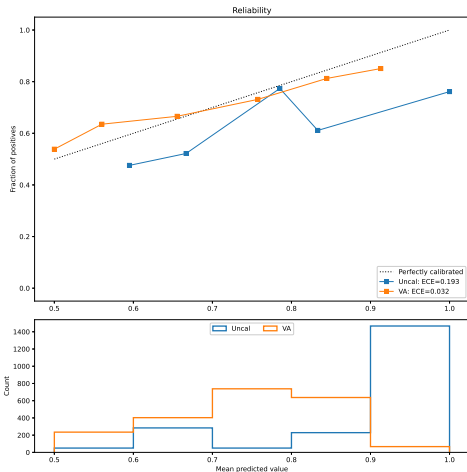


FET - Diabetes



- The uncalibrated FETs are often overconfident
- Here, the poorly calibrated FET is significantly improved by Venn-Abers: The ECE goes from 0.06 to 0.01
 - The uncalibrated FET has many estimates close to 1.0
 - With Venn-Abers, the estimated fidelities are often lower
 - (The overall fidelity is approximately 0.87)

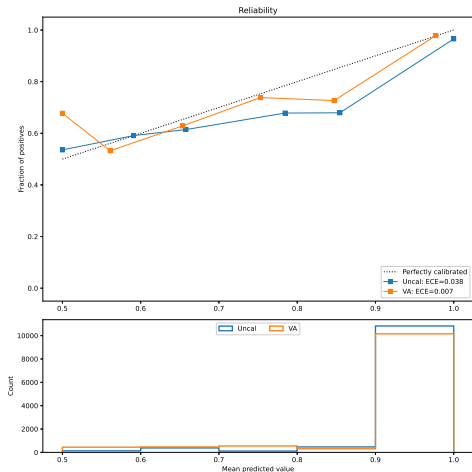




- This is one of the most extreme data sets, where the uncalibrated model is very overconfident
- Venn-Abers is very successful, although even the calibrated model is still slightly overconfident



FET - KC1



- An already rather well-calibrated FET is slightly improved by Venn-Abers



Calibration

	Difference		ECE		Log loss		Brier loss	
	Uncal	VA	Uncal	VA	Uncal	VA	Uncal	VA
colic	.079	-.013	.096	.030	8.97	.427	.322	.132
creditA	.063	-.004	.068	.014	15.84	.336	.514	.098
diabetes	.063	.002	.064	.007	21.10	.330	.660	.096
german	.075	.002	.075	.011	2.48	.404	.149	.126
haberman	.008	-.012	.009	.015	32.48	.058	.951	.013
heartC	.073	.003	.081	.026	14.99	.395	.504	.122
heartH	.051	-.014	.056	.031	17.63	.332	.600	.098
heartS	.081	-.010	.086	.029	16.39	.392	.536	.119
hepati	.089	-.005	.089	.019	23.31	.351	.752	.105
iono	.096	-.009	.097	.016	9.14	.339	.292	.101
je4042	.051	-.017	.053	.024	16.26	.284	.520	.081
je4243	.042	-.007	.046	.017	12.16	.295	.396	.085
kc1	.038	.004	.038	.007	29.39	.183	.880	.051
kc2	.032	-.015	.033	.020	26.67	.182	.800	.048
kc3	.025	-.013	.030	.019	29.31	.137	.899	.038
liver	.127	.005	.128	.013	6.43	.455	.272	.147
pc1req	.050	-.030	.061	.038	12.18	.409	.460	.128
pc4	.054	-.002	.057	.005	27.91	.224	.853	.062
sonar	.193	.010	.193	.032	10.90	.570	.385	.192
spect	.000	-.024	.025	.027	.32	.098	.030	.022
spectf	.119	-.010	.119	.013	3.91	.411	.162	.127
transfusion	.013	-.008	.013	.008	31.99	.083	.940	.021
ttt	.031	-.002	.032	.006	9.84	.257	.322	.073
vote	.036	-.003	.036	.006	21.71	.270	.699	.077
wbc	.011	-.010	.012	.013	21.37	.103	.635	.025
Mean	.060	-.007	.064	.018	16.91	.293	.541	.087
Mean rank			1.88	1.12	2.00	1.00	2.00	1.00



Fidelity estimation intervals for Venn-Abers

Venn-Abers fidelity estimates and corresponding empirical fidelity values

	VA fid. est.		Fid.		VA fid. est		Fid.
	Low	High	Emp.		Low	High	Emp.
colic	.799	.845	.824	kc2	.916	.949	.937
creditA	.863	.890	.874	kc3	.925	.963	.948
diabetes	.870	.893	.874	liver	.785	.829	.793
german	.822	.845	.827	pc1req	.764	.876	.822
haberman	.968	.990	.983	pc4	.914	.928	.919
heartC	.820	.871	.829	sonar	.726	.783	.733
heartH	.837	.893	.865	spect	.944	.982	.975
heartS	.818	.874	.841	spectf	.806	.852	.829
hepati	.835	.907	.859	transfusion	.962	.980	.974
iono	.851	.891	.870	ttt	.896	.916	.903
je4042	.866	.917	.894	vote	.891	.925	.902
je4243	.869	.910	.885	wbc	.955	.976	.968
kc1	.931	.945	.931	Mean	.865	.905	.882



Table of Contents

- 1 Introduction
- 2 Background
 - Rule extraction
 - Probabilistic prediction
 - Venn-Abers
- 3 Method
- 4 Results
 - Drug discovery case
 - Benchmark data sets
- 5 Concluding remarks



Concluding remarks

- We have in this paper introduced and evaluated rule extractors with well-calibrated fitness estimations
- Here, Venn-Abers was used for calibrating standard decision trees generated from pedagogic rule extraction
- The result is very informative models where each leaf contains a well-calibrated fidelity estimation probability interval.
- In our opinion, this solves the inherent problem with the potentially low test fidelity always present in black-box rule extraction.



Future work






- Dedicated rule extraction algorithms could be used instead of decision trees.
- More generally, we suggest outright comparisons between external explanation modules and well-calibrated rule extractors, investigating the quality of the explanations
- Finally, it should be noted that the fidelity trees introduced here, just like all pedagogic rule extractors, are of course agnostic to whether the opaque model is correct or not
 - An extracted model calibrated using a separate labeled data set can actually include information about the performance of the opaque model on these instances
 - We believe that investigating the exact construction and usability of such accuracy/fidelity estimation models would be very interesting



Thank you!
Questions?



References I

-  Lundberg, Scott M and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.
-  Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ““Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
-  Vovk, Vladimir, Alex Gammerman, and Glenn Shafer (2005). *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc.
-  Vovk, Vladimir and Ivan Petej (2012). “Venn-Abers predictors”. In: *arXiv preprint arXiv:1211.0025*.
-  Zadrozny, B. and C. Elkan (2001). “Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers”. In: *ICML*, pp. 609–616.

