Association Rule Mining for Error Detection in Predictive Models

Alon Harari, Idan Lazar

Abstract

Machine learning models can make systematic errors that follow specific patterns within the data. Error analysis methods typically focus on individual features and simple statistics, potentially missing complex feature interactions that lead to prediction errors. We tried to use apriori to create rules based on errors in data to get a better explanation on why certain errors happen and secondly, try to use these rules to recognize "problematic" rows, and change their prediction. Experimental results show that our approach can identify error patterns but the performance is still lacking.

Problem Description

Classic error analysis sometime lack in intuitive explanation why some errors occurred, it mainly focuses on individual features and less on complex interactions, it can be hard to understand exactly what should be done to improve the model based on the classic analysis, and the analyst based on the error should create manual insights and manually improve the model based on the insights.

Solution Overview

Our solution takes the model errors and runs the apriori algorithm over them. Given the rules we show the top 5 most relevant rules(using confidence and lift) that could explain the error. These rules are represented. After that, we take these rules and try to implement them on the existing model to try and fix the wrong predictions automatically.

The full solution using these steps:

1. First, we take an existing model(in our case xgBoost) and using it to predict the train data. then we find all the rows that the model made a mistake in the prediction.

2. Next, we format the data to be suitable for apriori algorithm, we are discretizing numerical features and formatting categorical features.

3. Now, the error data is ready for the rule mining. We run the apriori algorithm and we generate with it the rules, we only take rules with a minimum 0.1 support. Later we take the top 5 rules and print them.

4. Next, we take the already trained model, but adding to it another layer, after the original model prediction we find rows that correspond to the rules, and we change their prediction.

The solution gives us 2 things, it gives us new error analysis based on association mining rules but also uses them in practice to make the model better.

Experimental Evaluation

To evaluate our approach, we conducted experiments on four different datasets, adult income dataset, heart disease dataset, bank marketing dataset, diabetes disease dataset.

We used xgboost as a baseline model to run our experiment on. We checked the metrics before implementing the error rules and after. Metrics details:

Rule Quality Metrics:

Confidence: The probability that the rule captures the relationship correctly.

Lift: How much more likely the rule is to indicate an error compared to random chance

Coverage: The percentage of data that corresponds to the rules. We want the coverage to be close to the percentage of errors in the original model.

Model Performance Metrics:

Accuracy: The proportion of correct predictions

F1 Score: The harmonic mean of precision and recall

ROC AUC: Area under the ROC curve, it's good in our case because we are looking at binary classification.

Results

Our approach didn't work and just worsened the baseline. This is contrary to our expectations, the experimental results show that our rule-based error correction approach did not outperform the baseline across all datasets.

The results are shown in the tables below.

Performance Comparison

Dataset	Metric	Baseline	Our Approach	Change (%)
BANK_MARKETING	Accuracy	0.9178	0.8429	-7.49%
	F1 Score	0.9123	0.8549	-5.74%
	ROC AUC	0.7387	0.7067	-3.20%
ADULT	Accuracy	0.8695	0.8529	-1.66%
	F1 Score	0.8632	0.8477	-1.55%
	ROC AUC	0.7818	0.7685	-1.33%
DIABETES	Accuracy	0.7338	0.6364	-9.74%
	F1 Score	0.7361	0.6401	-9.60%
	ROC AUC	0.7202	0.6162	-10.40%
HEART	Accuracy	0.8333	0.4000	-43.33%
	F1 Score	0.8348	0.3818	-45.30%
	ROC AUC	0.8403	0.4375	-40.28%

Error fixed vs errors created

Dataset	Errors Fixed	Error created	Net Improvement	Test size	Rows modified	coverage
BANK_MARKETING	40	657	-617	8238	697	8.46%
ADULT	63	171	-108	6513	281	4.31%
DIABETES	8	23	-15	154	37	24.03%
HEART	5	31	-26	60	30	50.00%

Top Association Rules

For each dataset, we identified the top five association rules with the highest lift and then confidence values. These rules were used to determine if we should change the prediction or not.

BANK_MARKETING

Rule ID	Antecedent (IF)	Consequent (THEN)	Confidence	Lift	coverage
1	emp.var.rate=1, cons.price.idx=2, cons.conf.idx=1	month=nov, euribor3m=2	1.00	19.16	0
2	month=nov, euribor3m=2	emp.var.rate=1, cons.price.idx=2, cons.conf.idx=1	1.00	19.16	0
3	emp.var.rate=1, cons.price.idx=2, cons.conf.idx=1	nr.employed=2, month=nov	1.00	19.16	0
4	nr.employed=2, month=nov	emp.var.rate=1, cons.price.idx=2, cons.conf.idx=1	1.00	19.16	697/8238
5	cons.price.idx=2, cons.conf.idx=1, euribor3m=2	nr.employed=2, month=nov	1.00	19.16	0

ADULT

Rule ID	Antecedent (IF)	Consequent (THEN)	Confidence	Lift	coverage
1	sex=Female, native-country=United-States, marital-status=Married-civ-spouse, hours-per-week=0	race=White, relationship=Wife	0.90	11.48	281/6513
2	sex=Female, hours-per-week=0, capital-gain=0, native-country=United-States, marital-status=Married-civ-spouse	race=White, relationship=Wife	0.90	11.48	0
3	sex=Female, native-country=United-States, marital-status=Married-civ-spouse, hours-per-week=0	capital-gain=0, race=White, relationship=Wife	0.90	11.48	0
4	sex=Female, hours-per-week=0, capital-loss=0, native-country=United-	race=White, relationship=Wife	0.90	11.48	0

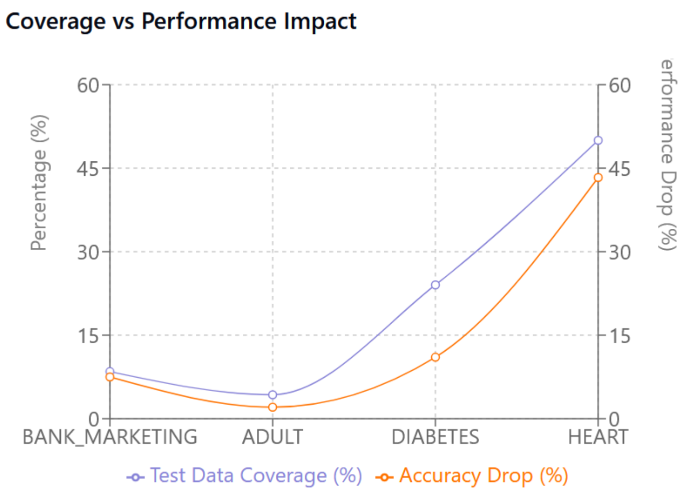
Rule ID	Antecedent (IF)	Consequent (THEN)	Confidence	Lift	coverage
	States, marital-status=Married-civ-spouse				
5	sex=Female, native-country=United-States, marital-status=Married-civ-spouse, hours-per-week=0	capital-loss=0, race=White, relationship=Wife	0.90	11.48	0

DIABETES

Rule ID	Antecedent (IF)	Consequent (THEN)	Confidence	Lift	coverage
1	glucose=1, age=2	blood=pressure_3, bmi=2	1.00	14.25	5/154
2	blood=pressure_3, bmi=2	glucose=1, age=2	0.75	14.25	9/154
3	pregnancies=0, insulin=1, age=1	diabetes=pedigree_3, skin=thickness_2	1.00	9.50	14/154
4	glucose=1, skin=thickness_2, pregnancies=0	diabetes=pedigree_3, insulin=1	1.00	9.50	7/154
5	diabetes=pedigree_3, skin=thickness_2	pregnancies=0, insulin=1, age=1	0.50	9.50	2/154

HEART

Rule ID	Antecedent (IF)	Consequent (THEN)	Confidence	Lift	coverage
1	age=0	chol=0	1.00	4.00	16/60
2	chol=0	age=0	1.00	4.00	10/60
3	age=0	thalach=3	1.00	4.00	0
4	thalach=3	age=0	1.00	4.00	4/60
5	trestbps=3	age=0	1.00	4.00	0



Analysis of results

With all the data we can see few things more clearly. First of all the rules can fit more rows then rows that had error in the baseline model(for example in heart there are only 4 errors in the original model, but our model change the prediction of 30 rows) this happens because the rules that were generated from the errors, actually captured a lot of data without any errors, which means the rules are too general. The high confidence of the rules supports it.

Also, our approach to fixing the rules was too simple. When a row satisfies the rule, we flip its prediction. This is suboptimal, it would be much more preferred if we changed it in a smarter way, maybe by giving some kind of weight for each rule and only rows with high enough weight will change, maybe by changing the prediction to the most common value of this rule. Maybe instead of changing the prediction we could add "new features" where each feature corresponds to rule that applied, and then the model can use it.

From the graph above that compare coverage to error it is clear that the rules that were generated captured a lot of good prediction, and the flipping of the prediction cause the high error. Is caused much more errors than fixes because the high accuracy of the model(more changes in prediction -> more errors).

From the rules tables we can see clear rules that supposed(because we couldn’t make it work) to explain errors in the data. we can see clear rules with most of them having high lift and confidence which means these rules explain the error data.

Related Work

Even though we didn't have the success we wanted, there are papers that did find success. this papers inspired our project.

1. **Rule Mining for Correcting Classification Models** (Das et al., 2022): This work did something very similar to our approach. The difference is that they created their own rules generator by using smart pruning technique while we used directly apriori algorithm, their approach discovers rules of the form X → δ, where X represents a condition that identifies a subpopulation and δ is an optimized correction amount to be added to the model's prediction score for instances in that subpopulation. While we find rules of X -> error where X is a subpopulation and error is the classification of the model that he got wrong, we then find these subpopulations and change their classification. We got inspiration on how to implement the rules mechanism from here.

2. **BoostClean: Automated Error Detection and Repair for Machine Learning** (Krishnan et al., 2017): BoostClean is a system that automates error detection and repair in machine learning pipelines. BoostClean uses statistical boosting to select an ensemble of error detection and repair combinations that maximize model accuracy. BoostClean uses boosting to combine different cleaning strategies, while our solution uses association rule mining to identify specific patterns in errors**.** our solution focus on rules which are more interpretable, Boostclean is more complex. This paper gave us the idea to improve the model automatically based on rules.

In general our approach is more straightforward than the papers, but supposed to give us simple rules that explain errors in the data.

Conclusion

In this research, we have developed and evaluated a framework for automatically discovering and explaining patterns in predictive model errors using association rule mining. Our experiments confirm that model errors can follow patterns that can be captured through association rules. These patterns frequently involve interactions between multiple features that might be missed by traditional error analysis methods. We thought that the solution will be more straightforward, but we discovered that using these rules to improve the model can be challenging, and requires more complex approach also we believe that with more research and work we can use the rules better by creating features based on rules and make more complicated decision on when to change the prediction.

References

1. Das, A., Du, X., Wang, B., Shi, K., Gu, J., Porter, T., & Cardie, C. (2022). Rule Mining for Correcting Classification Models. arXiv:2201.05017v3 [cs.CL].
2. Krishnan, S., Franklin, M. J., Goldberg, K., & Wu, E. (2017). BoostClean: Automated Error Detection and Repair for Machine Learning. arXiv:1711.01299v1 [cs.DB].
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).