Ensemble Methods for Label Noise Detection Under the Noisy at Random Model

Kecia G. de Moura
Ricardo B. C. Prudencio | George D. C. Cavalcanti

Centro de Informática - Universidade Federal de Pernambuco

kgm@cin.ufpe.br, rbcp@cin.ufpe.br, gdcc@cin.ufpe.br

October, 2018

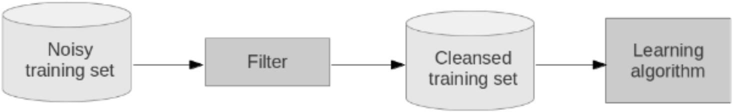
Contents

Introduction

- Noise in classification datasets: attributes and label
- Label noise detection:
  - Algorithms naturally robust
  - Algorithms that model label noise during learning process
  - Filter approaches: an algorithm or an ensemble of algorithms
- Limitation: previous work has evaluated noise detection often assuming noise completely at random in a dataset
- This paper investigates how noise detection is affected by other types of noise (e.g., different levels of noise per class)

Classification noise filtering

Process:



Ensemble filtering:

- Consensus voting: an instance is noisy if all the algorithms in the pool classify it incorrectly
- Majority voting: an instance is noisy if more than half of the algorithms in the pool classify it incorrectly

Label noise models

- NCAR: Noisy completely at random
- NAR: Noisy at random
- NNAR: Noisy not at random

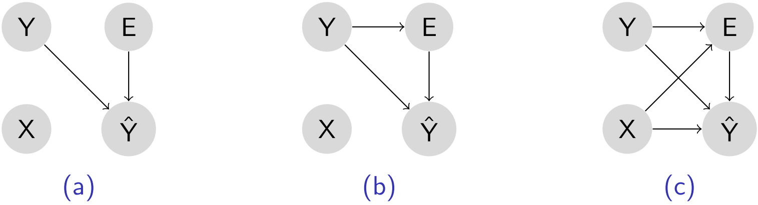


Figure: (a) NCAR, (b) NAR, and (c) NNAR. X = vector of features, Y = true class, $\hat{Y}$ = observed label, and E = error. Arrows report statistical dependencies.

Objectives

Limitations: most previous work assume the NCAR model
Some points investigated in the current work:

- Is noise detection affected by the kind of noise model assumed?
- Regarding the NAR model: how different ratios of noise distribution influence the detection (if so)?
- Is detection performance impacted when dealing with imbalanced datasets? How?

Developed Work

1. Evaluation of ensemble noise detection under the NAR model
2. Experiments varying (1) noise ratio per class; and (2) class imbalance ratio
3. Experiments using synthetic binary datasets: (P2) and ML Bench problems

Methodology - Dataset Generation

- 100 simulated datasets per problem with the following parameters:

N*	Class distribution		Noise distribution	Total percentage of noise in dataset			
	Class 1	Class 2	Class 1 : Class 2				
1	50%	50%	NCAR NAR (1 : 9) NAR (9 : 1)	5%	10%	15%	20%
2	30%	70%	NCAR NAR (1 : 9) NAR (9 : 1)	5%	10%	15%	20%
3	20%	80%	NCAR NAR (1 : 9) NAR (9 : 1)	5%	10%	15%	20%

*Experiment number.

Dataset: P2 problem

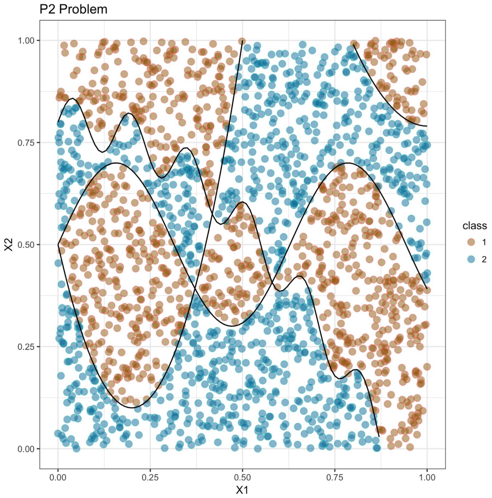


Figure: The P2 Problem.

Methodology - Ensemble Noise Detection

BRACIS 2018
IBM RESEARCH | SÃO PAULO, SP, BRAZIL

- Learning algorithms for classification noise filtering

Algorithm	**R Package**	**Algorithm**	**R Package**
CN2 (rule learner)	RoughSets	J48	RWeka
kNN (nearest neighbor)	class	JRip	RWeka
Naive Bayes	naivebayes	Multiplayer perceptron	RSNNS
Random forest	randomForest	Decision tree	party
SVM (RBF Kernel)	e1071	SMO (linear Kernel)	RWeka

- Majority was used as the voting approach

Centro de Informática U·F·P·E

Methodology - Performance Measures

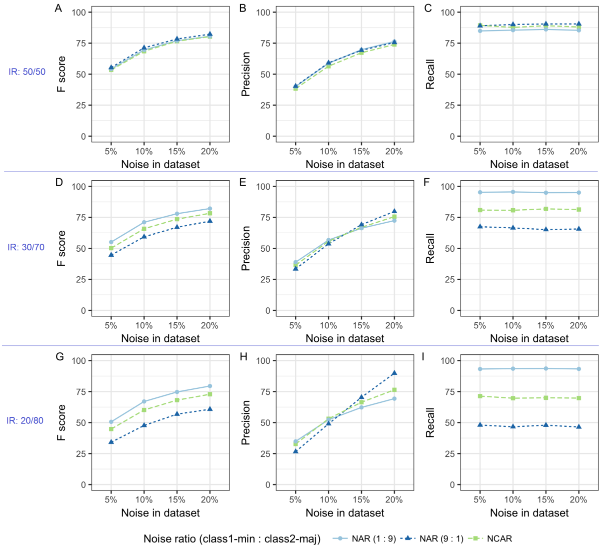
- Three measures to evaluate noise detection performance

$$\text{Precision} = \frac{\text{number of noisy cases correctly identified}}{\text{number of all noisy cases identified}}$$

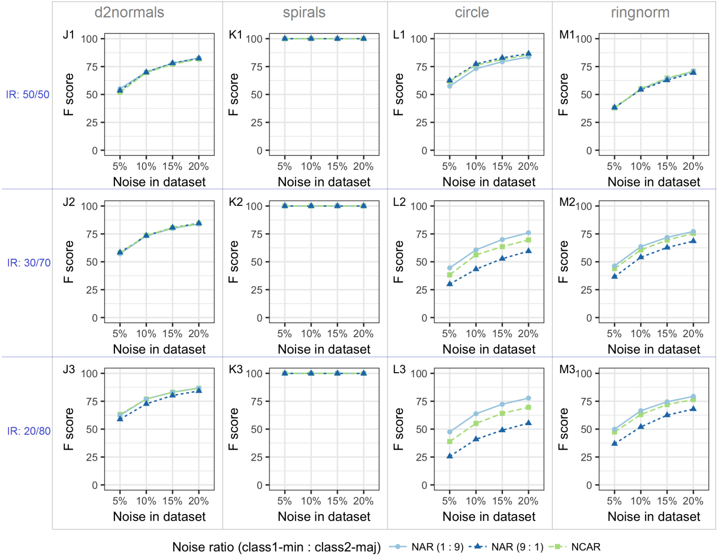
$$\text{Recall} = \frac{\text{number of noisy cases correctly identified}}{\text{number of all noisy cases in dataset}}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

Results for P2 problem



Results for mlbench datasets



Summary of Results

- Noise detection has the same performance for NAR and NCAR models when dealing with balanced datasets;
- Nevertheless, there is a significant difference in noise identification for imbalanced class datasets:
  - Both NAR and NCAR approaches harmed the detection of noise as class distribution disparity gets higher
  - When the minority class is noisier than the majority class, detectors' performance is decreased
  - When majority class is the noisiest one: noise identification is improved when class imbalance ratio is not too discrepant (for instance, "30/70")

Conclusion

- Noise detection is significantly impacted by the type of noise in the dataset
- This paper findings can be used as a guide to do experiments with properly noise injection.
- Limitations and future work:
  - Expand studies to cover also the NNAR model
  - Use real-world datasets
  - Apply a different ensemble of algorithms
  - Investigate the results for a different noise detection approach

Acknowledgments

BRACIS 2018
IBM RESEARCH | SÃO PAULO, SP, BRAZIL

Thank you!

This work was funded in part by the UFPE Information Center (CIn), the Science and Technology Support Foundation of Pernambuco State (FACEPE), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the National Council of Scientific and Technological Development (CNPq).

Centro de Informática U · F · P · E

Questions