Noise detection in classification problems

Luís P. F. Garcia[1,2], Ana C. Lorena[3], André C. P. L. F. de Carvalho[2]

[1]Department of Computer Science, University of Leipzig, Germany

[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil

[3]Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Brasil

garcia@informatik.uni-leipzig.de, aclorena@unifesp.br, andre@icmc.usp.br

Abstract. *Large volumes of data have been produced in many application domains. Nonetheless, when data quality is low, the performance of Machine Learning techniques is harmed. Real data are frequently affected by the presence of noise, which, when used in the training of Machine Learning techniques for predictive tasks, can result in complex models, with high induction time and low predictive performance. Identification and removal of noise can improve data quality and, as a result, the induced model. This thesis proposes new techniques for noise detection and the development of a recommendation system based on meta-learning to recommend the most suitable filter for new tasks. Experiments using artificial and real datasets show the relevance of this research.*

1. Introduction

In real world applications, there are many inconsistencies that affect data quality. Data cleansing investigates and employs techniques able to automatically treat data quality problems. This work is concerned with noise detection, which can be treated by noise filter (NF) techniques [Frenay and Verleysen 2014].

In class labeled datasets, noise can be present in the predictive attributes and in the target attribute [Zhu and Wu 2004]. In the second case, noise changes the true class label of an example. This has been regarded as the most harmful noise type. Several studies show that the use of NFs can improve the classification performance and reduce the complexity of predictive models [Brodley and Friedl 1996]. However, like in the "No Free Lunch" Theorem [Wolpert 1992], no NF is superior to all others for all datasets. This thesis investigates the hypothesis that noise removal could be more efficient if the a more suitable NF is recommended for a new dataset.

Meta-Learning (MtL) has been successfully used to support the recommendation of the most suitable technique(s) for a new dataset [Brazdil et al. 2009]. In this Thesis, given a set of frequently used NFs and a set of data complexity measures [Ho and Basu 2002] able to characterize datasets, a MtL-based system was proposed and experimentally evaluated to recommend the most suitable NF for a new dataset, thus supporting the selection of the most suitable NFs and improving label noise identification.

The main contributions of this thesis are: (i) proposal of two NFs, one based on ensembles of classifiers and the other based on a subset of data complexity measures [Garcia et al. 2015a, Garcia et al. 2012]; (ii) use of complexity measures to understand how the presence of noise in a dataset affects data complexity [Garcia et al. 2013,

Garcia et al. 2015a]; (iii) use of decomposition strategies to increase NF performance by decreasing the complexity of multi-class classification tasks [Garcia et al. 2015b]; (iv) soft adaptation of NFs for outputting a "Noise Degree Prediction" (NDP) value and the proposal of a new evaluation measure for soft NFs [Garcia et al. 2016b, Lorena et al. 2015]; (v) use of MtL to predict the NF with the best predictive performance and lower computational cost for a new task [Garcia et al. 2016a, Garcia et al. 2016b] and (vi) validation of the MtL-based recommender system in a real dataset from the ecology domain, with the support of a domain expert [Garcia et al. 2016b].

The rest of this paper is organized as follows. Section 2 has the main motivations for this Thesis, with an overview of NFs and MtL in recommendation systems. Section 3 presents the datasets adopted, the methodology followed to evaluate the NFs and the results obtained in the MtL analysis. Section 4 describes a case study where experimental results using an ecology dataset are validated by a domain expert. Section 5 presents the main conclusions from this Thesis and Section 6 enumerates the resulting publications.

2. Noise Filter Recommendation by Meta-Learning

Different information can be used by NFs, like neighborhood or density information [Tomek 1976, Garcia et al. 2015a], descriptors extracted from the dataset [Sluban et al. 2014] and classifiers [Garcia et al. 2012, Sluban et al. 2010, Brodley and Friedl 1996] for the noise identification. Table 1 shows the NFs investigated in the Thesis, which include NFs frequently used and two NFs proposed in the Thesis, DEF and GNN. DEF employs an ensemble of NF classifiers, whose classifiers are selected according to the dataset used [Garcia et al. 2012]. GNN models a dataset as a graph and identify noise by extracting a set of measures from the graph [Garcia et al. 2015a].

Table 1. List of NFs with acronym and reference.

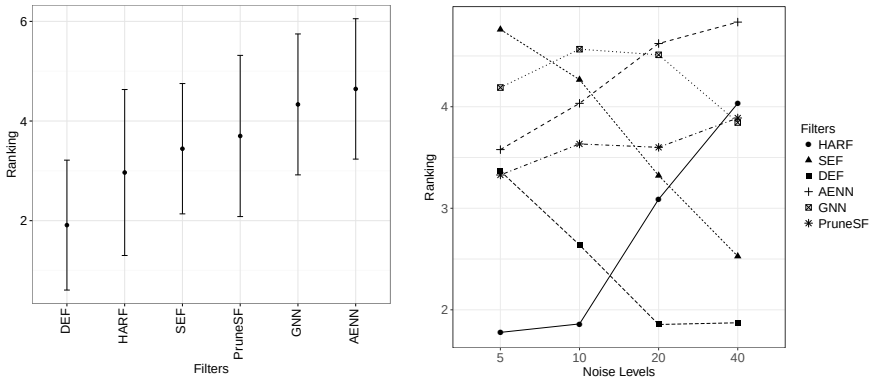
Filter	Acronym	Reference
All k-NN	AENN	[Tomek 1976]
Static Ensemble Filter	SEF	[Brodley and Friedl 1996]
High Agreement Random Forest Filter	HARF	[Sluban et al. 2010]
Dynamic Ensemble Filter	DEF	[Garcia et al. 2012]
Prune Saturation Filter	PruneSF	[Sluban et al. 2014]
Graph Nearest Neighbor	GNN	[Garcia et al. 2015a]

Since each NF has a bias, different NFs can present distinct noise identification performances in diverse datasets [Sluban et al. 2014]. This study investigates how MtL can be used to recommend the most suitable NF for new datasets [Brazdil et al. 2009]. The first step is the construction of a meta-dataset, where each meta-example is usually associated with a noisy version of a dataset, from which a set of characterization and complexity measures, named meta-features, are extracted. Each meta-example receives a label, which is the performance obtained by a set of NFs when applied to this dataset. Next, a ML technique is applied to the meta-dataset, as a conventional ML task, to induce a meta-model, which can be used as a recommendation system to select the most suitable NF for a new dataset. The use of complexity measures was motivated by experiments from this Thesis, when many of these measures captured the increasing complexity of a dataset in the presence of label noise [Garcia et al. 2015a]. Meta-regressors were built to predict the expected F_1 performance of various NFs, which were used to recommend NFs for new datasets.

3. Experimental

All techniques are evaluated in noisy versions of 90 benchmark datasets, created by using a random noise imputation method. For each dataset, random noise was added at rates of 5%, 10%, 20% and 40%. For each dataset and noise level, 10 different noisy versions were generated, resulting in 3600 datasets with class noise. The NFs were evaluated in noise identification using the F_β-score with $\beta = 1$. The Friedman statistical test [Demšar 2006] with 95% of confidence value compared their predictive performances.

Figure 1 shows the performance of the NFs. Figure 1(a) summarizes the F_1 predictive performance of all NFs. It shows the average ranking of each NF, regarding its predictive performance for all datasets, independently of the noise level introduced. Each value in the x-axis represents a NF. The y-axis shows the average and the standard deviation of the ranking of each NF. Figure 1(b) summarizes the ranking position of the NFs for all datasets for each noise level. The NF with the best predictive performance has the lowest average (and standard deviation) ranking values.



(a) Ranking of NFs according to F_1 performance.

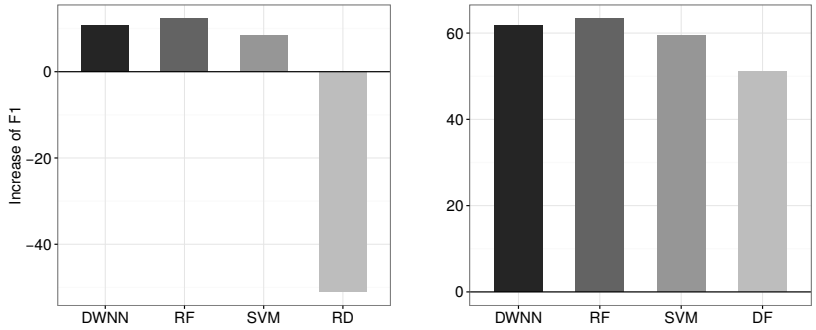
(b) Ranking of NFs according to F_1 performance per noise level.

Figure 1. Predictive performance of the NF techniques.

According to Figure 1(a), DEF was the best performing NF, followed by HARF and later SEF. PruneSF, GNN and AENN presented the worst performances. Considering the combined results of the F_1 performance shown in Figure 1(b) and of the statistical test performed, HARF was able to improve the F_1 values for low noise rates, while DEF was able to improve its performance for high noise rates. SEF was the worst NF for low noise rates, while AENN was the worst NF for high noise rates. GNN had the worst results for intermediate noise rates. Therefore, the choice of a NF depends on the expected noise level of a particular dataset. Overall, DEF should be preferred when a high noise level is expected, and HARF when the noise level is low. The dataset characteristics can also influence the results obtained, since each NF has a bias that can better fit specific cases. This motivated the use of MtL in the domain of label noise identification in the Thesis.

The performance of the MtL recommendation system is illustrated in Figure 2, which shows the increase of F_1 obtained by the NFs when the NF predicted as the best by the meta-regressors is used in noise detection (base-level) instead of the NF predicted by the baselines. The baselines are DF (default - Figure 2(a)) and RD (random - Figure 2(b)). The x-axis shows the meta-regressors and the y-axis the increase of F_1 when compared

with the corresponding baseline. Positive values indicate an increase of F_1 and negative values a decrease. In Figure 2(a), the increase in the base-level predictive performance obtained by using the meta-regressors DWNN (Distance Weight Nearest Neighbor), RF (Random Forest) and SVM (Support Vector Machine) were higher than using the DF baseline. RD had a high decrease of performance. All meta-regressors in Figure 2(b) increased their performance, including the DF baseline. The RF meta-regressor presented the best results in both cases.



(a) Difference of performance in the base-level when using DF as baseline.

(b) Difference of performance in the base-level when using RD as baseline.

Figure 2. MtL performance.

4. Case Study: Ecology Data

To validate the proposed approaches, the performance of the NF recommended by the MtL-based system was compared with the baseline NFs when both were applied to a real dataset from ecological niche modeling domain. The comparison was validated by a domain expert and the baseline had the lowest performance. The dataset shows the presence or absence of a non native specie *Hedychium coronarium* in georeferenced regions from protected areas of the Brazilian state of São Paulo. Both classes can present label noise, with the presence or absence misclassification.

Using the previous NFs adapted to a soft version as described in Chapter 3 and the MtL described in Chapter 4 of the thesis, 59 examples were detected as noisy, 12 in the absence class and 47 in the presence class. Only two examples were misclassified by the NF in the presence class. This example has native vegetation in riparian zone, which is favorable to the invasion. In the absence class, five of the examples misclassified are examples where the location and the conservation status do not favor the appearance of *H. coronarium*.

Overall, the filtering step efficiently identified potentially noisy examples. For data modeling, these examples should be removed to avoid their negative effect on the induced model. From the domain expert point of view, these examples should be monitored, since they represent areas in process of degeneration.

5. Conclusion

This Thesis investigated NFs for data cleansing and the use of MtL for NF recommendation. For such, the authors developed new NFs, proposed and investigated alternatives to

use MtL for NF recommendation and evaluated the use of MtL in a real dataset, whose results were validated by a domain expert. The main limitation of this Thesis is to not take into account the intrinsic noise levels in the real datasets used in this work, since it is usually not possible to assert that an example really has a noisy label. The parameters used by the NFs were those adopted in the reference literature. Future work includes fine tuning the hyper-parameters of the NFs, proposal of NFs specific for particular datasets and study the noisy patterns present in datasets. The authors would also like to investigate the influence of intrinsic dataset noise level in the performance of NFs.

6. Publications from this Thesis

This thesis resulted in publications in journals and conferences and in the implementation of publicly available R packages. The source codes of part of the experiments are also available in the GitHub[1] platform. Some of these publications resulted from internships abroad, in collaborations with Francisco Herrera from University of Granada and with Stan Matwin from Dalhousie University. Next, the papers and packages produced are listed.

Journal papers

- Garcia, L., de Carvalho, A., & Lorena, A. (2015). "Effect of label noise in the complexity of classification problems". Neurocomputing, 160:108 - 119. (JCR 2015 - 2.392)
- Garcia, L., Sáez, J., Luengo, J., Lorena, A., de Carvalho, A., & Herrera F. (2015). "Using the One-vs-One decomposition to improve the performance of class noise filters via an aggregation strategy in multi-class classification problems". Knowledge-Based Systems, 90:153 - 164. (JCR 2015 - 3.325)
- Garcia, L., de Carvalho, A., & Lorena, A. (2016). "Noise detection in the meta-learning level". Neurocomputing, 176:14 - 25. (JCR 2015 - 2.392)
- Garcia, L., Lorena, A., Matwin, S., & de Carvalho, A. (2016). "Ensembles of label noise filters: a ranking approach". Data Mining and Knowledge Discovery, 30(5):1192 - 1216. (JCR 2015 - 2.714)
- Morales, P., Luengo, J., Garcia, L., Lorena, A., de Carvalho, A., & Herrera F. (2017). "The NoiseFiltersR Package: Label Noise Preprocessing in R". The R Journal. *Accepted.* (JCR 2015 - 1.045)

Conference papers

- Garcia, L., Lorena, A., & de Carvalho, A. (2012). "A study on class noise detection and elimination". In Brazilian Symp. on Neural Networks (SBRN), 13 - 18.
- Garcia, L., de Carvalho, A., & Lorena, A. (2013). "Noisy data set identification". In Hybrid Artificial Intelligent Systems (HAIS), 629 - 638.
- Lorena, A., Garcia, L., & de Carvalho, A. (2015). "Adapting Noise Filters for Ranking". In Brazilian Conference on Intelligent Systems (BRACIS), 299 - 304.

R Packages in the CRAN Repository

- Morales, P., Luengo, J., Garcia, L., Lorena, A., de Carvalho, A., & Herrera F. (2016). "NoiseFiltersR: Label Noise Filters for Data Preprocessing in Classification". R package version 0.1.0. https://CRAN.R-project.org/package=NoiseFiltersR.
- Rivolli, A., Garcia, L., & de Carvalho, A. (2017). "mfe: Meta-Feature Extractor". R package version 0.1.0. https://CRAN.R-project.org/package=mfe.

[1] `https://github.com/lpfgarcia`

References

Brazdil, P., Giraud-Carrier, C., Soares, C., and Vilalta, R. (2009). *Metalearning - Applications to Data Mining*. Cognitive Technologies. Springer, 1 edition.

Brodley, C. and Friedl, M. (1996). Identifying and eliminating mislabeled training instances. In *13th National Conference on Artificial Intelligence (AAAI)*, pages 799–805.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Frenay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE Trans. on Neural Networks and Learning Systems*, 25(5):845–869.

Garcia, L., de Carvalho, A., and Lorena, A. (2013). Noisy data set identification. In *Hybrid Artificial Intelligent Systems (HAIS)*, volume 8073, pages 629–638.

Garcia, L., de Carvalho, A., and Lorena, A. (2015a). Effect of label noise in the complexity of classification problems. *Neurocomputing*, 160:108–119.

Garcia, L., de Carvalho, A., and Lorena, A. (2016a). Noise detection in the meta-learning level. *Neurocomputing*, 176:14–25.

Garcia, L., Lorena, A., and de Carvalho, A. (2012). A study on class noise detection and elimination. In *Brazilian Symposium on Neural Networks (SBRN)*, pages 13–18.

Garcia, L., Lorena, A., and de Carvalho, A. (2016b). Ensembles of label noise filters: a ranking approach. *Data Mining and Knowledge Discovery*, 30(5):1192 – 1216.

Garcia, L., Sáez, J., Luengo, J., Lorena, A., de Carvalho, A., and Herrera, F. (2015b). Using the one-vs-one decomposition to improve the performance of class noise filters via an aggregation strategy in multi-class classification problems. *Knowledge-Based Systems*, 90:153–164.

Ho, T. and Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):289–300.

Lorena, A., Garcia, L., and de Carvalho, A. (2015). Adapting noise filters for ranking. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 299–304.

Sluban, B., Gamberger, D., and Lavrač, N. (2010). Advances in class noise detection. In *19th European Conference on Artificial Intelligence (ECAI)*, pages 1105–1106.

Sluban, B., Gamberger, D., and Lavrač, N. (2014). Ensemble-based noise detection: noise ranking and visual performance evaluation. *Data Mining and Knowledge Discovery*, 28(2):265–303.

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Trans. on Systems, Man and Cybernetics*, 6(6):448–452.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Zhu, X. and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210.