Data quality in the context of classification tasks

Roxane Jouseau
roxane.jouseau@doctorant.uca.fr
UCA - LIMOS
Clermont-Ferrand, France

Sébastien Salva
sebastien.salva@uca.fr
UCA - LIMOS
Clermont-Ferrand, France

Chafik Samir
UCA - LIMOS
Clermont-Ferrand, France
chafik.samir@uca.fr

ABSTRACT

Data cleaning is an important step of a machine learning process to get the best results possible. The literature is rich, and there are many tools available, which makes choosing which tool to use complex. The objective of our work is to answer the question: Is it always best to repair data? We focus on numeric data for classification tasks. We decompose the question into five criteria, we propose a metric to measure how difficult using a repairing tool is. Then, we studied the impact of the degree of degradation of data, the type of errors present, the effectiveness of repairing tools, and the impact of different classification models. We found that error types such as missing values and outliers have more impact on accuracy and f1 score than other types of errors. Moreover, even though complex repairing tools were generally more effective, there is a point where data is so degraded that tools do not perform well. For low levels of errors, the tools also tend to have similar performances, the decision of which one to use can then be made according to their difficulty to use.

KEYWORDS

data quality, data repairing, data errors, classification

ACM Reference Format:
Roxane Jouseau, Sébastien Salva, and Chafik Samir. . Data quality in the context of classification tasks. In *Proceedings of Gestion de Données – Principes, Technologies et Applications (BDA 2022).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Data is a the heart of machine learning, having quality data is therefore crucial to ensure exploitable results. The field of data cleaning and repairing is very active, from the bibliographic work we did we extracted four key points: 1. there is a lot of repairing methods in the literature (e.g: HoloClean [12], Katara [5], ZeroER [13], CleanMl [9], REPAIR [10], OpenRefine [6], BoostClean [8], the data linter [7] and many others), 2. data holds different levels of degradation [9], 3. repairing methods require various metadata which can sometimes be very complex to produce [12], [5], [13], [9], [10] and, 3. data holds different types of errors [1], [4], [2].

From these points, we developed a research question: Is it always best to repair data? Our main contributions are a metric for the difficulty of using a repairing method and experimental results on classification tasks with deteriorated data.

2 SCOPE OF OUR WORK

Our work focuses on numeric data in the context of classification tasks with degradation at different levels of errors. We considered five types of errors: missing values ([9], [12], [5]), outliers, domain value violations ([8]), exact duplicates ([9], [7], [1]), and partial duplicates ([13]). We used 7 numerical datasets with various sizes, dimensions, and subjects: mnist, fashion-mnist, olivetti, iris, adult, breast cancer, and wine [11], [14], [3]. We also have decided to include the following classification models: Logistic regression, K-Nearest Neighbors, Decision tree, Random forest, Ada boost, Naïve Bayes, XGboost, Support vector classification, Gaussian process, Multi-layer perceptron, Stochastic gradient descent, and Gradient boosting [11].

We answer the research question through five criteria. With the first criteria C1, we studied the perceived difficulty of using a method according to experts. To answer these questions we propose a metric to evaluate how difficult using a repairing method is (C1). C2 focuses on studying the impact of the degradation of the data. With C3, we study the impact of the type of error present in data. We also investigate the effectiveness of repairing tools with C4 and the impact of the classification model used with C5. For C2 to C5 we designed an experiment that allow us to observe the impact of data repairing on classification tasks at different levels of degradation of the data. In this paper we will first present the metric we developed for C1, we will then present the experiments we conducted for C2, C3, C4, and C5 with a summary of our results and conclusions concerning these criteria. Finally, we conclude on our main research question using the five criteria we identified.

3 PRELIMINARY RESULTS

Traditionally, repairing methods are evaluated and compared on the difference between their accuracy and f1 score on unrepaired and repaired data. This kind of comparison is very limited other criteria should be considered such as producing complex metadata for a repairing method which can be both complex and time-consuming. To account for this, we propose to evaluate repairing methods based on how difficult they are to use according to data scientists. We first started by breaking down the repairing methods we found in the literature into elementary tasks describing each action needed to use them including creating the metadata needed. Given an error type, and a repairing method R_X, we built a tree expressing the steps of R_X. For any other repairing method R_Y we completed the tree when required. The final nodes of the tree are elementary tasks. Figure 1 is an example of the tree constructed for repairing missing values. We then asked a panel of eight industry data scientists to rank every elementary task individually on a four-level scale from easy to hard. These rankings allowed us to compute a weighted mean for the difficulty of each elementary task. We finally derive an overall difficulty ranking of every method by using a linear combination of the elementary tasks.

To study the criteria C2, C3, C4, and C5 we designed an experiment where for each of the seven datasets we included, we

BDA 2022, October 24 - 26 2022, Clermont-Ferrand, France

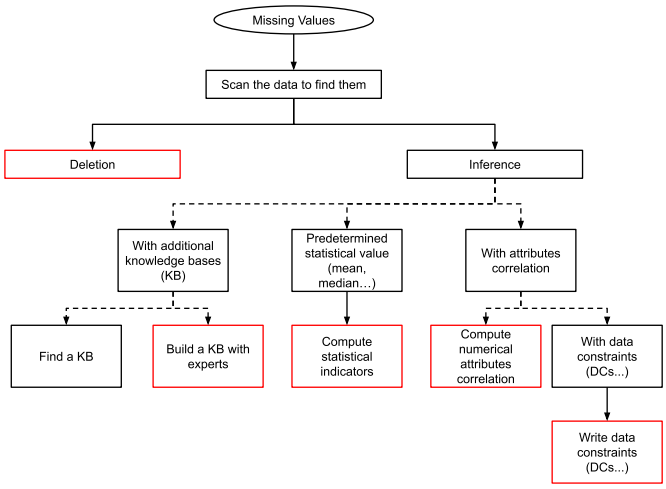

Figure 1: Elementary tree decomposition for missing values

split the dataset into training and test. The test remained intact throughout the experience, while the training was subjected to modifications. We first injected the training dataset with one type of error at a percentage varying from 0 to 95% with increments of 5%. We applied a panel of repairing methods to different copies of the deteriorated dataset to obtain cleaned datasets. We then used them to train several classification models. Finally, we computed the accuracy and f1-score on the tests. We executed the complete process 30 times to reduce the bias for each percentage level. Mainly our results allowed us to extract these observations. With C2 and C3 we identified 2 categories of error types: category 1 (exact duplicates, partial duplicates, and domain value violations) of data degradation has little to no impact on the accuracy, and category 2 (missing values and outliers) the level of data degradation seems to have a big impact. Studying C4 showed us that some methods do perform better but at high levels of degradation the effectiveness of the different repairing methods seems to be leveled out. Moreover, at levels of degradation under 10%, most repairing methods perform well which means that depending on the accuracy we aim to obtain a simpler method could be sufficient. C5 mainly showed us that the error type classification models are the most sensitive to is outliers.

4 CONCLUSION

Our work studied the impact of repairing errors on classification tasks to answer the question: Is it always best to repair data? We presented five criteria to answer this question. Firstly we identified two categories of error types: low-impact and high-impact errors. In the case of low-impact errors, it seems unnecessary to use a very complex repairing method as the impact of the error is relatively low. Secondly, with all the error types we studied, there was a point where the data was too deteriorated, and all repairing methods were performing equivalently poorly. We were also able to identify cases where a simple and a more complex method performed similarly. In these cases, our difficulty ratings of the methods are particularly relevant for comparing both.

Extensions of this work to applications other than classification tasks are possible directions for future work. Moreover, additional research could include more data types than numeric, especially more complex data types such as time series, which would imply more possible error types.

REFERENCES

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting Data Errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment* 9, 12 (2016), 993–1004.

[2] Otmane Azeroual, Gunter Saake, and Mohammad Abuosba. 2018. Data Quality Measures and Data Cleansing for Research Information Systems. *Journal of Digital Information Management* 16, 1 (Feb 2018).

[3] C. Blake and C. Merz. 1998. UCI repository of machine learning databases.

[4] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. *SIGMOD Proceedings of the 2016 international conference on management of data* (2016), 2201–2206. https://doi.org/10.1145/2882903.2912574

[5] Xu Chu, John Morcos, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. 2015. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 1247–1261.

[6] Kelli Ham. 2013. OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association: JMLA* 101, 3 (2013), 233.

[7] Nick Hynes, D. Sculley, and Michael Terry. 2017. *The Data Linter: Lightweight, Automated Sanity Checking for ML Data Sets.*

[8] Sanjay Krishnan, Michael J Franklin, Ken Goldberg, and Eugene Wu. 2017. Boostclean: Automated error detection and repair for machine learning. *arXiv preprint arXiv:1711.01299* (2017).

[9] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2021. CleanML: A study for Evaluating the IMpact of Data Cleaning on ML Classification Tasks. *36th IEEE International Conference on Data Engineering (ICDE 2020)(virtual)* (2021).

[10] Yi Li and Nuno Vasconcelos. 2019. REPAIR: Removing Representation Bias by Dataset Resampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 9572–9581.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[12] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proceedings of the VLDB Endowment* 10, 11 (2017).

[13] Renzhi Wu, Sanya Chaba, Saurabh Sawlani, Xu Chu, and Saravanan Thirumuruganathan. 2020. Zeroer: Entity resolution using zero labeled examples. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1149–1164.

[14] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747 [cs.LG]