Towards Smart Data Technologies for Big Data Analytics

María José Basgall[1,2,3][0000−0002−7024−847X], Marcelo Naiouf[2][0000−0001−9127−3212], Francisco Herrera[3][0000−0002−7283−312X], and Alberto Fernández[3][0000−0002−6480−8434]

[1] UNLP, CONICET, III-LIDI, La Plata, Argentina
mjbasgall@lidi.info.unlp.edu.ar
[2] Instituto de Investigación en Informática (III-LIDI), CIC-PBA, Facultad de Informática - Universidad Nacional de La Plata, Argentina
[3] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada, Spain

Abstract. Currently the publicly available datasets for Big Data Analytics are of different qualities, and obtaining the expected behavior from the Machine Learning algorithms is crucial. Furthermore, since working with a huge amount of data is usually a time-demanding task, to have high quality data is required. Smart Data refers to the process of transforming Big Data into clean and reliable data, and this can be accomplished by converting them, reducing unnecessary volume of data or applying some preprocessing techniques with the aim of improve their quality, and still to obtain trustworthy results. We present those properties that affect the quality of data. Also, the available proposals to analyze the quality of huge amount of data and to cope with low quality datasets in an scalable way, are commented. Furthermore, the need for a methodology towards Smart Data is highlighted.

Keywords: Big Data · Smart Data · Data Complexity · Data Quality.

1 Introduction

The Big Data term [13] refers to he enormous amount of data that is being generated increasingly and from several sources, with a strong relationship with both velocity and variety. However Big Data does not entail a good quality of data. In the field of Data Science, obtaining knowledge from datasets is the main task. Unfortunately, several data complexities can degrade the quality of the problems, and in turn yield to inaccurate results. Among others, class imbalance, overlapping, redundancy, outliers, missing values, can be stressed [5, 3].

Whether it is due to the nature of the problem, or due to the way in which data is obtained or generated, most of the publicly available big datasets have different qualities. This can be identified of one of the causes for being unable to replicate the good behavior of standard techniques in Big Data benchmarks [1, 2]. Because of all the above, it is crucial to identify the data complexity in order

to take action and to apply Big Data preprocessing techniques towards Smart Data [7], with the aim to learn from high quality data.

In this contribution, the data characteristics that affect the expected behavior of a knowledge extraction technique and how they are represented in the publicly available datasets for Big Data, are presented. Furthermore, we introduce the current proposals to cope with data quality, and the need for more technologies to turn Big Data into Smart Data is commented. This work is part of the ongoing doctoral thesis based on the analysis and design of preprocessing techniques for Imbalanced Big Data problems.

2 Towards Smart Data in Big Data Analytics

Identifying data complexities is very important in order to decide about which preprocessing technique or Machine Learning algorithm to apply. In data classification, class imbalance [9] refers to an uneven data distribution between the classes of a problem. The overlapping areas [5] of a problem are ambiguous regions of the data space, where there are instances belonging to different classes. In addition, if a dataset contains a subset of examples with different values in their features but representing the same concept, those instances are considered as redundant. Therefore redundancy is more than exact copies of the examples in a set of data [12, 10]. At the same time, outliers [4] relate to those instances that contain in their attributes very distant values from the common order of magnitude of the remainder. Finally, missing values [11] occurs, as its name suggests, when an instance has one or more of its features values lost.

In Academia, publicly available big datasets are of different qualities, presenting a variable type of undesirable characteristics. For instance, many Big Data classification problems are about biotechnology or related to new physics discoveries, and they usually pose a high degree of imbalance between its classes, in addition to other intrinsic complexities. Therefore, each data problem should be treated as an independent project in order to apply Smart Data technologies with the aim to improve the data quality. Therefore, if Smart Data is not taken into account, Machine Learning models could lead to misleading results.

In [1] and [2], we have studied the behaviour of the well-known oversampling technique called SMOTE [6] for Imbalanced Big Data Classification problems without following any additional Smart Data technology. The study was based using two different kind of design approaches in order to be scalable. SMOTE is one of the most widely used technique to balanced a dataset in small data scenarios due to its simplicity, but also for offering better results with respect to the standard random solutions. Those work contributions showed that the results achieved were not as good as the ones in small datasets, and one reason was pointed out as the lack of data quality.

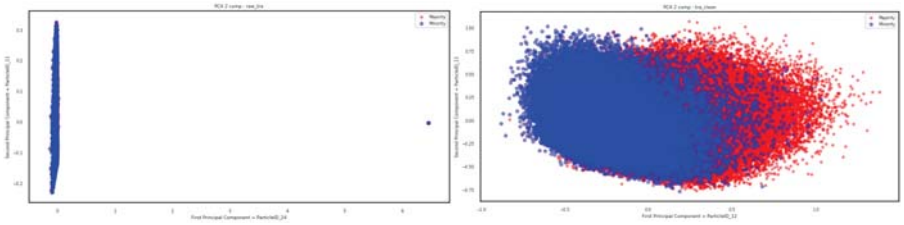
Regarding the current proposals towards Smart Data we may found two main papers. On the one hand, in [10] two novel metrics to describe the quality of a big dataset in a scalable way were proposed, jointly to another basic metrics as a Spark-package. Authors have found a redundancy of information in most of

the Big Data Classification problems. The main conclusion was that randomly decreasing the datasets up to 25% do not affect significantly the performance of the classifiers applied. On the other hand, in [8] a Smart Data based ensemble for Big Data Imbalanced Classification problems was proposed. Authors reached an improvement for data quality by combining different preprocessing techniques.

3 A case study

In this section, a brief analysis over the `MiniBooNE` dataset as a case study is presented. The objective is to briefly show the incidence of some of several intrisinc data characteristics in a Big Data classification task. Specifically, the most straightforward characteristics to discover in a dataset are the imbalanced degree, the replicated instances (as part of the redundancy property), missing values, and rare values of features, among others.

The `MiniBooNE` dataset has more than 130,000 instances and a high dimensionality (49 continuous features). It represents a binary classification problem from the physic field and it aims to distinguish signal from background, where the signal is the 36.5% of the dataset. Using a pipeline of standard techniques, we have been able to determine that `MiniBooNE` does not contain missing values, and a 0.36 % of the data are replicated instances (which we have removed). Furthermore, by means of a graphical analysis we have detected values far from the common order of magnitude of the `ParticleID_19` feature, and the instances with those extreme values were removed from the raw dataset. The incidence of this action can be seen in Fig. 1, where the two principal components (PC) for the raw dataset (Fig. 1a) and for the new version of it (Fig. 1b) are shown.



(a) Principal Components for raw dataset (b) Principal Components for clean dataset

Fig. 1: Principal component analysis (PCA) for `MiniBooNE` dataset before and after cleaning stage

After the aforementioned pipeline, a Decision Tree classifier to learn from the raw and from the clean version of the dataset was applied. Table 1 shows the widespread metrics used for imbalanced classification problems. A slight improvement in detecting the minority class instances can be seen. Considering that no further preprocessing techniques have been applied to the dataset except

for these basic ones, a trend of improving classification results is evident as more Smart Data technologies are applied.

Table 1: Decision Tree classifier results for the `MiniBooNE` dataset

	GM	AUC	TPR	TNR
raw	0.8556	0.8590	0.7825	0.9355
clean	0.8636	0.8659	0.8025	0.9293

4 Conclusions

The data quality analysis of the Big Data sets is almost an uncharted territory, and Smart Data is also a fledgling topic. An exhaustive study of the data properties, together with the application of the proper preprocessing techniques, has become mandatory for all Data Science projects in both industry and academia. By considering a case study on Big Data classification, we have determined that even straightforward data transformation allowed at improving the modeling process, thus stressing the need towards the use and development of Smart Data technologies.

References

1. Basgall, M.J., et al.: Smote-bd: An exact and scalable oversampling method for imbalanced classification in big data. JCS&T **18**(03), e23 (Dec 2018)
2. Basgall, M.J., et al.: An analysis of local and global solutions to address big data imbalanced classification: A case study with smote preprocessing. In: VII JCC&BD. vol. 1050, pp. 75–85. Springer (2019)
3. Das, S., et al.: Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition **81**, 674–693 (2018)
4. Devi, D., et al.: Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. Pattern Recognit. Lett. **93**, 3–12 (Jul 2017)
5. Fernandez, A., et al.: Learning from Imbalanced Data Sets. Springer (2018)
6. Fernández, A., García, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)
7. García-Gil, D., et al.: Enabling smart data: Noise filtering in big data classification. Inf. Sci. **479**, 135–152 (2019)
8. García-Gil, D., et al.: Smart data based ensemble for imbalanced big data classification (2020)
9. López, V., et al.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**(20), 113–141 (2013)
10. Maillo, J., et al.: Redundancy and complexity metrics for big data classification: Towards smart data. IEEE Access pp. 1–1 (2020)
11. Montesdeoca, B., et al.: A first approach on big data missing values imputation. In: IoTBDS (2019)
12. ur Rehman, M.H., et al.: Big Data Reduction Methods: A Survey. DSE **1**(4), 265–284 (Dec 2016)
13. Wu, X., et al.: Data mining with big data. IEEE TKDE **26**(1), 97–107 (Jan 2014)