Emerging topics and challenges of learning from noisy data in non-standard classification: A survey beyond binary class noise

Ronaldo C. Prati
Center of Mathematics, Computer Science and Cognition
Federal University of ABC
Santo André, São Paulo, Brazil
ronaldo.prati@ufabc.edu.br

Julián Luengo
Data Science and Computational Intelligence Institute
University of Granada
Granada, Spain
julianlm@decsai.ugr.es

Francisco Herrera
Data Science and Computational Intelligence Institute
University of Granada
Granada, Spain
herrera@decsai.ugr.es

***Abstract*—This is a summary of our article published in Knowledge and Information Systems [1] to be part of the MultiConference CAEPIA'18 Key Works.**

***Index Terms*—Data preprocessing, non-standard classification, noise data, multiclass classification, multi-instance learning, multi-label learning, multitask problems, ordinal classification, data streams, non-stationary environments**

I. Summary

Learning from noisy data is an important topic in machine learning, data mining and pattern recognition, as real world data sets may suffer from imperfections in data acquisition, transmission, storage, integration and categorization. Indeed, over the last few decades, noisy data has attracted a considerable amount of attention from researchers and practitioners, and the research community has developed numerous techniques and algorithms in order to deal with the issue [2]–[4].

These approaches include the development of learning algorithms which are robust to noise as well as data pre-processing techniques that remove or "repair" noisy instances. Although noise can affect both input and class attributes, class noise is generally considered more harmful to the learning process, and methods for dealing with class noise are becoming more frequent in the literature [3].

Class noise may have many reasons, such as errors or subjectivity in the data labeling process, as well as the use of inadequate information for labeling. For instance, in some medical applications, the true status of some diseases can only be determined by expensive or intrusive procedures, some of which can only be carried out after a patient's death. Another reason is that data labeling by domain experts is generally costly, and several applications use labels which are automatically defined by autonomous taggers (e.g., sentiment analysis polarization [5]), or by non-domain experts. This approach is common in, e.g., social media analysis [5], where hashtags used by users or information provided by a pool of non-domain experts (crowdsourcing) are used to derive labels.

Even though class noise is predominant in the literature (see [2], [6] for recent surveys and comparison studies), most of the research has been focused on noise handling in binary class problems. However, new real-life problems have motivated the development of classification paradigms beyond binary classification [7]. These paradigms include ordinal class [8], multiclass [9], multilabel [10] and multi-instance [11] as well as learning from data streams and non-stationary environments [12] and joint exploiting related tasks [13]. Due to the ubiquity of noise, it is of fundamental importance to better understand the relationships and implications of class noise within these paradigms. Each paradigm has its own particularities which impose new challenges and research questions for noise handling. Although research for class noise handling in these paradigms is somewhat present in the literature, it remains quite scarce and requires general discussion of issues, challenges and research practices regarding it.

The related paper aims to discuss open-ended challenges and future research directions for learning with class noise data, focusing on the aforementioned non-binary classification domains. The main contributions of such a paper are:

- We discuss some current research, as well as the need of adaptation or development of new techniques for handling class noise within non-binary classification paradigms.
- We also discuss issues related to the simulation of noise scenarios (inclusion of artificial noise) within these paradigms, an experimental artifact frequently adopted

This work have been partially supported by the São Paulo State (Brazil) research council FAPESP under project 2015/20606-6, the Spanish Ministry of Science and Technology under project TIN2014-57251-P and the Andalusian Research Plan under project P12-TIC-2958.

for analysis of noise dealing techniques. These issues are important for simulating noise scenarios that may occur in real world applications, and can serve as the basis for uniforming procedures by providing an objective ground in order to assess the robustness of the learning methods.

- We present some important open-ended issues and offer some possible solutions to the existing problems.

We are aware of some studies already considering multiclass noise problems. Different multiclass noise patterns impose numerous challenges, some of them infrequently addressed in the literature. Even state of the art methods for dealing with binary class noise present considerable variation in performance when considering different multiclass noise patterns at the same noise ratio. Despite this, these issued are seldom considered in the literature. In the related paper we focus on some of aspects that could be studied further, providing a guideline of open challenges for researchers, such as:

- Would these different types of noise patterns pose the same or different challenges when dealing with multiple class noise?
- Which one would be more difficult to tackle?
- Which aspects of the problem would be more affected by considering different noise multiclass pattern?
- How do existing methods behave considering these different noise patterns?

One interesting topic for further research is how to extend methods, originally developed only for binary class, to the multiclass case. Some data transformation approaches for transforming multiclass to binary problems, e.g., One-versus-One (OVO) or One-versus-ALL (OVA), could be applied [14]. However, research on this topic generally involves random noise completely at random, with uniform class noise distribution. Investigating how these approaches are affected by different noise patterns is an interesting topic for research. For instance, when applying a filter using a OVA decomposition, does the order in which class noise is removed matter? If so, is this influence stronger for different noise distribution among the classes?

Another open-ended problem is the relationship with imbalanced classification [15] and multiclass noise. It is reported in the literature that noise in minority classes is more harmful than in majority classes [16]. However, multiclass imbalance [17] has further issues to consider, as multiple predominant or infrequent classes may occur. It is unclear what learning difficulties multiclass noise can cause under highly imbalanced class distributions, and how to handle it effectively is an open-ended issue. Furthermore, different noise patterns can change the observed class ratio, which may influence the behavior of class imbalance techniques. Uniform class noise, for instance, may mask the observed class ratio of multiple rare classes even for low noise levels. Default class may also introduce an artificial predominant class, thus generating an artificial imbalanced problem due to the presence of noise. Possible ways to handle noise in imbalanced problems include cost sensitive noise handling [18], [19], attributing and the development of class ratio aware filtering approaches [20] considering the multiclass context.

We believe this discussion will encourage researchers and practitioners to explore the problem of class noise handling in new scenarios and different learning paradigms in more detail.

References

[1] R. C. Prati, J. Luengo, and F. Herrera, “Emerging topics and challenges of learning from noisy data in non-standard classification: A survey beyond binary class noise,” *Knowledge and Information Systems*, vol. in press, 2018.

[2] B. Frénay and M. Verleysen, “Classification in the presence of label noise: A survey,” vol. 25, no. 5, pp. 845–869, 2014.

[3] X. Zhu and X. Wu, “Class noise vs. attribute noise: A quantitative study,” *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 177–210, 2004.

[4] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[5] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[6] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, “A study of the effect of different types of noise on the precision of supervised learning techniques,” *Artif. Intell. Rev.*, vol. 33, no. 4, pp. 275–306, 2010.

[7] J. Hernández-González, I. Inza, and J. A. Lozano, “Weak supervision and other non-standard classification problems: a taxonomy,” *Pattern Recognit. Lett.*, vol. 69, pp. 49–55, 2016.

[8] P. A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernández-Navarro, and C. Hervas-Martinez, “Ordinal regression methods: survey and experimental study,” vol. 28, no. 1, pp. 127–146, 2016.

[9] T. G. Dietterich and G. Bakiri, “Solving multiclass learning problems via error-correcting output codes,” *Journal of artificial intelligence research*, vol. 2, pp. 263–286, 1995.

[10] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Springer, 2016.

[11] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, *Multiple Instance Learning: Foundations and Algorithms*. Springer, 2016.

[12] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, “Learning in nonstationary environments: a survey,” vol. 10, no. 4, pp. 12–25, 2015.

[13] T. Evgeniou, C. A. Micchelli, and M. Pontil, “Learning multiple tasks with kernel methods,” *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 615–637, 2005.

[14] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, “Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition,” *Knowledge and information systems*, vol. 38, no. 1, pp. 179–206, 2014.

[15] R. C. Prati, G. E. A. P. A. Batista, and D. F. Silva, “Class imbalance revisited: a new experimental setup to assess the performance of treatment methods,” *Knowl. Inf. Syst.*, vol. 45, no. 1, pp. 247–270, 2015.

[16] J. Van Hulse and T. Khoshgoftaar, “Knowledge discovery from imbalanced and noisy data,” *Data Knowl. Eng.*, vol. 68, no. 12, pp. 1513–1542, 2009.

[17] S. Wang and X. Yao, “Multiclass imbalance problems: Analysis and potential solutions,” vol. 42, no. 4, pp. 1119–1130, 2012.

[18] X. Zhu and X. Wu, “Cost-guided class noise handling for effective cost-sensitive learning,” in *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2004, pp. 297–304.

[19] X. Zhu, X. Wu, T. M. Khoshgoftaar, and Y. Shi, “An empirical study of the noise impact on cost-sensitive learning.” in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 7, 2007, pp. 1168–1173.

[20] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Inf. Sci.*, vol. 291, pp. 184–203, 2015.