applied sciences

MDPI

Editorial

Machine Learning Methods with Noisy, Incomplete or Small Datasets

Cesar F. Caiafa [1,*], Zhe Sun [2], Toshihisa Tanaka [3], Pere Marti-Puig [4] and Jordi Solé-Casals [4,*]

[1] Instituto Argentino de Radioastronomía—CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina
[2] Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, Wako-Shi 351-0198, Japan; zhe.sun.vk@riken.jp
[3] Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan; tanakt@cc.tuat.ac.jp
[4] Data and Signal Processing Research Group, University of Vic-Central University of Catalonia, 08500 Barcelona, Spain; pere.marti@uvic.cat
* Correspondence: ccaiafa@gmail.com (C.F.C.); jordi.sole@uvic.cat (J.S.-C.)

Abstract: In this article, we present a collection of fifteen novel contributions on machine learning methods with low-quality or imperfect datasets, which were accepted for publication in the special issue "Machine Learning Methods with Noisy, Incomplete or Small Datasets", Applied Sciences (ISSN 2076-3417). These papers provide a variety of novel approaches to real-world machine learning problems where available datasets suffer from imperfections such as missing values, noise or artefacts. Contributions in applied sciences include medical applications, epidemic management tools, methodological work, and industrial applications, among others. We believe that this special issue will bring new ideas for solving this challenging problem, and will provide clear examples of application in real-world scenarios.

Keywords: artificial intelligence; imperfect dataset; imperfect dataset; machine learning

Citation: Caiafa, C.F.; Sun, Z.; Tanaka, T.; Marti-Puig, P.; Solé-Casals, J. Machine Learning Methods with Noisy, Incomplete or Small Datasets. *Appl. Sci.* **2021**, *11*, 4132. https://doi.org/10.3390/app11094132

Received: 20 April 2021
Accepted: 28 April 2021
Published: 30 April 2021

1. Introduction

In many machine learning applications, available datasets are sometimes incomplete, noisy or affected by artifacts. In supervised scenarios, it could happen that label information is of low quality, which includes unbalanced training sets, noisy labels and other problems. Moreover, in practice, it is very common that available data samples are not enough to derive useful supervised or unsupervised classifiers. All these issues are commonly referred as the *low-quality data problem*. Machine learning researchers and practitioners have been working on various strategies to correctly handle the low-quality problem in recent years. Far from being solved, this problem still represents a fundamental and classic challenge in the artificial intelligence community.

The aim of this Special Issue was to collect novel contributions on machine learning methods for low-quality datasets, to contribute to the dissemination of new ideas to solve this challenging problem, and to provide clear examples of application in real scenarios. Despite the COVID-19 crisis and lockdowns in most countries, this Special Issue attracted great attention among researchers worldwide. A total number of twenty-one papers were submitted and fifteen of them were accepted after appropriate revisions. We were pleasantly surprised by the diversity of nationalities of contributors and the variety of the addressed problems in applied sciences ranging from medical and health applications through specific industrial case study examples. The authors of the published papers are from nine countries located in Europe, America, Africa and Asia.

In the following sections, the accepted papers and their corresponding most relevant contributions are summarized, which are grouped in the following categories: medical applications, epidemics management tools, methodological papers, industrial applications, and others.

2. Medical Applications

Interestingly, the majority of the contributions are related to specific applications in medicine. Three papers addressed different problems or diseases in Neuroscience. For example, in [1], Caiafa et al. (Argentina–Spain–Japan) reviewed recent approaches to deal with incomplete or noisy measurements by applying signal decomposition methods and showed their usefulness in epileptic intracranial electroencephalogram (iEEG) signals classification, among other applications. Finding epileptic focus with iEEG is usually difficult mainly because available datasets labeled by expert medical doctors are scarce. In [2], Tong et al. (China–South Africa) proposed a few-shot learning method for the severity assessment of Parkinson's disease based on a small gait dataset. The proposed algorithm solves the small-data problem by using permutation-variable importance (PVI) and persistent entropy of topological imprints; as well as applying a support vector machine (SVM) classifier to achieve the severity classification of Parkinson disease patients. In [3], Wang et al. (China) addressed the problem of small and unbalanced datasets in functional magnetic resonance imaging (fMRI) for neuroscience studies. Their technique combines Independent Component Analysis (ICA) for dimensionality reduction, data augmentation to balance data and a convolution-gated recurrent unit (GRU) network. Results on episodic memory evaluation are reported.

The other papers that addressed medical applications are described as follows. In [4], Yasutomi et al. (Japan) introduced a deep learning method based on an auto-encoder architecture to detect and remove shadow artifacts in ultrasound images. The model can be trained on unlabeled data (unsupervised) or with few pixel labels available (semi supervised). The method has been applied to fetal heart diagnosis. In [5], Ahmad et al. (Saudi Arabia) investigated a machine learning approach to predict diabetes mellitus based on a handful set of features obtained by simple laboratory tests, allowing a cost-effective and rapid screening tool. They compared different machine learning classifiers and provided a set of recommendations based on those analyses. In [6], Qiao et al. (China) proposed a method to measure the length of the root canal length, which is crucial for an effective treatment of endodontics and periapicalitis. The authors employed a neural network on multifrequency impedance measurements.

3. Epidemics Monitoring and Management Tools

Machine learning has been demonstrated to have an important role in dealing with infectious diseases and epidemics. In this collection, two contributions are devoted to the development of tools to deal with some aspects of COVID-19 and dengue epidemics. More specifically, in [7], Gibert Oliveras et al. (Spain) reported the results of a project developed in Catalonia, Spain, owing to help in the COVID-19 crisis. The project allowed for quick territory screening providing relevant information to support informed decision-making and strategy and policy design. The authors proposed a data-driven methodology in order to deal with small subgroups of the population for statistical secrecy preserving. In [8], Silitonga et al. (Indonesia) developed prediction models to estimate the severity level of dengue based on the laboratory test results of the corresponding patients using artificial neural network (ANN) and discriminant analysis (DA) applied to very small datasets.

4. Methodological Articles

Four contributions proposed general methods for machine learning with low-quality datasets. In [1], the authors provided a unified review of decomposition methods, which includes linear decomposition, low-rank matrix/tensor factorization, sparse matrix/tensor decomposition and empirical mode decomposition (EMD) models. This paper illustrates the ability of these decomposition models to impute missing features, denoising and to artificially generate additional data samples (data augmentation) with examples to the brain–computer interface (BCI) and epileptic EEG analysis, among others. In [9], Lee et al. (South Korea) developed feature extraction methods based on the non-negative matrix factorization (NMF) algorithm and it is applied in weakly supervised sound event detection.

The algorithm considers learning from strongly and weakly labeled data. On the other side, in [10], Gil et al. (Spain) investigated the use of optimization in the preprocessing step of time series joining. More specifically, the authors proposed an error function to measure the adequateness of the joining and demonstrated the effectiveness of the proposed method on the synthetical datasets and real industrial process scenario. Finally, in [11], Wang et al. (China–Japan) proposed a novel multi-label feature selection approach by embedding label correlations (dubbed ELCs) in order to eliminate irrelevant and redundant, features, also referred as noisy features.

5. Applications to the Industry

This Special Issue also includes two papers studying the application of machine learning to specific practical problems in different industries: the fishing and smart buildings industries. In [12], Marti-Puig et al. (Spain) addressed the problem of distinguishing between different Mediterranean demersal species of fish that share a remarkably similar form and that are also used for the evaluation of marine resources. The authors employed both a binary and a multi-class classification problem based on very small datasets with unreliable labels. In [13], Ge et al. (Japan–China) proposed a unified and practical framework for knowledge inference inside the smart building.

6. Other Applications

Two very important machine learning problems face recognition and natural language processing. These two problems were addressed in this Special Issue for cases with low-quality datasets. In [14], Lee et al. (Korea) studied the problem of training a facial recognition system provided that only one sample per identity is available. The authors proposed a data augmentation technique by introducing changes in pixels in face images associated with variations by extracting the binary weighted interpolation map (B-WIM) from neutral and variational images in the auxiliary set. In [1], the EMD method was applied to remove noise in face images, thus improving the classification accuracy of a machine learning classifier. Finally, in [15], Mouratidis et al. (Greece) provided an application to natural language processing. They developed a deep learning schema for machine translation evaluation (English–Greek and English–Italian), based on different categories of information (linguistic features, natural language processing metrics and embeddings), by using a model for machine learning based on noisy and small datasets.

7. Conclusions

The correct handling of noisy, incomplete or small datasets remains an open problem in the artificial intelligence community. However, this Special Issue collects fifteen research papers providing general approaches to some low-quality datasets problems and clear practical examples in different applied sciences disciplines. This collection of papers represents a good reference for the current *state-of-the-arts*, also providing an excellent starting point for developing new advanced methods in the future.

Funding: This research received no external funding.

Institutional Review Board Statement: Does not apply.

Informed Consent Statement: Does not apply.

Data Availability Statement: Does not apply.

Conflicts of Interest: The author declares no conflict of interests.

References

1. Caiafa, C.F.; Solé-Casals, J.; Marti-Puig, P.; Zhe, S.; Tanaka, T. Decomposition Methods for Machine Learning with Small, Incomplete or Noisy Datasets. *Appl. Sci.* **2020**, *10*, 8481. [CrossRef]
2. Tong, J.; Zhang, J.; Dong, E.; Du, S. Severity Classification of Parkinson's Disease Based on Permutation-Variable Importance and Persistent Entropy. *Appl. Sci.* **2021**, *11*, 1834. [CrossRef]

3. Wang, S.; Duan, F.; Zhang, M. Convolution-GRU Based on Independent Component Analysis for fMRI Analysis with Small and Imbalanced Samples. *Appl. Sci.* **2020**, *10*, 7465. [CrossRef]
4. Yasutomi, S.; Arakaki, T.; Matsuoka, R.; Sakai, A.; Komatsu, R.; Shozu, K.; Dozen, A.; Machino, H.; Asada, K.; Kaneko, S.; et al. Shadow Estimation for Ultrasound Images Using Auto-Encoding Structures and Synthetic Shadows. *Appl. Sci.* **2021**, *11*, 1127. [CrossRef]
5. Ahmad, H.F.; Mukhtar, H.; Alaqail, H.; Seliaman, M.; Alhumam, A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Appl. Sci.* **2021**, *11*, 1173. [CrossRef]
6. Qiao, X.; Zhang, Z.; Chen, X. Multifrequency Impedance Method Based on Neural Network for Root Canal Length Measurement. *Appl. Sci.* **2020**, *10*, 7430. [CrossRef]
7. Gibert, K.; Angerri, X. The INSESS-COVID19 Project. Evaluating the Impact of the COVID19 in Social Vulnerability While Preserving Privacy of Participants from Minority Subpopulations. *Appl. Sci.* **2021**, *11*, 3110. [CrossRef]
8. Silitonga, P.; Bustamam, A.; Muradi, H.; Mangunwardoyo, W.; Dewi, B.E. Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset. *Appl. Sci.* **2021**, *11*, 943. [CrossRef]
9. Lee, S.; Kim, M.; Shin, S.; Park, S.; Jeong, Y. Data-Dependent Feature Extraction Method Based on Non-Negative Matrix Factorization for Weakly Supervised Domestic Sound Event Detection. *Appl. Sci.* **2021**, *11*, 1040. [CrossRef]
10. Gil, A.; Quartulli, M.; Olaizola, I.G.; Sierra, B. Learning Optimal Time Series Combination and Pre-Processing by Smart Joins. *Appl. Sci.* **2020**, *10*, 6346. [CrossRef]
11. Wang, J.; Xu, Y.; Xu, H.; Sun, Z.; Yang, Z.; Wei, J. An Effective Multi-Label Feature Selection Model Towards Eliminating Noisy Features. *Appl. Sci.* **2020**, *10*, 8093. [CrossRef]
12. Marti-Puig, P.; Manjabacas, A.; Lombarte, A. Automatic Classification of Morphologically Similar Fish Species Using Their Head Contours. *Appl. Sci.* **2020**, *10*, 3408. [CrossRef]
13. Ge, H.; Peng, X.; Koshizuka, N. Applying Knowledge Inference on Event-Conjunction for Automatic Control in Smart Building. *Appl. Sci.* **2021**, *11*, 935. [CrossRef]
14. Lee, Y.; Choi, S.-I. Training Set Enlargement Using Binary Weighted Interpolation Maps for the Single Sample per Person Problem in Face Recognition. *Appl. Sci.* **2020**, *10*, 6659. [CrossRef]
15. Mouratidis, D.; Kermanidis, K.L.; Sosoni, V. Innovatively Fused Deep Learning with Limited Noisy Data for Evaluating Translations from Poor into Rich Morphology. *Appl. Sci.* **2021**, *11*, 639. [CrossRef]