applied sciences

MDPI

Editorial

Special Issue on Data Preprocessing in Pattern Recognition: Recent Progress, Trends and Applications

José Salvador Sánchez [1,*,†] **and Vicente García** [2,†]

[1] Department of Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, Av. de Vicent Sos Baynat s/n, 12071 Castelló de la Plana, Spain
[2] División Multidisciplinaria en Ciudad Universitaria, Universidad Autónoma de Ciudad Juárez, Av. José de Jesús Delgado 18100, Ciudad Juárez 32310, Chihuahua, Mexico
* Correspondence: sanchez@uji.es
† These authors contributed equally to this work.

The availability of rich data sets from several sources poses new opportunities to develop pattern recognition systems in a diverse array of industry, government, health, and academic areas. To reach accurate pattern recognizers on a given task is crucial to prepare the proper raw data set, converting inconsistent data into reliable data. In a pattern recognition project, 80% of the effort is focused on preparing data sets. Therefore, data preprocessing is vital to producing high-quality data and building models with excellent generalization performance. With the main aim is sharing and disseminating the most recent findings on data preprocessing, this Special Issue was launched to be a reference source for researchers, scholars, students, and professionals interested in transforming raw data into a meaningful format.

A total of ten high-quality and peer-reviewed papers form this Special Issue, covering the following topics: class imbalance [1–6], big data preprocessing [1], prototype selection [7,8], variable selection [9] and clustering data on arbitrary shape [10].

When the prior probabilities are unequal in a classification problem, the learning process is always biased towards the predominant classes. Rendon et al. [1] propose to mitigate the unbalance of multi-class big datasets using a hybrid method, conformed by a well-known oversampling technique and a prototype selection method, applied in the artificial neural network's output domain as well as the feature space. Duan et al. [2] propose a two-step solution for two-class problems using a novel classifier ensemble framework based on K-means and the oversampling technique called ADASYIN. Rangel-Díaz-de-la-Vega et al. [3] performed an experimental study on the behavior of four associative classifiers trained on resampled imbalanced credit scoring datasets. Gul et al. [4] deal with the class imbalance problem for a theft electricity detection problem using a five-step framework incorporating several data preprocessing techniques. Guzmán-Ponce et al. [5] propose a two-stage under-sampling technique that combines the DBSCAN clustering algorithm to remove noisy samples and a minimum spanning tree algorithm to face the class imbalance. Rivera et al. [6] develop an architecture of a real-world traffic incident classification system capable of dealing with the imbalance that exists between the classes of traffic incidents and not traffic accidents.

Prototype selection methods have faced noise and high storage requirements, two of the weaknesses affecting the performance of the k-nearest neighbor classifiers. González et al. [7] propose a novel method to simultaneously address the prototype selection and the label-specific feature selection preprocessing techniques using a search method based on evolutionary algorithms that obtain a solution to both problems in a reasonable time. For a string-based space, Valero et al. [8] present the adaptation of the generation-based reduction algorithm that generates a reduced version of the initial dataset.

Homocianu et al. [9] apply different approaches, techniques, and applications for a real-world problem focused on the job satisfaction behavior of Romanian people aged 50.

check for updates

Citation: Sánchez, J.S.; García, V. Special Issue on Data Preprocessing in Pattern Recognition: Recent Progress, Trends and Applications. *Appl. Sci.* **2022**, *12*, 8709. https://doi.org/10.3390/app12178709

Received: 26 August 2022
Accepted: 29 August 2022
Published: 30 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Finally, Niu et al. [10], in order to improve the strength and quality of the clustering task, propose a new ensemble clustering algorithm using multiple k-medoids clustering algorithms.

Funding: This research received no external funding.

Acknowledgments: We would like to express our thanks to all the authors who contributed to this Special Issue. Additionally, we would like to recognize the invaluable work of reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rendón, E.; Alejo, R.; Castorena, C.; Isidro-Ortega, F.J.; Granda-Gutiérrez, E.E. Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem. *Appl. Sci.* **2020**, *10*, 1276. [CrossRef]
2. Duan, H.; Wei, Y.; Liu, P.; Yin, H. A Novel Ensemble Framework Based on K-Means and Resampling for Imbalanced Data. *Appl. Sci.* **2020**, *10*, 1684. [CrossRef]
3. Rangel-Díaz-de-la Vega, A.; Villuendas-Rey, Y.; Yáñez-Márquez, C.; Camacho-Nieto, O.; López-Yáñez, I. Impact of Imbalanced Datasets Preprocessing in the Performance of Associative Classifiers. *Appl. Sci.* **2020**, *10*, 2779. [CrossRef]
4. Gul, H.; Javaid, N.; Ullah, I.; Qamar, A.M.; Afzal, M.K.; Joshi, G.P. Detection of Non-Technical Losses Using SOSTLink and Bidirectional Gated Recurrent Unit to Secure Smart Meters. *Appl. Sci.* **2020**, *10*, 3151. [CrossRef]
5. Guzmán-Ponce, A.; Valdovinos, R.M.; Sánchez, J.S.; Marcial-Romero, J.R. A New Under-Sampling Method to Face Class Overlap and Imbalance. *Appl. Sci.* **2020**, *10*, 5164. [CrossRef]
6. Rivera, G.; Florencia, R.; García, V.; Ruiz, A.; Sánchez-Solís, J.P. News Classification for Identifying Traffic Incident Points in a Spanish-Speaking Country: A Real-World Case Study of Class Imbalance Learning. *Appl. Sci.* **2020**, *10*, 6253. [CrossRef]
7. González, M.; Cano, J.R.; García, S. ProLSFEO-LDL: Prototype Selection and Label- Specific Feature Evolutionary Optimization for Label Distribution Learning. *Appl. Sci.* **2020**, *10*, 3089. [CrossRef]
8. Valero-Mas, J.J.; Castellanos, F.J. Data Reduction in the String Space for Efficient kNN Classification Through Space Partitioning. *Appl. Sci.* **2020**, *10*, 3356. [CrossRef]
9. Homocianu, D.; Plopeanu, A.P.; Florea, N.; Andrieș, A.M. Exploring the Patterns of Job Satisfaction for Individuals Aged 50 and over from Three Historical Regions of Romania. An Inductive Approach with Respect to Triangulation, Cross-Validation and Support for Replication of Results. *Appl. Sci.* **2020**, *10*, 2573. [CrossRef]
10. Niu, H.; Khozouie, N.; Parvin, H.; Alinejad-Rokny, H.; Beheshti, A.; Mahmoudi, M.R. An Ensemble of Locally Reliable Cluster Solutions. *Appl. Sci.* **2020**, *10*, 1891. [CrossRef]