Complex Networks and Machine Learning: From Molecular to Social Sciences

David Quesada [1,2], Maykel Cruz-Monteagudo [3,4,5], Terace Fletcher [1,3], Aliuska Duardo-Sanchez [6] and Humbert González-Díaz [7,8],*

[1] Department of Mathematics, Wolfson Campus, Miami Dade College (MDC), Miami, FL 33132, USA; dquesada@mdc.edu (D.Q.); teFletcher@westcoastuniversity.edu (T.F.)
[2] School of Science, Technology, and Engineering Management, St. Thomas University (STU), Miami, FL 33054, USA
[3] Department of Science, Department of General Education, West Coast University, Miami Campus, Miami, FL 33136, USA; mCruz@westcoastuniversity.edu
[4] Center for Computer Science (CCS), University of Miami, Miami, FL 33136, USA
[5] Department of Chemistry and Biochemistry, University of Porto, 4099-002 Porto, Portugal
[6] University of The Basque Country UPV/EHU, Great Bilbao, Biscay, Basque Country, 48940 Leioa, Spain; aliuska.duardo@ehu.es
[7] Department of Organic Chemistry II, University of the Basque Country (UPV/EHU), 48940 Leioa, Great Bilbao, Biscay, Basque Country, Spain
[8] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Basque Country, Spain
*** Correspondence: humberto.gonzalezdiaz@ehu.es

check for updates

Abstract: Combining complex networks analysis methods with machine learning (ML) algorithms have become a very useful strategy for the study of complex systems in applied sciences. Noteworthy, the structure and function of such systems can be studied and represented through the above-mentioned approaches, which range from small chemical compounds, proteins, metabolic pathways, and other molecular systems, to neuronal synapsis in the brain's cortex, ecosystems, the internet, markets, social networks, program's development in education, social learning, etc. On the other hand, ML algorithms are useful to study large datasets with characteristic features of complex systems. In this context, we decided to launch one special issue focused on the benefits of using ML and complex network analysis (in combination or separately) to study complex systems in applied sciences. The topic of the issue is: Complex Networks and Machine Learning in Applied Sciences. Contributions to this special issue are highlighted below. The present issue is also linked to conference series, MOL2NET International Conference on Multidisciplinary Sciences, ISSN: 2624-5078, MDPI AG, SciForum, Basel, Switzerland. At the same time, the special issue and the conference are hosts for the works published by students/tutors of the USEDAT: USA–Europe Data Analysis Training Worldwide Program.

Combining complex networks analysis methods with machine learning (ML) algorithms have become a very useful strategy for the study of complex systems in applied sciences. Noteworthy, the structure and function of such systems can be studied and represented through the above-mentioned approaches, which range from small chemical compounds, proteins, metabolic pathways, and other molecular systems, to neuronal synapsis in the brain's cortex, ecosystems, the internet, markets, social

networks, program's development in education, social learning, etc. On the other hand, developed algorithms within ML are permitting discover patterns and make predictions among large datasets of potential associations, many of which are characteristic features of complex systems. ML methods include: regression, classification, clustering, dimensionality reduction, ensemble methods, neural networks and deep learning, transfer learning, reinforcement learning, natural language processing, which can be implemented through either artificial neural networks, or support vector machines, or some hybrid versions of them. In addition, descriptors of complex networks at the local and global scales (degree distribution, average degree, diameter of the network, average shortest path, clustering coefficient, connectedness, node centrality, and node influence) can be used as input variables to train ML algorithms in order to predict the properties of these systems. Motivated by the large amount of results and the need of a summary aimed at pointing future unsolved questions, we decided to launch one special issue focused on the benefits of using ML and complex network analysis (in combination or separately) to study complex systems in applied sciences. The topic of the issue is: Complex Networks and Machine Learning in Applied Sciences. Contributions to this special issue are highlighted below.

In *Bortko* et al. [1] contribution "From the Hands of an Early Adopter's Avatar to Virtual Junkyards: Analysis of Virtual Goods' Lifetime Survival" it is pointed out that the measurement and prediction of value creation, distribution, and lifetime is a major goal study in economics, logistics, and business analysis at present. However, the analysis of large data sets about commerce and transactions is an important challenge to be overcome. Therefore, models that predict circulation of goods can be tested and confirmed before their introduction to "real life" and other scenarios. The present study is focused on the characteristics of early-stage adopters for virtual goods, and how they predict the lifespan of the goods. The authors used ML algorithms (including also decision trees) to build up their predictive models. Results provide evidence that the prediction of the lifespan of virtual objects is possible and is based on data from early holders of those objects.

In Audenaert et al. [2] contribution, "Regular Equivalence for Social Networks", they began by mentioning that assigning different roles to the nodes of the graphs used to represent complex network systems of real-life communities and their interactions may help to effectively cluster them in equivalence classes. After that, a novel formal definition of regular equivalence of graph nodes was introduced in this paper. In addition, the authors studied the connection to alternative equivalence types. Lastly, a new algorithm able to detect all regularly equivalent roles in large-scale complex networks was defined. The authors used their algorithm to study Barabási–Albert network models along with social networks.

In Diao et al. [3] contribution, "Data Fusion of Multivariate Time Series: Application to Noisy 12-Lead ECG Signals", the authors mentioned that 12-lead electrocardiograph (ECG) signal fusion is crucial for further ECG signal processing. In this paper, based on the idea of the local weighted linear prediction algorithm, a novel fusion data algorithm was proposed, which was applied in data fusion of the 12-lead ECG signals. In order to analyze comprehensively the signal quality, its characteristics should be adequately retained in the final fused result. In the authors' algorithm, the values for the weighted coefficient of state points were closely related to the final fused result. Thus, two fuzzy inference systems were designed to calculate the weighted coefficients. For the sake of assessing the performance of their method, synthetic ECG signals and realistic ECG signals were applied in the experiments. Experimental results indicated that their method can fuse the 12-lead ECG signals effectively to inherit the quality characteristics of original ECG signals inherited properly.

Mato et al. [4] used ML algorithms known as artificial neural networks (ANN) for detecting patients with depression-related mild cognitive impairment, which is common mostly among elderly people. The authors mentioned, however, that associating late-life depression (LLD) and mild cognitive impairment (MCI) is difficult. They trained an ANN algorithm to classify 96 MCI patients (42 with depression and 54 without) with scores from a neurological examination with high sensitivity and specificity.

Tang et al. [5] presented a multi-view object detection approach based on ML algorithms called Deep Learning. They also evaluated the object retrieval ability and object detection accuracy of both the multi-view and classic ML methods. Multi-view YOLO (You Only Look Once) and Multi-view SSD (Single Shot Multibox Detector) achieved better results than classic ML versions.

Matta et al. [6] examined a scheme for graph-theoretic clustering using node-based resilience measures. Node-based resilience measures optimize an object based on a critical set of nodes whose removal causes some severity of disconnection in the network. Beyond presenting a general framework for the usage of node-based resilience measures for variations of clustering problems, the authors experimentally validated the usefulness of such methods in accomplishing the following: (i) clustering a graph in one step without knowing the number of clusters a priori; (ii) removing noise from noisy data; and (iii) detecting overlapping communities. They demonstrated that this clustering scheme can be applied successfully using a wide range of data, including both real and synthetic networks, both natively in graph form and also expressed as point sets.

Liu and You [7] focused on a roadmap modeling and assessment approach for defense technology system. Advanced defense technology plays a crucial role in safeguarding national security and economic interests. Aiming to handle the problems of current research and development (R and D) management approaches faced with the rocketing complexities of multicomponent systems. The authors proposed a novel roadmap modeling and assessment methodology through studying the driving forces of general technology development and analyzing realistic requirements of defense technology management. First, a requirement decomposition framework was designed based on multi-view theories, and text-mining tools were used to construct a multi-layer knowledge-flow network model. Second, the contribution of required elements at different levels were evaluated using a multi-criteria decision-making approach, and the node importance was assessed based on the topological structure of multi-layer networks. Third, results from the last approach were utilized to demonstrate the effectiveness of the proposed methodologies. It includes examples from technology requirements in maritime security strategy investigation and a dual-layer knowledge-flow network that consists of patents that belong to the "Coherent Light Generator (CLC)" classification from the United States Patent and Trademark Office (USPTO) database and the related academic papers from Web of Science. Finally, the contributions, potential applications, and drawbacks of this work were discussed and research outlooks were provided.

In Chen et al. [8] contribution "Ensemble Classification of Data Streams Based on Attribute Reduction and a Sliding Window", authors address how the increasing volume and dimensionality of data, result in a failure of classification algorithms because they are unable to satisfy the demands of the practical classification applications of data streams. To deal with noise and concept drift in data streams, the authors proposed an ensemble classification algorithm based on attribute reduction and a sliding window in this paper. Using mutual information, an approximate attribute reduction algorithm based on rough sets was used to reduce data dimensionality and increase the diversity of reduced results in the algorithm. A double-threshold concept drift detection method and a three-stage sliding window control strategy were introduced to improve the performance of the algorithm when dealing with both noise and concept drift. The classification precision was further improved by updating the base classifiers and their nonlinear weights. Experiments on synthetic datasets and actual datasets demonstrated the performance of the algorithm in terms of classification precision, memory use, and time efficiency.

De Julian-Ortíz et al. [9] contributed with the manuscript, "Modeling Properties with Artificial Neural Networks and Multilinear Least-Squares Regression: Advantages and Drawbacks of the Two Methods". The mean molecular connectivity indices (MMCI) proposed in previous studies were used in conjunction with well-known molecular connectivity indices (MCI) to model 11 properties of organic solvents. The MMCI and MCI descriptors selected by the stepwise multilinear least-squares (MLS) procedure were used to perform artificial neural network (ANN) computations, with the aim of detecting the advantages and limits of the ANN approach. The MLS procedure can replicate the

obtained results for as long as is needed, a characteristic not shared by the ANN methodology, which, on the one hand, increases the quality of a description, and, on the other hand, also results in overfitting. The present study also revealed how ANN methods prefer MCI relative to MMCI descriptors. Four types of ANN computations showed that: (i) MMCI descriptors are preferred with properties with a small number of points; (ii) MLS is preferred over ANN when the number of ANN weights is similar to the number of regression coefficients; and, (iii) in some cases, the MLS modeling quality is similar to the modeling quality of ANN computations. Both the common training set and an external randomly chosen validation set were used throughout the paper.

Ren et al.'s [10] focus was on small object detection in optical remote sensing images via modified Faster R-CNN. The PASCAL VOC challenge performance has been significantly boosted by the prevalently CNN-based pipelines like Faster R-CNN. However, directly applying the Faster R-CNN to the small remote sensing objects usually renders poor performance. To address this issue, this paper investigated how to modify Faster R-CNN for the task of small object detection in optical remote sensing images. First of all, the authors not only modified the RPN stage of Faster R-CNN by setting appropriate anchors, but also leveraged a single high-level feature map of a fine resolution by designing similar architecture adopting top-down and skip connections. In addition, they incorporated context information to further boost small remote sensing object detection performance while they applied a simple sampling strategy to solve the issue about the imbalanced numbers of images between different classes. At last, they introduced a simple yet effective data augmentation method named "random rotation" during training. Experimental results showed that their modified Faster R-CNN algorithm improved the mean average precision by a large margin on detecting small remote sensing objects.

Lastly, we would like to add that the present issue is also linked to MOL2NET-05, the International Conference on Multidisciplinary Sciences, ISSN: 2624-5078, MDPI AG, SciForum, Basel, Switzerland, 2017, 2018, and 2019 eds. [11,12]. This means that, contributing authors to the conference and the students who participated in the training school were both allowed to submit full versions of their manuscripts to the present special issue. The short communications of the conference have been published by the Sciforum platform, supported by MDPI AG editorial. These communications have been presented online and/or in person in more than 10 associated specialized workshops held at universities in the USA, Spain, Portugal, Brazil, etc. The conference publishes 100–300 communications each year, authored by 200–700 authors from more than 20 countries worldwide. The members of the committee are guest editors of more than 10 issues in MDPI AG journals with JCR impact factor in the range 2–5. Please, visit the different editions of the MOL2NET conference, 2018, and 2019 links: http://sciforum.net/conference/mol2net-04, http://sciforum.net/conference/mol2net-05. At the same time, the special issue and the conference are hosts for the works published by students/tutors of the USEDAT: USA–Europe Data Analysis Training Worldwide Program [13]. This Transatlantic initiative joins PhD/MSc programs, capstone courses, summer schools, or computing boot camps of more than 10 universities and research centers in USA and Europe. This training program focused on Applied Sciences with an emphasis on both the Introduction to Experimental Data Recording (NMR, MS, IR, 2DGE, EEG, etc.) and/or posterior Computational Data Analysis (machine learning, complex networks, etc.). It includes applications on Cheminformatics, Bioinformatics, Medicinal Chemistry, Nanotechnology, Systems Biology, Biomedical Engineering, etc. The school also promotes training and knowledge of ethical and legal regulatory issues in the USA and Europe (GDPR, REACH, OECD, FDA, etc.) about data use and data protection in chemistry and biomedical research. See the USEDAT workshop link: http://sciforum.net/conference/mol2net-05/usedat-07.

References

1. Bortko, K.; Pazura, P.; Hamari, J.; Bartków, P.; Jankowski, J. From the Hands of an Early Adopter's Avatar to Virtual Junkyards: Analysis of Virtual Goods' Lifetime Survival. *Appl. Sci.* **2019**, *9*, 1268. [CrossRef]
2. Audenaert, P.; Colle, D.; Pickavet, M. Regular Equivalence for Social Networks. *Appl. Sci.* **2019**, *9*, 117. [CrossRef]
3. Diao, C.; Wang, B.; Cai, N. Data Fusion of Multivariate Time Series: Application to Noisy 12-Lead ECG Signals. *Appl. Sci.* **2019**, *9*, 105. [CrossRef]
4. Mato-Abad, V.; Jiménez, I.; García-Vázquez, R.; Aldrey, J.M.; Rivero, D.; Cacabelos, P.; Andrade-Garda, J.; Pías-Peleteiro, J.M.; Rodríguez-Yáñez, S. Using Artificial Neural Networks for Identifying Patients with Mild Cognitive Impairment Associated with Depression Using Neuropsychological Test Features. *Appl. Sci.* **2018**, *8*, 1629. [CrossRef]
5. Tang, C.; Ling, Y.; Yang, X.; Jin, W.; Zheng, C. Multi-View Object Detection Based on Deep Learning. *Appl. Sci.* **2018**, *8*, 1423. [CrossRef]
6. Matta, J.; Obafemi-Ajayi, T.; Borwey, J.; Sinha, K.; Wunsch, D.; Ercal, G. Node-Based Resilience Measure Clustering with Applications to Noisy and Overlapping Communities in Complex Networks. *Appl. Sci.* **2018**, *8*, 1307. [CrossRef]
7. Lu, H.; You, H. Roadmap Modeling and Assessment Approach for Defense Technology System of Systems. *Appl. Sci.* **2018**, *8*, 908. [CrossRef]
8. Chen, Y.; Li, O.; Sun, Y.; Li, F. Ensemble Classification of Data Streams Based on Attribute Reduction and a Sliding Window. *Appl. Sci.* **2018**, *8*, 620. [CrossRef]
9. de Julián-Ortiz, J.V.; Pogliani, L.; Besalú, E. Modeling Properties with Artificial Neural Networks and Multilinear Least-Squares Regression: Advantages and Drawbacks of the Two Methods. *Appl. Sci.* **2018**, *8*, 1094. [CrossRef]
10. Ren, Y.; Zhu, C.; Xiao, S. Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]
11. Gonzalez-Diaz, H.; Arrasate, S.; Sotomayor, N.; Lete, E.; Cossío, F.P.; Sylla, M.; Bonchev, D.; Basak, S.C.; Quesada, D.; Moreira, I.; et al. Mol2net: From Molecules to Networks (Conf. Proc. Book). In *MOL2NET 2016, International Conference on Multidisciplinary Sciences*; MDPI AG: Basel, Switzerland, 2019; Volume 4, p. 2985. ISBN 978-3-03842-820-6. ISSN: 2624-5078. [CrossRef]
12. Gonzalez-Diaz, H.; Arrasate, S.; Sotomayor, N.; Lete, E.; Cossío, F.P.; Dominguez, E.; Bonchev, D.; Basak, S.C.; Reitz, A.; Chou, J.J.; et al. Mol2net: From Molecules to Networks (Conf. Proc. Book). In *MOL2NET 2016, International Conference on Multidisciplinary Sciences*; MDPI AG: Basel, Switzerland, 2018; Volume 3, p. 679. ISBN 978-3-03842-819-0. ISSN: 2624-5078. [CrossRef]
13. Leszczynski, J.; Quesada, D.; Muratov, E.; Montemore, M.; Green, J.R.; Basak, S.; Rasulev, B.; Munteanu, C.R.; Pazos, A.; Perez-Riverol, Y.; et al. USEDAT: USA-Europe Data Analysis Training Worldwide Program. In *MOL2NET 2019, International Conference on Multidisciplinary Sciences*; MDPI AG: Basel, Switzerland, 2019; ISSN: 2624-5078. [CrossRef]