Tutorial

"Classification of unbalanced datasets for the detection of rare patterns: stateof the art and current challenges with a special focus on industrial problems"

Valentina Colla, Pd.D., Marco Vannucci, Ph. D.,
Scuola Superiore Sant'Anna, Pisa, Italy
colla@sssup.it,mvannucci@sssup.it

In many practical fields the detection of particular and rare patterns among a waste amount of data is a key issue since such patterns represent interesting situations whose correct detection is fundamental in the problem context. In the industrial field, for instance, machine malfunctions belong to this category of events: they occur rarely with respect to *normal* situation, nevertheless their identification is essential as it could avoid the manufacturing of defective products or repairing costs.

Despite the importance of the detection of the rare patterns, the class unbalance of the training datasets prevents their correct identification through the use of standard and soft-computing techniques due to numerous interacting causes that include the scarcity of samples to be used for their characterization, the complexity of the problem and the presence of noise and outliers. Due to the relevance of the problem in literature it is possible to find numerous works related to the classification of unbalanced dataset by means of the design of new algorithms that directly tackle the class unbalance (internal methods) or by performing pre-processing steps on the training dataset aiming at mitigating its effect and favouring the identification of the rare events (external methods).

In the proposed tutorial the main issues related to the classification of unbalanced datasets will be described and the main factors that affect the performance of standard classifier will be analysed. Subsequently the state-of-the-art of the approaches for coping with this problems will be presented including internal, external and hybrid methods (i.e. a combination of internal and external techniques). Particular attention will be given to emerging methods based on the use of Artificial Neural Networks (ANN) which, due to their characteristics of flexibility, robustness and generalization capabilities, are becoming the leading technology in this framework.

The performance of the introduced approaches will be evaluated on exemplar case studies coming

(mainly) from the industrial field in order to put into evidence strong and weak points of each method in relation to the characteristics of the different considered tasks. Finally the open challenges and future directions in the research on the classification of unbalanced datasets and detection of rare patterns within large datasets will be outlined.

Valentina Collaobtained a Master Degree in Telecommunication Engineer at the University of Pisa in 1994 and a PhD at Scuola Superiore Sant'Anna (SSSA) in Robotics in 1998. She is currently Technical Research Manager at SSSA and she is the responsible of the Center of Information and Communication Technologies for Complex Industrial Systems and Processes (ICT-COISP) of the TeCIP Institute of SSSA. Her research fields include standard and Artificial Intelligence-based data processing, data mining and machine learning tools and techniques. She is deeply involved in research activities related to modeling, simulation, optimization and control of industrial processes, with a particular focus on manufacturing industry and process industry. She is also active in the field of simulation of complex industrial processes and application of multi-objective optimization techniques aimed at improving resource efficiency and reducing production costs and environmental impact of process industries.
She has a considerable experience in the process industry and in manufacturing fields. She has been involved in more than 45 EU funded projects and in many projects supported by industries. She is coordinator of 3 projects supported by the EU through the Research Fund for Coal & Steel.
Marco Vannucci obtained a Master Degree in computer science at the University of Pisa in 2001 and a PhD at Scuola Superiore San'Anna in industrial engineering in 2006. He is currently a research assistant at TeCIP Institute of SSSA since 2006.He has experience in the field of application of Artificial Intelligence techniques for industrial data processing, data mining and intelligent systems for industrial automation. His expertise includes mathematical modelling of complex systems, artificial neural networks and optimization methods. A significant part of his research regards the steel and manufacturing sectors: he participated to more than 20 projects funded by the EU and is involved in several thirs parties collaborations with Italian and European companies. He is co-author of more than 90 papers published in international journals and conferences as well as book chapters. He is reviewer for several journals in the field of artificial intelligence and industrial automation and has experience as proposal evaluator in RFCS TGS9 Factory-wide and environmental control.