Title: “Big Data: Fundamentals and Paradigms. A Case Study on Imbalanced Big Data Classification”

In the era of the information technology, the problem of managing Big Data applications is becoming the main focus of attention in a wide variety of disciplines such as science, business, industry, and so on. Data and the ability to process and extract knowledge from it are the "new gold" in the digital economy in which we move. As a result, it has emerged an area called Data Science. It collects all scenarios in which data has a starring role with the aim of turning it into knowledge. Data Science encompasses the areas known as machine learning, data mining, social mining, Big Data, and so on.

Addressing Big Data becomes a very interesting and challenging task where we must consider new paradigms to develop scalable algorithms. The MapReduce framework, introduced by Google, allows us to carry out the processing of these large amounts of information. Its open source implementation, named Hadoop, led the development of a popular platform with a wide use. Recently, new alternatives to the standard Hadoop-MapReduce framework have arisen to improve the performance in this scenario, being the most relevant one the Apache Spark project. The MapReduce framework implies that existing algorithms have to be redesigned or that new ones need to be developed in order to take advantage of their capabilities in the big data context.

In this tutorial we will first provide a gentle introduction to the problem of Big Data as well as the presentation of recent technologies (Hadoop ecosystem, Spark). Then, we will dive into the field of Big Data analytics, introducing Machine Learning libraries such as Mahout and MLlib.

Afterwards, we will present a case study on imbalanced Big Data classification. The aim is to show a real application that illustrates both the benefits and challenges of this task. Software from spark-packages and MLlib will be provided to ease the replication of the proposed analysis.

Alberto Fernández Bio:

Alberto Fernández received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 2005 and 2010, respectively. He is currently an Assistant Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, Spain.

He has published more than 100 papers in highly rated JCR journals and international conferences, collecting up to 5000 citations (according to Google Scholar, May 2017). He has 6 research works included in the “highly cited papers” list from Web of Science (May 2017). In 2013, 2014 and 2017

Dr. Fernández received the University of Granada Prize for Scientific Excellence Works in the field of Engineering.

His research interests include classification in imbalanced domains, fuzzy rule learning, evolutionary algorithms and evolutionary fuzzy systems, multiclassification problems with ensembles and decomposition techniques, and data science in Big Data applications