Abstract

With the Big Data explosion, companies have the opportunity to access a massive amount of data that can improve their efficiency in terms of decision-making, adopted solutions, customer care, and so on. By conveniently structuring the Knowledge Extraction processes, companies can easily convert information into opportunities. However, in continuously evolving contexts, a significant analysis should be dedicated to the data quality assessment to deal with unreliable information. Furthermore, designed decision-making solutions should be aware of data drift and (re)adapt themselves along their lifecycle.

In this sense, the thesis work proposes Data Mining methodologies that take into account Veracity and Value challenges underlying Big Data. The meaning of Veracity in the context of Big Data concerns with the truthful of a data set and how trustworthy the data source, type, and processing is. However, the Value of Big Data is strictly related to the Veracity (or quality) of treated data. In fact, integrity awareness about data and its sources is crucial if we are trying to extract information from huge amounts of data. Some of the main achievements of this thesis work are summarized following:

- The application of the well-known theory of Formal Concept Analysis and its variants for extracting conceptualization models from different data streams contents (i.e., social media, papers, etc.).

- The definition and experimentation of a method for cross-relating data sources, with different velocity, size, and credibility levels, by joining conceptualization models to support information trustworthiness (i.e., Veracity) and enable an information filtering system.

- The definition and experimentation of a drift-aware deep learning model based on LSTM for adaptively recognizing and distinguishing evolving

energy consumption behaviors pruning the risk of false-positive alarm about frauds.

- The definition of a consistency measure based on Fuzzy Consensus model, a method widely used in Group Decision Making, to support the training data value assessment before applying a machine learning algorithm for extracting a predictive model.

Presented methodologies are supported by the application and experimentation on several real-world application scenarios giving an idea of their applicability and effectiveness. Faced problems include recommendations, anomaly detection, fake news detection, pharmacovigilance, Emergency Department overcrowding, etc.