Chapter 7
Causal Feature Selection

Walisson Ferreira Carvalho
Centro Universitario Una, Brazil

Luis Zarate
Pontificia Universidade Catolica de Minas Gerais, Brazil

ABSTRACT

Feature selection is a process of the data preprocessing task in business intelligence (BI), analytics, and data mining that urges for new methods that can handle with high dimensionality. One alternative that have been researched to deal with the curse of dimensionality is causal feature selection. Causal feature selection is not based on correlation, but the causality relationship among variables. The main goal of this chapter is to present, based on the issues identified on other methods, a new strategy that considers attributes beyond those that compounds the Markov blanket of a node and calculate the causal effect to ensure the causality relationship.

INTRODUCTION

Year after year, the volume of data has proliferated at remarkable speed. However, large volumes and variety of data do not necessarily translate into quality and, due to this exponential growth, researchers are dealing with new challenges on the process of discovering knowledge. These challenges involve: the comprehension and modeling of the problem being considered, that quality of data, and identifying relevant data. One well-known problem is the Curse of Dimensionality. The Curse of Dimensionality is a term presented by Bellman in 1957 to describe a problem caused by an exponential increase in volume, especially complications when it comes to analyzing and organizing data in high-dimensional spaces (Keogh & Mueen, 2017).

The more data is available, the greater the need to analyze it in order transform it into knowledge, and then convert knowledge into information. Three areas of knowledge are currently dealing with this very subject: Business Intelligence (BI), Analytics, and Data Mining.

Business Intelligence can be defined as the process of transforming data into information and, consequently, into knowledge. Analytics can be defined as the process of transforming data into insights.

DOI: 10.4018/978-1-7998-5781-5.ch007

Whereas Data Mining is the process of discovering potentially useful and unknown information from a collection of data. All three processes have the same input: data. Their shared aim to produce information and knowledge to support decisions' makers.

Despite their minor differences, all three processes are dependent of the quality of data, not only on the volume that enters the pipeline. Therefore, quality data is a critical factor of success. This quality of data can be understood from the concept of Smart Data, which refers to the process of transforming raw data into quality data. The process of discovering smart data is defined by the Gartner Group as "a next-generation data discovery capability that provides business users or citizen data scientists with insights from advanced analytics."

It is well known that the pipeline for transforming raw data into knowledge and, consequently, in information (or insights) includes the preprocessing stage. According to Garcıa et al. (2015) preprocessing is the most important stage in data mining and is affected by the volume of data as well. In the event raw data is not ready to be analyzed, it is necessary to prepare it before being processed by learner's model algorithm. The preprocessing phase is responsible for transforming data and includes data cleaning, integration, normalization, and dealing with missing data.

One strategy used during the preprocessing stage is dimensionality reduction, a technique that can be feature extraction, feature selection, or instance selection. Feature extraction is associated with constructing new features as functions of existing ones. Transformation, discretization, and Principal Components Analysis (PCA) are techniques of feature extraction. Meanwhile, feature selection aims to reduce the number of features by selecting the more representative subset of variables in a given problem.

The reduction of dimensionality can also consider attributes and samples in a process known as hybrid partitioning. In other words, the data set can be reduced in terms of column (attributes) or rows (samples). The reduction of sample is known as Instance Selection and is a technique used to select the best subset of examples and naturally improves the performance of the learning's algorithm, but the focus of this chapter is on feature selection because it facilitates the learning task and aims to select the optimal subset of features that best represents a problem.

Triguero et al. (2019) emphasized that data preprocessing is one of the most important stages in the process of transforming data into information and Feature Selection is a data preprocessing strategy that should be applied to mitigate problems in the data pipeline.

Take, for instance, the Analytics' process that, despite of its growth, is still prone to some challenges such as how to handle the amount of data, the lack of quality in data, computational resources, and high dimensionality. Analytics can be classified as Descriptive, Predictive, and Prescriptive. Descriptive is related to historical data. In this preliminary stage, the question to be answered is "What is happening?". Predictive is related to the future, using data from the past to predict the future to answer such questions as "What will happen in the future?". Prescriptive is dedicated to trying to answer the question "What should be done?". In general, applying a satisfactory Analytics process requires having smart data that can answer these questions.

Besides mitigating computational complexity, feature selection results in predictive models that are easier to understand due to the reduced number of attributes. According to Garcıa et al. (2015) feature selection is a family of methods with that have the following immediate positive effects on the data analysis:

1. Improve data quality;
2. Increases performance of Data Mining Algorithms;
3. Makes the results easier to understand.