Abstract.

In recent years, after the challenges of data collection were somehow overcome, the main question now is how to process this huge amount of data. In order to exploit and explore knowledge from this huge amount of data, it is necessary to create and use the appropriate tools and infrastructure.

According to the 5-step model used for the process of knowledge extraction and exploration, the most essential and fundamental part is the preprocessing and preparation of data, the result of which is the quality of the output and the next steps of the model are guaranteed. Preprocess, integrate, merge, convert, and store when there are many data sources, including online (streaming) and offline (non-streaming are issues we face. The purpose of the systematic preprocessing framework is to create a structure to define the types of input sources, the types of data requirements, the types of preprocessing components, the types of data conditions, etc. In fact, through this framework, the user can create and define different and complex structures for their specific needs to determine the preprocessing structure for streaming and non-streaming data sources.

In this dissertation, a systematic framework for preprocessing streaming and non-streaming data, an intelligent algorithm for performing user-defined preprocessing procedures according to the resources available and the conditions and requirements defined by the user is also provided with an adaptation algorithm to enable self-correction and adaptation of the user's preprocessing routine. The smart city and Internet of Things datasets were used to evaluate the proposed framework.

The proposed framework was evaluated from both online and offline perspectives, and in cases where it is possible to compare with other works, comparison methods were performed. Furthermore, the proposed method was evaluated in different scenarios. In other words, using the proposed framework, various applications were implemented in the field of big data to determine the effect of using the proposed framework and the timing component in reducing execution time and reducing the complexity of implementation.