Abstract

In this thesis, our main objective is to develop methods to explore non-linear relationships between features. For this, we have developed a feature relationship index (named MIDI) that can be used to quantify non-linear feature relationships, including the non-monotonous ones. We have also designed three biclustering methods that find biclusters based on non-linear feature relationships. All these methods have been applied to real-world datasets, allowing us to learn the patterns in the data.

This thesis introduces a novel feature relationship index, MIDI (Mutual Information-based Dependence Index), designed to quantify both linear and non-linear relationships between features. MIDI estimates feature dependence using mutual information derived from histogram-based density estimates. It employs an innovative bin length selection method based on the maximal spacing between consecutive data points, ensuring consistent and reliable density estimation. The resulting density estimate is used to compute a normalized mutual information score. MIDI does not assume any specific functional form, making it capable of detecting a wide range of complex relationships. Theoretical properties of MIDI are also discussed, and power statistics demonstrate its effectiveness in identifying non-linear relationships. Additionally, MIDI has been used to develop a novel feature selection method that shows promising results compared to existing techniques across several datasets.

This thesis presents three feature-relationship-based biclustering methods. The first of these is a density-based biclustering method. Unlike many existing density-based biclustering algorithms, which find clusters using spatial proximity, the existence of common high-density regions is used for grouping features by this method. This method is called Connectedness-Based Subspace Clustering (CBSC). In contrast to conventional density-based biclustering algorithms, for CBSC the user does not need to provide the parameter values for density estimation. These values are calculated for each pair of features using the data distribution in the space corresponding to the particular pair of features. Thus, the proposed approach can adapt to different data distributions in two-dimensional spaces, allowing it to find feature relation-based biclusters with greater accuracy. This approach allows CBSC to find biclusters based on non-linear and even non-monotonous relationships between features.

Another method for finding biclusters based on non-linear relationships, named RelDenClu, has also been presented. RelDenClu uses the local variations in the marginal and joint densities of each pair of features to find the subset of observations

that forms the basis of the relation between them. Then it finds the set of features connected by a common set of observations, resulting in a bicluster. This approach allows RelDenClu to adapt to local variations in density, leading to improved performance on synthetic and real-world datasets. It can find biclusters based on feature relations, including non-linear and non-monotonous ones.

In this thesis, a novel parameter-free biclustering method named PF-RelDenBi has also been presented. Like RelDenClu, PF-RelDenBi also uses local variations in marginal and joint densities to find observations that form dense regions for pairs of features. But unlike CBSC and RelDenClu, it is a parameter-free method. The thresholds required to find the biclusters are calculated from the data on the fly. In addition, it uses the feature relationship index called MIDI to find sets of related features from the dense regions obtained using relative density. Thus, it uses MIDI to find biclusters based on feature relationships, including non-linear and non-monotonous relations.

The biclustering methods presented in this thesis have been used to study the relationship between lifestyle factors in different countries and the COVID-19 infection rates in these countries. They have also been used to find clusters in three UCI-ML datasets. The biclustering methods have also been used to enhance UCI-ML datasets in order to perform classification of these UCI-ML datasets with higher accuracy. This is done by generating new features using the biclustering methods.

Broadly, these methods offer a powerful approach for analyzing data from complex systems composed of multiple interacting processes and may serve as a foundation for subsequent causal analysis. As datasets from diverse sources continue to be integrated, further application areas for these techniques are likely to emerge.