Abstract

Pattern recognition primarily involves clustering, classification, and feature selection/dimensionality reduction. The success of clustering and classification heavily depends on data representation, i.e., the features used. More features do not necessarily improve performance, as some may be detrimental. A computational framework for these tasks can be established using statistical approaches, neural networks, and fuzzy set theoretic methods. Classification requires labeled data and falls under supervised learning, whereas clustering is unsupervised. Feature selection/dimensionality reduction can be performed in both frameworks. This thesis primarily addresses unsupervised pattern recognition using a fuzzy set theoretic framework.

Chapter 1 reviews the literature related to the addressed problems. Chapter 2 introduces an unsupervised feature selection method based on regularized weighted fuzzy C-means (WRFCM) clustering. The objective is to select a feature subset that yields a partition matrix similar to that obtained using the full feature set. A novel objective function within the WRFCM framework ensures feature selection while maintaining the FCM-based target partition. The proposed method is evaluated using Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and KM-index, demonstrating its effectiveness in selecting informative features. This is the first known attempt at unsupervised feature selection matching a target partition.

Chapter 3 explores the disparity between clustering in the original feature space and in the kernel space using fuzzy set theoretic methods, specifically KFCM-F (Kernel FCM with prototypes in feature space) and KFCM-K (Kernel FCM with prototypes in kernel space). The study highlights that kernel clustering may impose unintended structures, leading to counterintuitive results. It argues that kernel clustering could be useful only with appropriate kernel parameters, a challenge in unsupervised settings.

Chapter 4 critically examines clustering in transformed spaces, emphasizing that clustering should identify natural subgroups inherent in the data. Transformations should preserve or enhance these structures rather than impose new, irrelevant ones. The study underscores the difficulty of determining the utility of kernel clustering in high-dimensional spaces, proposing Sammon's nonlinear projection for visualization. The discussion extends to the challenges of selecting appropriate kernel parameters and their interaction with clustering algorithms.

Chapter 5 extends the conventional Fuzzy C-Means (FCM) algorithm by incorporating neighborhood information in non-image datasets within Euclidean space. This enhancement improves clustering performance, as demonstrated on synthetic and real datasets.

Chapter 6 revisits feature selection, focusing on synergistic pairs of features, particularly in prostate cancer research. The study addresses the challenge of clustering 80 million gene pairs to extract fuzzy rules representing synergistic relations. A fuzzy-rule-based approach is proposed to identify synergy networks of gene interactions relevant to disease progression. This is the first attempt to apply fuzzy modeling for synergy network discovery. The proposed method is validated on prostate cancer datasets, with results shown to be statistically significant and biologically relevant.

Finally, Chapter 7 discusses the limitations of the proposed methods and outlines potential directions for future research.