Abstract

This thesis explores advanced methodologies for clustering numerical and categorical data, particularly in the context of inherent uncertainties such as ambiguity, vagueness, and noise. With the growing availability of diverse data sources, traditional clustering algorithms face significant challenges in effectively managing the complexities of numerical and categorical dataset. The research highlights the limitations of conventional type-1 fuzzy set theory in addressing these uncertainties and proposes the use of type-2 fuzzy set theory, which incorporates a secondary membership value to better handle fuzziness in pattern recognition.

To address the inadequacies of existing clustering techniques, the thesis introduces a novel ensemble-based rough fuzzy clustering method that integrates rough set and fuzzy set principles. This approach aims to enhance clustering performance by employing multi-phase learning strategies, specifically through Simulated Annealing-based Rough Fuzzy K-Modes and Genetic Algorithm-based Rough Fuzzy K-Modes. These methods optimize clustering by addressing local optima issues and utilizing Random Forest for the allocation of peripheral points to crisp clusters.

Additionally, the thesis presents a semi-supervised clustering technique that leverages credibilistic measures to tackle the challenges of coincident clustering and to accurately identify cluster memberships. This technique is particularly beneficial in scenarios with scarce labelled data, allowing for the effective categorization of unlabelled or uncertain data instances.

The effectiveness of the proposed methodologies is validated through extensive experiments on synthetic and real-world datasets, demonstrating superior performance compared to state-of-the-art techniques. The thesis concludes by summarizing the findings, acknowledging limitations, and suggesting future research directions to further enhance clustering techniques for numerical and categorical data.