UGC/FDS11/E03/24

Title of Project:

(English) Towards Building Trustworthy and Practical Deep Clustering Net

(Chinese) 建構可信賴且實用的深度聚類網絡

Abstract of Research Comprehensible to a Non-specialist

Deep clustering is a powerful machine learning technique that combines deep learning and clustering algorithms to perform unsupervised learning. Unlike the traditional clustering methods that rely on manual feature engineering or shallow representations, deep clustering utilizes deep neural networks to automatically learn intricate feature representations from raw data. By leveraging the hierarchical and non-linear capabilities of deep neural networks, deep clustering can effectively capture complex patterns and structures in high-dimensional and unstructured data, such as images, text, and audio. Deep clustering has been applied in diverse domains, including computer vision, natural language processing, and bioinformatics. Deep clustering enables the discovery of meaningful and coherent groups within the data without relying on labeled information, making it particularly valuable in scenarios where labeled data is limited or unavailable. However, the existing deep clustering methods usually suffer from two significant limitations. First, pseudo labeling, which is commonly used to fine-tune deep clustering networks, introduces the risk of overconfidence where the network's predicted confidence exceeds its actual accuracy. This overconfidence hinders the application of deep neural networks in decision-making systems, such as medical diagnosis. Second, previous methods assume balanced clusters, which is not realistic in real-world scenarios. In this project, we aim to address these limitations and investigate trustworthy and practical deep clustering frameworks. First, to solve the over-confidence problem and build a wellcalibrated unsupervised deep clustering net, we will explore methods to measure network calibration and develop techniques to align the constructed features with the network's output, reducing overconfidence. By establishing a well-calibrated deep clustering net, we can enhance the quality of pseudo labels and improve overall clustering performance. Second, to construct a practical deep clustering net capable of handling cluster imbalance scenarios, we will employ techniques to assess the imbalance level in clustered samples and propose effective approaches to tackle imbalanced clustering. Additionally, we will leverage the representation ability of the foundation model to alleviate challenges associated with cluster imbalance.
The development of a well-calibrated and reliable clustering model will have significant implications for various downstream tasks. In this project, we will focus on two specific tasks. First, we aim to reduce labeling costs by utilizing the well-calibrated unsupervised clustering net to guide the selection of important samples for labeling, thus optimizing

overall labeling cost. Second, we will address the challenge of label noise. By utilizing the deep clustering net as a preprocessing step, we can learn a discriminative representation that is robust to label noise. This project will contribute to fundamental research in unsupervised learning, deep learning, and clustering. It will also have practical applications in various domains such as general data compression and activities related to uncertainty estimation, including financial investment and disease diagnosis.