統計科學研

INSTITUTE (

統計所學術演　中研院統計

學 術 演 講

講　題：Unveiling Progress: Reports from Seed Projects on Robust Statistical Inference for High Dimensional Data and Deep Models

講　者：Dr. Tso-Jung Yen, Dr. Chen-Hsiang Yeang, Dr. Henghsiu Tsai, Dr. Su-Yun Huang, and Dr. Hsin-Chou Yang
（顏佐榕博士、楊振翔博士、蔡恆修博士、陳素雲博士與楊欣洲博士）
（中央研究院統計科學研究所）

時　間：2023年12月4日(星期一)，11:30-12:00

地　點：統計所B1演講廳

Abstract

In this presentation, we will provide a brief overview of our selected publications and notable achievements, with a specific focus on the following two key contributions.

● Fatty liver classification via risk-controlled neural networks trained on grouped ultrasound image data. Ultrasound imaging is a widely used technique for fatty liver screening as it is practically affordable and can be quickly deployed by using suitable devices. When it is applied to a patient, multiple images of the targeted organs are produced. By jointly looking at these images, physicians can deliver a quick diagnosis to the patient. In this paper we propose a machine learning model for fatty liver screening from multiple ultrasound images. This model first extracts features of the ultrasound images by using a pre-trained image encoder. It further yields a summarized embedding of these features by using a graph-based aggregate

encoder. The summarized embedding is used as input for a classifier of fatty liver screening. We trained the machine learning model on an ultrasound image dataset provided by Taiwan Biobank. The results show that the classifier can achieve good performance on fatty liver diagnosis. We also carry out risk control on the machine learning model by constructing conformal prediction sets for the output of the machine learning model. Under the risk control procedure, the machine learning model can further improve its results with high probabilistic guarantees.

- Clustering image data with a fixed embedding. Clustering unlabeled image data using deep neural network (DNN) models is under active investigation. Most existing approaches transform the data through embedding operations and cluster the embedded data, and the embedding is learned to fit the data. In some applications, the embedding model is explicitly given due to the concerns of generalizability, transferability, privacy and security. Despite rapid progress in self-supervised learning, clustering data with a fixed embedding is rarely explored. We propose an Merge & Expand (ME) algorithm to cluster image data using a fixed embedding and a DNN classification model. ME achieves a comparable level of accuracy with some state-of-the-art algorithms. It further demarcates the "clean" and "unclean" images where their geometric relations in the embedded space are compatible and incompatible with their cluster structure respectively. We further exploited the heterogeneity information and modified ME to improve clustering accuracy by introducing the second embedding. Moreover, we gave intuitive explanations about the source of confusion in merging seed regions. To sum up, ME enables users to better understand the relation between geometry of the embedding space and the underlying cluster structure.Unlike existing approaches, we deconvolve bulk-level RNAseq data by several methods and compared their likelihood scores on single-cell RNAseq data.

※ 實體演講，不開放線上視訊。