Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Master thesis project ”Semisupervised machine learning for cell typing”

Background

One of the key problems in medical care is that many patients do not respond to treatment. Bulk analysis of whole tissues or groups of cells from patients have shown that this likely depends on involvement of thousands of genes. However, each tissue or group involves many different cell types. Which cell type and genes are most important to target with drugs? A new technique, single cell RNA-sequencing (scRNA-seq) allows genome wide analysis of mRNAs in individual cell types, one by one. However, to identify the most important cell types we need to develop methods to classify them. In this project, you will develop machine learning methods to classify cell types for medical research. Bulk and scRNA-seq data from known cell types will be used to develop these methods.

Semisupervised learning is widely used when different classes in the data are organized as clusters, i.e. there is a gap between classes. In this situation, unlabeled data can be used together with the labeled data to get improved decision boundaries.

Data

- Single cell data describing gene expressions (thousands of cells, thousands of genes)
- Smaller single cell data with known cell types
- Very small bulk data with known cell types

Aims

- Find an appropriate semisupervised learning approach that can be used to classify unlabeled single cell data based on single cell data with known cell types and/or bulk data
- Apply the selected semisupervised approach to the given single cell data
- Compare the results with other known statistical approaches for cell typing such as RCA [1]
- Study the effect of data denoising on the classification results

Prerequisites

- Good knowledge of Machine learning and Statistics
- Good programming skills

Research Team

- Oleg Sysoev, STIMA, Linköping University
- Mikael Benson, Center for Personalized Medicine, Linköping University.

Contact and application

- Oleg Sysoev, oleg.sysoev@liu.se

References

[1] Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., ... & Wong, M. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, *49*(5), 708.