HPI Hasso Plattner Institut
Digital Engineering
Universität Potsdam

Identifying Cluster Discriminant Features In Cancer Data

Motivation

Cancer is classified according to the affected tissue, such as breast or lung cancer. Interestingly, these cancer types also differ in their molecular profiles, i.e. what genes are expressed and thus affect the metabolic processes in the cell. Research has already ascertained multiple markers, i.e. genes and gene expression profiles, which are directly linked to specific cancer types, e.g. BRCA1 and BRCA2 for breast cancer. Identifying such markers eventually leads to improved treatments.
However, cancer spans up into various subtypes, e.g. non-small cell lung cancer, which again can be characterized by specific markers. Identifying those markers for all cancer types remains a tedious task that requires time and cost intensive experiments and biological expertise. In turn, initiatives like The Cancer Genome Atlas collect vast amounts of genetic cancer data and make it publicly available. These data sets are more and more used for research on efficient computational analysis methods, which mainly concentrates on strategies for the "traditional" analysis procedure of feature selection/extraction and clustering/classification. Identifying discriminant features of existing clusters is kind of a backward analysis, but remains a classical classification problem.

Project Goals

Your task will be to design and implement a technique that identifies markers for given clusters of cancer types. You will use state-of-the-art and/or extended machine learning techniques to analyze genetic cancer data, e.g. gene expression profiles, that have been grouped prior into cancer types. For each cluster, i.e. cancer type, your technique will identify the cluster-discriminant features, e.g. those genes whose expression pattern is unique for the respective cluster. You will work with real-world data that has been curated by experts and labeled prior according to different cancer types, e.g. from The Cancer Genome Atlas. We aim at publishing the results of this master project in a scientific paper.

Technology and Skills

Prior experience with machine learning techniques will be helpful but is not mandatory. Biological expertise is not required; you will learn about the few necessary biological details during the project. The system will be built from scratch; decisions on the applied technologies (programming language, tools, etc.) will be taken in joint consideration.

Contact

You are welcome to visit us in the "Villa" office V-0.01 or reach out to one of the contacts listed below. For further information, we also invite you to an **upfront meeting at room V2.16 on Friday, 21st July, at 10 AM**.

Dr. Matthias Uflacker (matthias.uflacker@hpi.de)
Cindy Perscheid (cindy.perscheid@hpi.de)
Milena Kraus (milena.kraus@hpi.de)