UMEÅ UNIVERSITET

Cancer subtype identification using cluster analysis on high-dimensional omics data

Linda Vidman

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i N460, Naturvetarhuset,
fredagen den 7 februari, kl. 09:15.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Erik Kristiansson,
Institutionen för Matematiska Vetenskaper, Chalmers tekniska högskola och Göteborgs universitet, Göteborg, Sverige.

Department of Mathematics and Mathematical Statstics

Organization
Umeå University
Department of Mathematics
and Mathematical Statistics

Document type
Doctoral thesis

Date of publication
17 January 2020

Author
Linda Vidman

Title
Cancer subtype identification using cluster analysis on high-dimensional omics data

Abstract
Identification and prediction of cancer subtypes are important parts in the development towards personalized medicine. By tailoring treatments, it is possible to decrease unnecessary suffering and reduce costs. Since the introduction of next generation sequencing techniques, the amount of data available for medical research has increased rapidly. The high dimensional omics data produced by various techniques requires statistical methods to transform data into information and knowledge.
All papers in this thesis are related to distinguishing of disease subtypes in patients with cancer using omics data. The high dimension and the complexity of sequencing data from tumor samples makes it necessary to pre—process the data. We carry out comparisons of feature selection methods and clustering methods used for identification of cancer subtypes. In addition, we evaluate the effect that certain characteristics of the data have on the ability to identify cancer subtypes. The results show that no method outperforms the others in all cases and the relative ranking of methods is very dependent on the data. We also show that the benefit of receiving a more homogeneous data by analyzing genders separately can outweigh the possible drawbacks caused by smaller sample sizes. One of the major challenges when dealing with omics data from tumor samples is that the patients are generally a very heterogeneous group. Factors that lead to heterogeneity include age, gender, ethnicity and stage of disease. How big the effect size is for each of these factors might affect the ability to identify the subgroups of interest.
In omics data, the feature space is often large and how many of the features that are informative for the factors of interest will also affect the complexity of the problem. We present a novel clustering approach that can identify different clusters in different subsets of the feature space, which is applied on methylation data to create new potential biomarkers. It is shown that by combining clinical data with methylation data for patients with clear cell renal carcinoma, it is possible to improve the currently used prediction model for disease progression.
Using unsupervised clustering techniques, we identify three molecular subtypes of prostate cancer bone metastases based on gene expression profiles. The robustness of the identified subtypes is confirmed by applying several clustering algorithms with very similar results.

Keywords
Cancer, classification, clustering, feature selection, subtype identification

Language
English

ISBN
978-91-7855-172-9

ISSN
1653-0829

Number of pages
22 + 4 papers