Analyzing How Protein Interaction Networks Improve Classification Performance in Gene Expression Data Analysis

Adrin Jalali and Nico Pfeifer
Department of Computational Biology and Applied Informatics
Max Planck Institute for Informatics
Saarbrücken, Germany

Motivation

Unsupervised two-dimensional cluster analysis of 98 breast tumours [1]

Problem statement:
- Input: Gene expression data
- Output: Prognosis (Poor vs. Good), Metastases
- Goal: Classify samples and find important genes
- Issue: Hard to classify due to large number of features (genes) compared to number of samples \( (\sim 22000 > 98) \)

Method

1. Dual Problem

\[
\min_{w,\alpha} \left\{ \frac{1}{2} ||w||^2 + \frac{1}{2} \sum_{i,j \in \mathcal{E}} \alpha_i \alpha_j y_i y_j w_i w_j \right\}
\]

s.t.
\[
\forall i \in \{1, \ldots, n\} \colon (w_i, \alpha_i) y_i \geq 1
\]

2. Dual SVM modified objective function [3]

\[
\max_{\alpha \geq 0} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j w_i w_j \right\}
\]

s.t.
\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]
\[
\alpha_i \geq 0
\]

Lagrange matrix
\[
B = D - A
\]

Dual to Primal

\[
w = (4 + \beta B)^{-1} \sum_{i} \alpha_i y_i w_i
\]

What we do:
- Reverse engineer the learned machine to extract important genes after using the network information.
- Solve SVM problem for original and transformed data.
- Calculate w for both models.
- Compute for each pair of nodes, for each model:
  \[
  \text{Score}(i, j) = \frac{1}{2} \left( 1 - \frac{d_{ij}}{d_{thres}} \right)
  \]
- Report pairs with highest scores for both trained models.

Results

References and Acknowledgment

Acknowledgments:
Prof. Dr. Dr. Thomas Lengauer, Suntech Nakert, Nora Speicher, Anna Feldmann.

References: