KDD CUP 2001
Task 3 Localization

Hisashi Hayashi
Jun Sese
Shinichi Morishita

Department of Computer Science
University of Tokyo

Overview

Task
- Predict the localization of a given gene in a cell among 15 distinct positions

Data
- Relation table with six categorical attributes
  Essential, Class, Complex, Phenotype, Motif, Chromosome Number
- Interaction matrix listing all the interactions between genes

Challenges
- How to use interactions ?
- How to deal with missing values ?

Characteristic of Dataset

•*Class, Complex, Motif*, and *Interaction* are highly correlated with localization (evaluated by entropy).

• Each attribute however has many missing values.
70% of Class, 50% of Complex, 50% of Motif

• Four attributes together complement each other to fill missing values.
Only 14 among 381 test records are isolated.

The Winning Approach

Examined three approaches:
• Decision tree with correlated association rules
• Boosting correlated association rules
• Nearest neighbor strategy

Nearest neighbor worked best against the training dataset.

The crux was the definition of "neighborhood."

Definition of Neighborhood

Two records *agree on* an attribute *A* iff

A's values of both records are defined and equal.

Example of the Relational Table

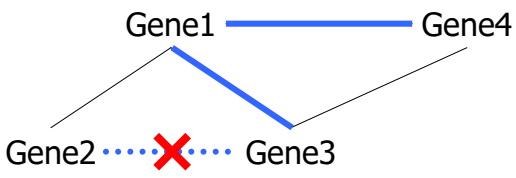
	Complex	Class	Motif
Gene 1	Translocon	actins	?
Gene 2	?	actins	?
Gene 3	Translocon	?	PS00012
Gene 4	Translocon	?	?

Definition of Neighborhood – Cont'd

Two records *agree on* the interaction matrix iff

these records are interacted.

Example of the Interaction Matrix



Definition of Neighborhood – Cont'd

X: a test gene Y: a training gene

If X and Y agree on attribute A,
associate the positive weight of the agreement w_A to A.
Otherwise, $w_A = 0$.

Y is a nearest neighbor of X if Y maximizes the sum of weights;
$w_{Class} + w_{Complex} + w_{Motif} + w_{Interaction}$

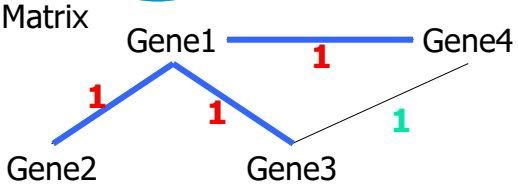
When X and Y agree on all the attributes,
$w_{Complex} >> w_{Class} >> w_{Motif} >> w_{Interaction}$
(ex. 1000 >> 100 >> 10 >> 1)

Nearest Neighbors - Example

The Relational Table

W_A		Complex **1000**	Class **100**	Motif **10**	Sum of Weight
Test	Gene 1	Translocon	actins	?	
Training	Gene 2	?	actins	?	**101**
Training	Gene 3	Translocon	?	PS00012	**1001**
Training	Gene 4	Translocon	?	?	**1001**

The Interaction Matrix



Prediction

1. Given a test gene *X*.

2. Predict the localization of *X*
   by a majority vote
   among the nearest neighbors of *X*.

Conclusion

• Data mining machinery automatically selects biologically meaningful four attributes.

• The step of handling missing values was most elaborated and time-consuming.