Experience with Weka by Predictive Classification on Gene-Expression Data

Matěj Holec

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Intelligent Data Analysis lab
http://ida.felk.cvut.cz

March 17, 2011

Outline

1. Introduction
   - Motivation
   - Biological Background and Data
   - Tools: Weka and R
2. Experiments
   - Integrating Multiple-Platform Expression Data
   - XGENE.ORG
   - Comparative Evaluation of Set-Level Techniques
3. Future Work
4. References
   - Software
   - Bibliography

Outline

1. Introduction
   - Motivation
   - Biological Background and Data
   - Tools: Weka and R
2. Experiments
   - Integrating Multiple-Platform Expression Data
   - XGENE.ORG
   - Comparative Evaluation of Set-Level Techniques
3. Future Work
4. References
   - Software
   - Bibliography

Outline

1. Introduction
   - Motivation
   - Biological Background and Data
   - Tools: Weka and R
2. Experiments
   - Integrating Multiple-Platform Expression Data
   - XGENE.ORG
   - Comparative Evaluation of Set-Level Techniques
3. Future Work
4. References
   - Software
   - Bibliography

Outline

1. Introduction
   - Motivation
   - Biological Background and Data
   - Tools: Weka and R
2. Experiments
   - Integrating Multiple-Platform Expression Data
   - XGENE.ORG
   - Comparative Evaluation of Set-Level Techniques
3. Future Work
4. References
   - Software
   - Bibliography

Motivation

Motivation Bridging the gap between system biology and machine learning.

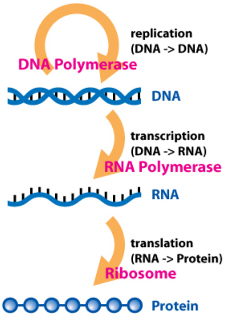
Biological Databases

- NCBI National Center for Biotechnology Information
- EBI European Bioinformatic Institute
- GenomeNet Japanese network of databases and computational services for genome research
- The Gene ontology (GO) vocabulary of terms for describing gene product characteristics and annotation data
- …

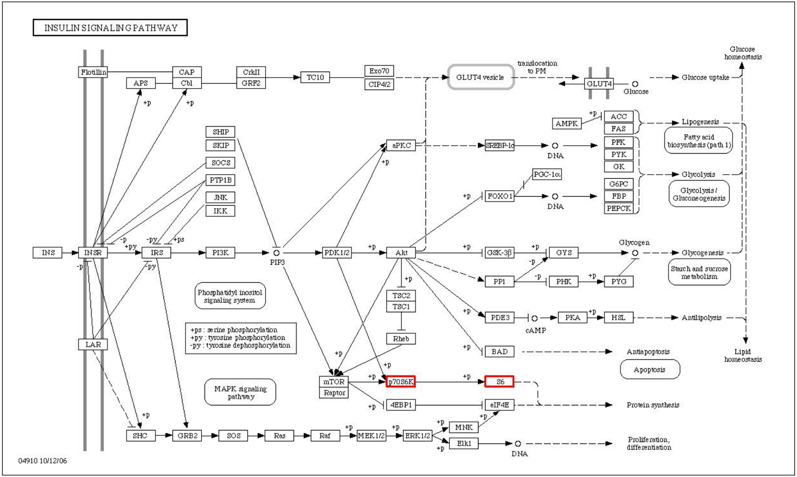
Short Introduction to Biology

- Human cell genome consist of ∼30.000 genes.
- Cell is an integrated device of several thousand types of interacting proteins.
- Cell respond to internal and external environmental signals by producing appropriate proteins.

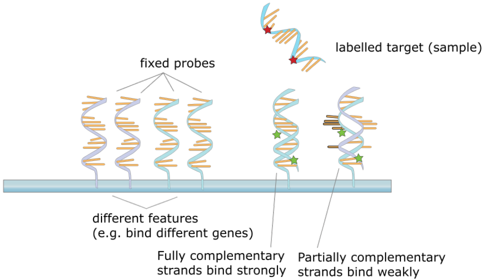
Central dogma of molecular biology

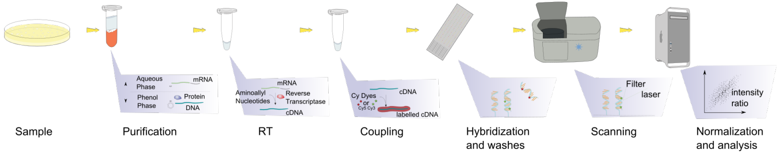


Cellular Pathway and a Fully Coupled Flux Example



DNA Microarrays





Pitfalls of Microarray Technology

- Problem to interpret results (‘Gene list’ syndrome).
- Curse of dimensionality of MA data (tens of thousands genes in tens of samples).
- Noise in microarray data.
- Experiments are still expensive.

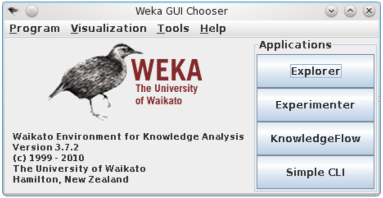
Set-level Approach

- Use prior knowledge

WEKA (Waikato Environment for Knowledge Analysis)

- Machine learning software written in Java
- Licensed under GNU GPL
- Versions: book 3.4.18, stable 3.6.4, developer 3.7.3

Allows data pre-processing, classification, regression, clustering, association rules, visualization


Waikato Environment for Knowledge Analysis
Version 3.7.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Using Weka in Java Code

```
import weka.core.Instances;
import . . . ;
// Input data
DataSource source = new DataSource(”iris.arff”);
Instances instances = source.getDataSet();
. . .
// Create classifier with options
SMO classifier = new SMO();
// train and evaluate the classifier
classifier.buildClassifier(train);
Evaluation eval = new Evaluation(train);
eval.evaluateModel(classifier, test);
// Print summary on the testing instances
System.out.print(eval.toSummaryString());
```

Using Weka in R

```
library(RWeka)
file=”dataset.arff”
splitR=66
instances=read.arff(file)
# shuffle instances
instances=instances[sample(nrow(instances)),]
#get training and testing data
ntrain=round(nrow(instances)*splitR/100)
ntest=nrow(instances)-ntrain
train=instances[1:ntrain,]
test=instances[(ntrain+1):(ntest+ntrain),]
#train and evaluate the classifier
cl=SMO(Class ∼ .,data=train,control = NULL)
evaluate_Weka_classifier(cl,newdata=test)
```

Integrating Multiple-Platform Expression Data through Gene Set Features

Goals:

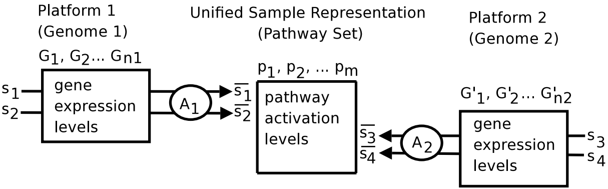
- Integration of data from heterogeneous platforms using gene sets.
- Are the biologically defined gene sets more informative then random gene sets.

Gene set features used for the integration process:

1. Gene ontology terms
2. Cellular pathways
3. Fully coupled fluxes (strongly co-expressed genes)

Integrating Multiple-Platform Expression Data (contd)

1. Preparation (Quantile normalization)
2. Gene set features construction and data integration
3. Analysis by learning curves (Weka Experimenter)



Integrating Multiple-Platform Expression Data Results

(Q1) Single gene based classifiers vs. biologically meaningful gene sets

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

Integrating Multiple-Platform Expression Data Results

(Q1) Single gene based classifiers vs. biologically meaningful gene sets

- Accuracy is not sacrificed by controverting from gene representation of features to the gene-set features.

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

Integrating Multiple-Platform Expression Data Results

(Q1) Single gene based classifiers vs. biologically meaningful gene sets

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

- No of the genuine gene sets strictly outperformed its random counterparts.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

Integrating Multiple-Platform Expression Data Results

(Q1) Single gene based classifiers vs. biologically meaningful gene sets

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

- No of the genuine gene sets strictly outperformed its random counterparts.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

Integrating Multiple-Platform Expression Data Results

(Q1) Single gene based classifiers vs. biologically meaningful gene sets

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

- Assembling of multiple-platform data did not have a detrimental effect on classification performance.

Integrating Multiple-Platform Expression Data Results

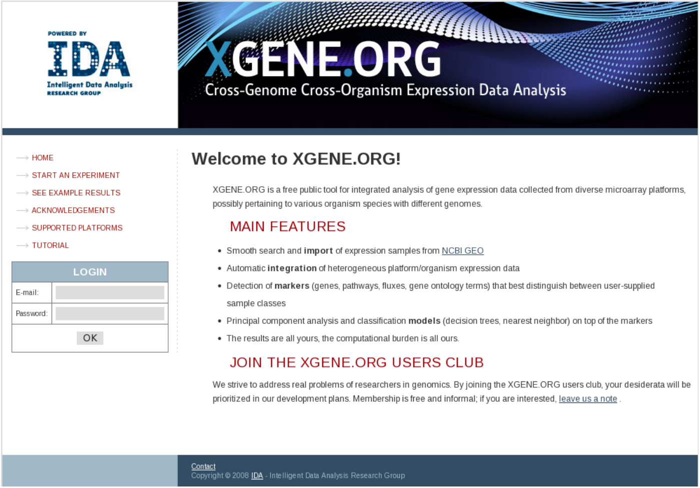
(Q1) Single gene based classifiers vs. biologically meaningful gene sets

(Q2) Classifiers based on the biologically meaningful gene sets vs. based on the gene sets constructed randomly.

(Q3) Classifiers learned from single-platform data vs. learned from the data integrated from heterogeneous platforms

- Assembling of multiple-platform data did not have a detrimental effect on classification performance.

Main Page



Welcome to XGENE.ORG!

XGENE.ORG is a free public tool for integrated analysis of gene expression data collected from diverse microarray platforms, possibly pertaining to various organism species with different genomes.

MAIN FEATURES

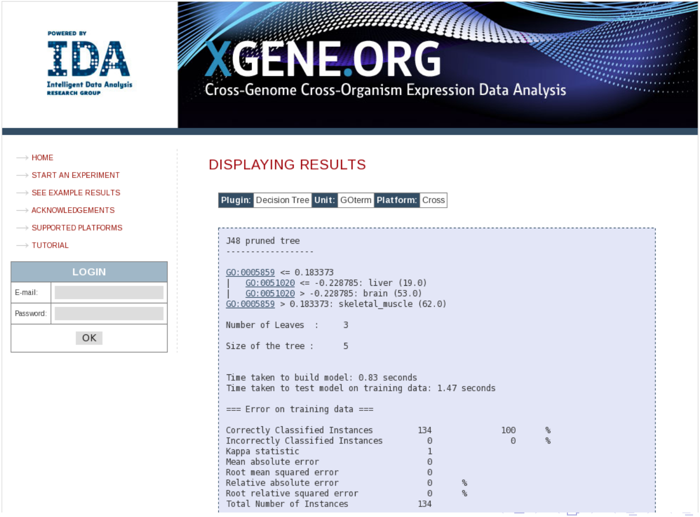
- Smooth search and **import** of expression samples from NCBI GEO
- Automatic **integration** of heterogeneous platform/organism expression data
- Detection of **markers** (genes, pathways, fluxes, gene ontology terms) that best distinguish between user-supplied sample classes
- Principal component analysis and classification **models** (decision trees, nearest neighbor) on top of the markers
- The results are all yours, the computational burden is all ours.

JOIN THE XGENE.ORG USERS CLUB

We strive to address real problems of researchers in genomics. By joining the XGENE.ORG users club, your desiderata will be prioritized in our development plans. Membership is free and informal; if you are interested, leave us a note .

Contact
Copyright © 2008 IDA - Intelligent Data Analysis Research Group

Results



DISPLAYING RESULTS

Plugin: Decision Tree **Unit:** GOterm **Platform:** Cross

```
J48 pruned tree
------------------

GO:0005859 <= 0.183373
|   GO:0051020 <= -0.228785: liver (19.0)
|   GO:0051020 > -0.228785: brain (53.0)
GO:0005859 > 0.183373: skeletal_muscle (62.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.83 seconds
Time taken to test model on training data: 1.47 seconds

=== Error on training data ===

Correctly Classified Instances         134              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              134
```

Short Description

- Web application for cross-genome multiple-platform analysis of gene expression.
- Functionality is done by easy-to-extend plugin system (R, Weka, ...).
- Executes tasks in a grid environment (not working now).

Comparative Evaluation of Set-Level Techniques in Predictive Classification of Gene Expression Samples

- Set-level analysis typically yields more compact and interpretable results.
- Set-level strategy can be adopted by ML algorithms.

Q1 Which one state-of-the-art set-level analysis technique can be used for a better classification.

Q2 How the classification accuracy depends on the functionally defined gene sets in compare to random.

Q3 How accurate are classifiers based on the set-level features in compare to the gene-based.

Experimental settings

Input

- Data

  Microarray experiment data NCBI-GEO
  Functionally defined gene sets (KEGG, KO)

- Algorithms

  Feature selection Globaltest (log. regression), GSEA (Kolmogorov statistic), SAM-GS (Euclidean distance)

  Aggregation avg (average expression), svd (principal component), setsig (transformation using samples class)

Output

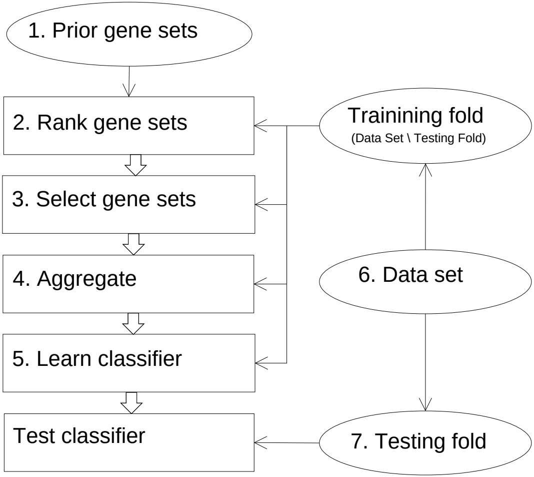
Predictive accuracies on the testing data

Factors

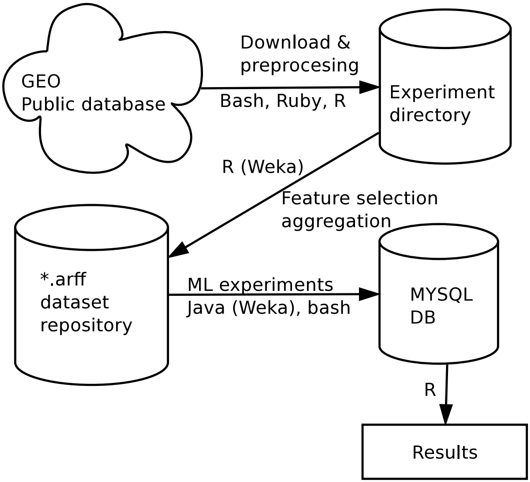
Analyzed factors	*Alternatives*	*#Alts*
1. Gene sets	{genuine, random}	2
2. Ranking algo	{gsea, sam-gs, global, ig}	4
*3. Sets forming features**	$\{1, 2, \ldots 10,$ $n - 9, n - 8, \ldots n,$ $1{:}10, n - 9 : n\}$	22
4. Aggregation	{svd, avg, setsig, none}	4

Auxiliary factors	*Alternatives*	*#Alts*
5. Learning algo	{svm, 1-nn, 3-nn, nb, dt}	5
6. Dataset	$\{d_1 \ldots d_{30}\}$	30
7. Testing Fold	$\{f_1 \ldots f_{10}\}$	10

Data Flow


1. Prior gene sets
2. Rank gene sets
3. Select gene sets
4. Aggregate
5. Learn classifier
Test classifier
Trainining fold
(Data Set \ Testing Fold)
6. Data set
7. Testing fold

Experiment Settings



ML Experiments in Weka – technical summary

- 30 datasets
- 6 Weka algorithms (SMO, J48, 1-NN, 3-NN, NB, ZeroR)
- Total number of ML experiments is 1.470.600
- Speed of Weka experiments execution

$$\frac{30 \times 49020}{105 \times 60} \approx 233[\frac{experiments}{sec}]$$

Analysis

Results were obtained by (two-sided) Wilcoxon test (on level of signif. 0.05, Bonferroni-Dunn adjustment)

Factor	*Alternatives*	
	Better	*Worse*
1. Gene sets	genuine	random
2. Ranking algo	global, ig	sam-gs, gsea
3. Sets forming features	high ranking	low ranking
3. Sets forming features	1:10	1
*4. Aggregation**	setsig, svd	avg

* Difference not significant if Factor 3 is 1:10.

Conclusion

1. Study determined suitability of various set-level methods.
2. Classifiers based on aggregated gene-set features outperform baseline experiments.
3. Gene-set based features allows easier interpretability and data compression.
4. Still are ignored dependencies among gene set members.

Future Work

- XGENE.ORG ver 0.2
  - Support of semiautomatic workflows allowing to define complicated ML tasks.
  - Full support of grid environment.
  - Easy to debug environment (based on Java).
- Experimental analysis of pathway modes (elementary pathways).
- Improve set-level techniques to take into account structural knowledge.

WEKA

WEKA http://www.cs.waikato.ac.nz/ml/weka/

- Documentation http://weka.wikispaces.com/
- Using Weka in Java code http://weka.wikispaces.com/Use+Weka+in+your+Java+code
- Related projects http://www.cs.waikato.ac.nz/ml/weka/index_related.html
- RWeka http://cran.r-project.org/web/packages/RWeka/index.html

R

R `http://www.r-project.org/`

- Bioconductor `http://www.bioconductor.org/`
- RCPP (facilitates integration R and C++)
  `http://dirk.eddelbuettel.com/code/rcpp.html`
  `http://cran.r-project.org/web/.../Rcpp/`

- Set-level analysis
  - Subramanian A et al.: Gene set enrichment analysis: A knowledge-based approach for inter- preting genome-wide expression profiles, *PNAS*, 2005.
  - Jelle Goeman and Peter Buhlmann. Analyzing gene expression data in terms of genesets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007
  - Mramor Minca et al. On utility of gene set signatures in gene expression-based cancer class prediction. *Machine Learning in Systems Biology*, 2010.
- Biological databases
  - The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25, 2000.
  - Minoru Kanehisa et al. KEGG for representation and analysis of molecular networks involving diseases and drugs.*Nucleic acids research*, 38:355–360, 2010

Thank you for your attention