Master’s Thesis Defense:

Automatic Cancer Diagnostic Decision Support System for Gene Expression Domain

Alexander Statnikov

Committee Members:
• Dr. Constantin F. Aliferis (advisor)
• Dr. Douglas P. Hardin
• Dr. Shawn Levy
• Dr. Ioannis Tsamardinos (advisor)

Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
07/06/2005
An automated system for development and evaluation of high-quality cancer diagnostic models and biomarker discovery from microarray gene expression data.
Building Cancer Diagnostic Models from Microarray Gene Expression Data

Clinical Researcher

- samples
- microarray data
- report, dx model

Microarray Lab

- microarray data

Biostatistician

- microarray data, hypothesis

Bioinformatician / Computer Scientist

- experimental design

Programmer

- execution of experiments

- microarray data; plan of experiments

Write-up of the report

Analysis of results

Results, dx model

Weeks or Months
Automated Diagnostic System **GEMS:** Gene Expression Model Selector

- Outputs high quality models for cancer diagnosis from gene expression data;
- Produces reliable performance estimates;
- Allows to apply models to unseen patients;
- Discovers target biomarker candidates.

- Implement best known diagnostic methodologies
- Use sound techniques for model selection & performance estimation

**Clinical Researcher**
- reports, dx model
- microarray data, hypothesis

**Microarray Lab**
- samples
- microarray data

- Week or Months
- Minutes or hours
There exist many good software packages for supervised analysis of microarray data, but...

- **Neither system provides a protocol for data analysis that precludes overfitting.**

- **A typical software either offers an overabundance of algorithms or algorithms with unknown performance. Thus is it not clear to the user how to choose an optimal algorithm for a given data analysis task.**

- **The software packages address needs of experienced analysts. However, there is a need to use this software (and still achieve good results) by users who know little about data analysis (e.g., biologists and clinicians).**
What Does Prior Research Suggest About the Best Performing Methods?

- Limited range of methods & datasets per study
- No description of parameter optimization of learners
- Different experimental designs are employed
- Overfitting [Ntzani et al., Lancet 2003]:
  - 74% no validation
  - 13% incomplete cross-validation
  - 13% implemented cross-validation correctly
- The available meta-analyses are not aimed at identification of best performing methodologies

193 primary studies
2 meta-analyses

Cannot specify a small set of best performing diagnostic algorithms; Have to to perform evaluation *de novo*
Algorithmic Evaluations to Inform Development of the System
Main Goal: Investigate which ones among the many powerful classifiers currently available for gene expression diagnosis perform the best across many datasets and cancer types.

Methods at a Glance

**Classifiers (11)**
- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS
- KNN
- Backprop. NN
- Prob. NN
- Decision Trees
- One-Versus-Rest
- One-Versus-One

**Ensemble Classifiers (7)**
- Majority Voting
- MC-SVM OVR
- MC-SVM OVO
- MC-SVM DAGSVM
- Decision Trees
- Based on MC-SVM outputs
- Based on outputs of all classifiers

**Cross-Validation Designs (2)**
- 10-Fold CV
- LOOCV

**Performance Metrics (2)**
- Accuracy
- RCI

**Statistical Comparison**
- Randomized permutation testing

**Gene Selection Methods (4)**
- S2N One-Versus-Rest
- S2N One-Versus-One
- Non-param. ANOVA
- BW ratio

**Methods at a Glance**

- Decision Trees
- Majority Voting
- Based on MC-SVM outputs
- Based on outputs of all classifiers

- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS
- KNN
- Backprop. NN
- Prob. NN
- Decision Trees

- MC-SVM OVR
- MC-SVM OVO
- MC-SVM DAGSVM
- Decision Trees

- 10-Fold CV
- LOOCV

- Accuracy
- RCI

- Randomized permutation testing

- S2N One-Versus-Rest
- S2N One-Versus-One
- Non-param. ANOVA
- BW ratio
## Datasets Used in Evaluation

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Number of</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Samples</td>
<td>Variables (genes)</td>
</tr>
<tr>
<td>11_Tumors</td>
<td>174</td>
<td>12533</td>
</tr>
<tr>
<td>14_Tumors</td>
<td>308</td>
<td>15009</td>
</tr>
<tr>
<td>9_Tumors</td>
<td>60</td>
<td>5726</td>
</tr>
<tr>
<td>Brain_Tumor1</td>
<td>90</td>
<td>5920</td>
</tr>
<tr>
<td>Brain_Tumor2</td>
<td>50</td>
<td>10367</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>72</td>
<td>5327</td>
</tr>
<tr>
<td>Leukemia2</td>
<td>72</td>
<td>11225</td>
</tr>
<tr>
<td>Lung_Cancer</td>
<td>203</td>
<td>12600</td>
</tr>
<tr>
<td>SRBCT</td>
<td>83</td>
<td>2308</td>
</tr>
<tr>
<td>Prostate_Tumor</td>
<td>102</td>
<td>10509</td>
</tr>
<tr>
<td>DLBCL</td>
<td>77</td>
<td>5469</td>
</tr>
</tbody>
</table>

**Total:**
- 1291 samples
- 74 diagnostic categories
- 41 cancer types and 12 normal tissue types
Performance of Algorithms

Accuracy, %

- OVR
- OVO
- DAGSVM
- WW
- CS
- KNN
- NN
- PNN
- MC-SVM

Categories:
- 9 Tumors
- 14 Tumors
- Brain Tumor1
- Brain Tumor2
- 11 Tumors
- Leukemia1
- Leukemia2
- Lung Cancer
- SRBCT
- Prostate Tumor
- DLBCL
Comparison with Literature

- Multiclass SVMs (this study)
- Multiple specialized classification methods (original primary studies)
Conclusions of Evaluation

• Multi-class SVMs are the best family among the tested algorithms outperforming KNN, NN, PNN, DT, and WV.

• Gene selection in some cases improves classification performance of all classifiers, especially of non-SVM algorithms;

• Ensemble classification does not improve performance;

• Obtained results favorably compare with literature.
Main Goal: Determine feature selection algorithms (applicable to high-dimensional microarray gene expression or mass-spectrometry data) that significantly reduce the number of predictors, maintaining optimal classification performance.

Markov Blanket techniques (e.g., HITON) provide the smallest subsets of predictors that achieve optimal classification performance.
Algorithms Implemented in GEMS

**Classifiers**
- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS

**Gene Selection Methods**
- S2N One-Versus-Rest
- S2N One-Versus-One
- Non-param. ANOVA
- BW ratio
- HITON_MB
- HITON_PC

**Cross-Validation Designs**
- N-Fold Cross-Validation
- LOOCV
- Nested N-Fold Cross-Validation
- Nested LOOCV

**Performance Metrics**
- Accuracy
- RCI
- AUC ROC

**Normalization Techniques**
- \([a, b]\)
- \((x – \text{MEAN}(x)) / \text{STD}(x)\)
- \(x / \text{STD}(x)\)
- \(x / \text{MEAN}(x)\)
- \(x / \text{MEDIAN}(x)\)
- \(x / \text{NORM}(x)\)
- \(x – \text{MEAN}(x)\)
- \(x – \text{MEDIAN}(x)\)
- \(\text{ABS}(x)\)
- \(x + \text{ABS}(x)\)

**Normalization Techniques**
- \([a, b]\)
- \((x – \text{MEAN}(x)) / \text{STD}(x)\)
- \(x / \text{STD}(x)\)
- \(x / \text{MEAN}(x)\)
- \(x / \text{MEDIAN}(x)\)
- \(x / \text{NORM}(x)\)
- \(x – \text{MEAN}(x)\)
- \(x – \text{MEDIAN}(x)\)
- \(\text{ABS}(x)\)
- \(x + \text{ABS}(x)\)
System Description

http://www.gems-system.org
Inputs & Outputs

1. Dataset & outcome or diagnostic labels
2. Optional: gene names and/or accession numbers
3. Various choices of parameters for the analysis (defaults)
4. In application mode: previously saved model

1. Classification model
2. Performance estimate
3. In application mode: the model’s diagnoses/predictions & overall performance
4. A reduced set of genes
5. Links from the genes to literature and other resources.
Performance estimation by Cross-Validation

What if classifier is parametric?

Model selection / parameter optimization by (nested) cross-validation
Software Architecture

Client (Wizard-Like User Interface)

Computational Engine

- Estimate classification performance
- Generate a classification model and estimate its performance
- Generate a classification model
- Apply existing model to a new set of patients

Cross-Validation Loop for Performance Est.
- N-Fold CV
- LOOCV

Cross-Validation Loop for Model Selection
- N-Fold CV
- LOOCV

Performance Computation
- Accuracy
- RCI
- AUC ROC

Normalization

Gene Selection
- S2N One-Versus-Rest
- S2N One-Versus-One
- Non-param. ANOVA
- BW ratio
- HITON PC
- HITON MB

Classification by MC-SVM
- One-Versus-Rest
- One-Versus-One
- DAGSVM
- Method by WW
- Method by CS

Report Generator
User Interface

Step 4/5: Classification algorithm(s)

- SVM classification algorithms (select at least one):
  - One-versus-rest (OVR)
  - One-versus-one (OVO)
  - DAGSVM
  - Method by Weston and Watkins (MW)
  - Method by Grammer and Singal (CS)

- Select kernel for SVM algorithms
  - Polynomial kernel (including linear)
  - Gaussian kernel (RBF)

- SVM parameters
  - No need to optimize, use the following default values:
    - Cost: 100
    - Degree of polynomial: 1
  - Optimize parameters by cross-validation. Search the following grid:
    - Cost: 0.001 to 100 with multiplicative step 10
    - Degree: 1 to 4 with step 1

Project Summary

- Task: Generate a classification model
  - Model file: model.model

Dataset Specification
  - Dataset file: D:\heading\Data\iOR\EPM\SampleData\data.data.txt

Cross-Validation Design
  - Cross-validation design: N-fold cross-validation
  - Number of CV folds: 10
  - Generate CV sample splits: Yes
  - Save CV sample splits: Yes
  - Filename for saving CV sample splits: splits.txt

Normalization
  - Normalization method: x = [0,1]

Classification
  - MC-SVM classification algorithms: OVR, OVO, DAGSVM
  - SVM kernel: Polynomial
  - Optimize SVM parameters: Yes
  - SVM costs: 0.001 to 100, with multiplicative step 10
  - SVM degrees: 1 to 4, step = 1

Variable Selection
  - Variable selection method: Your_SVM_OVR, S2n_OVO
  - Hiton FCHITON_MB
  - Optimize variable selection: Yes
  - Number of variables: 100 to 1000, step = 50
  - Optimize threshold: Yes
  - Threshold: 0.01 to 0.99, step = 0.01

Performance Metric
  - Performance metric: Accuracy

Complexity: 41161 models

Project: Untitled
Number of samples: 83
Number of variables/gene: 2309
Number of categories: 4
Steps in User Interface

- Task selection
- Dataset specification
- Cross-validation design
- Normalization
- Logging
- Performance metric
- Gene selection
- Classification
- Report generation
- Analysis execution
An Evaluation of the System:

• Apply GEMS to datasets not involved in algorithmic evaluation and compare results with ones obtained by human analysts and published in the literature;

• Verify generalizability of models produced by GEMS in cross-dataset applications.

Evaluation Using New Datasets

Datasets

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Number of Samples</th>
<th>Variables (genes)</th>
<th>Categories</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>6_Tumors</td>
<td>353</td>
<td>7069</td>
<td>6</td>
<td>Shedden, 2003</td>
</tr>
<tr>
<td>Leukemia3</td>
<td>248</td>
<td>12135</td>
<td>6</td>
<td>Yeoh, 2002</td>
</tr>
<tr>
<td>Lung_Cancer2</td>
<td>96</td>
<td>7129</td>
<td>2</td>
<td>Beer, 2002</td>
</tr>
<tr>
<td>Lung_Cancer3</td>
<td>181</td>
<td>12533</td>
<td>2</td>
<td>Gordon, 2003</td>
</tr>
<tr>
<td>DLBCL2</td>
<td>210</td>
<td>32404</td>
<td>2</td>
<td>Savage, 2003</td>
</tr>
</tbody>
</table>

Comparison with literature

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>GEMS classification accuracy</th>
<th>Published classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>6_Tumors</td>
<td>97.2%</td>
<td>96.0%</td>
</tr>
<tr>
<td>Leukemia3</td>
<td>98.4%</td>
<td>98.4%</td>
</tr>
<tr>
<td>Lung_Cancer2</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Lung_Cancer3</td>
<td>99.4%</td>
<td>99.3%</td>
</tr>
<tr>
<td>DLBCL2</td>
<td>87.1%</td>
<td>83.9%</td>
</tr>
</tbody>
</table>

Analyzes were completed within 10-30 minutes with GEMS.
## Verify Generalizability of Models in Cross-Dataset Applications

<table>
<thead>
<tr>
<th>Dataset used for construction of a classification model</th>
<th>Performance estimate of the model* (AUC, %)</th>
<th>Dataset used for independent validation of the classification model</th>
<th>Performance on the independent dataset (AUC, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td><strong>Distribution of samples</strong></td>
<td><strong>Author</strong></td>
<td><strong>Distribution of samples</strong></td>
</tr>
<tr>
<td>Lung cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Bhattacharjee</em></td>
<td>186 tumors 17 normals</td>
<td><em>Beer</em></td>
<td>86 tumors 10 normals</td>
</tr>
<tr>
<td>Leukemia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Armstrong</em></td>
<td>24 ALL 28 AML</td>
<td><em>Golub</em></td>
<td>47 ALL 25 AML</td>
</tr>
</tbody>
</table>

*This performance estimate was obtained by nested cross-validation on the dataset used for construction of the model.*
Live Demonstration of GEMS
Scenario 1:

Binary classification model development and evaluation using a lung cancer microarray gene expression dataset.
Live Demo of GEMS (Scenario 1)

Binary classification model development and evaluation

Lung cancer dataset from Bhattacharjee, 2001:

- **Diagnostic task:** Lung cancer vs normal tissues
- **Microarray platform:** Affymetrix U95A
- **Number of oligonucleotides:** 12,600*
- **Number of patients:** 203

User

Various parameters

Classification model

List of genes

Cross-validation performance estimate (AUC)
Scenario 2:

Multicategory classification model development and evaluation using a small round blood cell tumor microarray gene expression dataset.
Live Demo of GEMS (Scenario 2)

Multicategory classification model development and evaluation

Lung cancer dataset from Khan, 2001:

- **Diagnostic task:**
  Ewing Sarcoma vs rhabdomyosarcoma vs Burkitt Lymphoma vs neuroblastoma
- **Microarray platform:**
  cDNA
- **Number of probes:**
  2,308
- **Number of patients:**
  63

Cross-validation performance estimate (AUC)
Scenario 3:

Validating the reproducibility of genes selected in Scenario 1 using another lung cancer microarray gene expression dataset.
Live Demo of GEMS (Scenario 3)
Are selected genes reproducible in another dataset?

Lung cancer dataset from Beer, 2002:
- **Diagnostic task:** Lung cancer vs normal tissues
- **Microarray platform:** Affymetrix HuGeneFL
- **Number of oligonucleotides:** 7,129*
- **Number of patients:** 96

User
Various parameters
Classification model
Cross-validation performance estimate (AUC)
Scenario 4:

Verifying generalizability of the classification model produced in Scenario 1 using another lung cancer microarray gene expression dataset.
Live Demo of GEMS (Scenario 4)

Is constructed classification model generalizable in another microarray dataset?

Classification model
From Scenario 1 (Bhattacharjee’s data)

Lung cancer dataset from Beer, 2002:
- **Diagnostic task:** Lung cancer vs normal tissues
- **Microarray platform:** Affymetrix HuGeneFL
- **Number of oligonucleotides:** 7,129*
- **Number of patients:** 96

Various parameters
User

Performance estimate (AUC)
**GEMS in a Nutshell**

1. The system is fully automated, yet provides many optional features for the seasoned analyst.

2. The system is based on a nested cross-validation design that avoids overfitting.

3. GEMS’s algorithms were chosen after the two extensive algorithmic evaluations.

4. After the system was built, it was validated in cross-dataset applications and also using new datasets.

5. GEMS has an intuitive wizard-like user interface which abstracts data analysis process.

6. GEMS possesses a convenient client-server architecture.
Acknowledgements

Members of my MS Committee:
• Dr. Constantin F. Aliferis (advisor)
• Dr. Douglas P. Hardin
• Dr. Shawn Levy
• Dr. Ioannis Tsamardinos (advisor)

• Yerbolat Dosbayev
• Vanderbilt University Department of Biomedical Informatics faculty and students

NIH grants for funding of this project:
• R01 LM007948-01
• P20 LM007613-01
References

Journal Papers:


Papers in Conference Proceedings (peer-reviewed):


Posters in Conference Proceedings (peer-reviewed):


Software Demonstrations in Conference Proceedings (peer-reviewed):


• Statnikov A, Tsamardinos I, Aliferis CF. Using the GEMS System for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *12th National Conference on Artificial Intelligence (AAAI)*, 2005.