Using the GEMS System for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data

Alexander Statnikov, Ioannis Tsamardinos, Constantin F. Aliferis

Discovery System Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232
{alexander.statnikov, ioannis.tsamardinos, constantin.aliferis}@vanderbilt.edu

Abstract

We will demonstrate the GEMS system for automated development and evaluation of high-quality cancer diagnostic models and biomarker discovery from microarray gene expression data. The development of GEMS was informed by the results of an extensive algorithmic evaluation using 11 microarray datasets. The system was further evaluated in two cross-dataset applications and using 5 microarray datasets. The performance of models produced by GEMS is comparable or better than the results obtained by human analysts, and these models generalize well to independent samples in cross-dataset applications. The system is freely available for download from http://www.gems-system.org for non-commercial use.

Introduction

Development of cancer diagnostic models and biomarker discovery from microarray gene expression data is of great importance in bioinformatics and medicine. Currently, building of cancer diagnostic models from gene expression data has at least three challenging components: collection of samples, assaying, and statistical analysis. A typical statistical analysis process takes from a few weeks to several months and involves many specialists: clinical researchers, statisticians, bioinformaticians, and programmers. As a result, statistical analysis is a serious bottleneck in the development of cancer decision support models, and its enhancement by an automated system will benefit research significantly. To this end, we have built a system called GEMS (Gene Expression Model Selector) for the automated development and evaluation of high-quality cancer diagnostic models and biomarker discovery from gene expression data (Statnikov et al. 2004).

System Description

Given a microarray gene expression dataset as input, GEMS constructs in a supervised fashion classification models that can be used for cancer detection and determination of correct disease subtype. During construction of these models, GEMS allows selection of a subset of genes of minimal size that are as good as or better than the full gene set for the diagnosis. The selection of biomarkers or genes is also useful for discovery purposes, since they suggest plausible causes and treatments of various types of cancer. Finally, GEMS provides estimates of the models' performance in future applications (i.e., when applied to patients not used to build the models but who come from the same patient population) and readily allows users to apply the models to individual patients and other datasets.

We implemented in GEMS only the best-performing methodologies according to the results of an extensive algorithmic evaluation using 11 publicly available cancer microarray datasets with the total of 74 diagnostic categories and 1291 patients (Statnikov et al. 2005). In addition to popular univariate gene selection algorithms, the system also implements two multivariate techniques HITON_PC and HITON_MB which return the set of parents and children (HITON_PC) and Markov blanket (HITON_MB) variables in the causal graph (Aliferis et al. 2003). The above two algorithms possess well-defined properties, theoretical guarantees for correctness and excellent empirical performance. The algorithms currently implemented in the system are shown in Figure 1.

In a preliminary evaluation of the system with 5 cancer gene expression datasets (with the total of 1088 patients) not employed for the algorithmic comparison, GEMS completed the analysis of each dataset within 10-30 minutes (on a standard PC with Intel Pentium-IV 2.4 GHz CPU) and the output model performed as well as or better than previously published models obtained by human analysts. Also, we used this system to perform cross-dataset analysis of cancer diagnostic models using two pairs of different datasets corresponding to two different diagnostic tasks. We found that the diagnostic models obtained by GEMS in one dataset generalize well to data from a different laboratory and that nested cross-validation performance estimates well approximate the error obtained by the independent validation.

GEMS provides an intuitive wizard-like user interface abstracting the microarray data analysis process and not requiring users to be experts in data analysis (Figure 2). To guide the user's choices according to the available

- **Model selection & performance estimation**
    - N-fold cross-validation
    - Leave-one-out cross-validation
    - Nested N-fold cross-validation
    - Nested leave-one-out cross-validation
- **Classification**

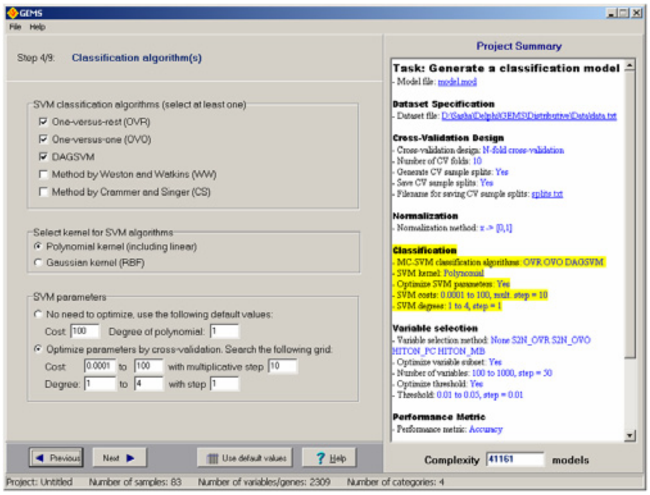
  Multicategory Support Vector Machines:
    - One-versus-rest
    - One-versus-one
    - DAGSVM
    - Method by Crammer and Singer
    - Method by Weston and Watkins
- **Normalization/Rescaling:** 11 methods
- **Gene selection**

  Univariate:
    - Kruskal-Wallis non-parametric ANOVA
    - Signal-to-noise ratio: one-versus-rest
    - Signal-to-noise ratio: one-versus-one
    - Ratio of genes between-categories to within-categories sum of squares

  Multivariate:
    - HITON_PC
    - HITON_MB
- **Performance metrics**
    - Accuracy
    - Relative classifier information
    - Area under ROC curve

Figure 1: Algorithms currently implemented in GEMS.

computational power and time, the system outputs the number of models to be generated for parameters specified by the user. Each step in the interface consists of a form with options for the specific analysis stage. Depending on the task selected by the user, which can either be (1) generate a classification model, or (2) estimate performance of a classification model, or (3) perform tasks (1) and (2) simultaneously, or (4) apply existing classification model to new data, the system will provide a specific sequence of steps applicable for that task. The steps corresponding to generation of a classification model (task 1) are shown in Figure 3.



Step 4/9: Classification algorithm(s)

SVM classification algorithms (select at least one)

☑ One-versus-rest (OVR)
☑ One-versus-one (OVO)
☑ DAGSVM
☐ Method by Weston and Watkins (WW)
☐ Method by Crammer and Singer (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 0.0001 to 100 with multiplicative step 10
Degree: 1 to 4 with step 1

Project Summary

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

Dataset Specification
- Dataset file: D:\Sasha\Delphi\GEMS\Distributive\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: 0.0001 to 100, mult. step = 10
- SVM degrees: 1 to 4, step = 1

Variable selection
- Variable selection method: None S2N_OVR S2N_OVO HITON_PC HITON_MB
- Optimize variable subset: Yes
- Number of variables: 100 to 1000, step = 50
- Optimize threshold: Yes
- Threshold: 0.01 to 0.05, step = 0.01

Performance Metric
- Performance metric: Accuracy

Complexity 41161 models

Project: Untitled Number of samples: 83 Number of variables/genes: 2309 Number of categories: 4

Figure 2: Screen-shot of GEMS. The left part of the screen contains options for the current analysis step (classification algorithm). The summary of the entire project is shown in the right part of the screen.

- overall task selection	- gene selection
- dataset specification	- performance estimation
- cross-validation design	- logging
- normalization	- report generation
- classification	- execution of analysis

Figure 3: User interface steps corresponding to generation of a classification model (task 1).

The system implements a client-server architecture and is made of a computational engine (server) and an interface client. The computational engine is separated from the client and incorporates functional units corresponding to different aspects of analysis. The current version of GEMS runs on MS Windows platforms.

Conclusion

We demonstrate the system GEMS for automated development and evaluation of supervised classification models and biomarker discovery (variable selection) from microarray gene expression data. The system has the following distinguishing features which make it competitive to other available systems for supervised analysis of microarray data: (1) GEMS's classification algorithms were chosen after an extensive algorithmic evaluation, (2) After the system was build, it was validated in cross-dataset applications and also using new datasets, (3) GEMS implements state-of-the-art gene selection and causal discovery algorithms, (4) The system is based on a nested cross-validation design that avoids overfitting, (5) GEMS has intuitive wizard-like interface which abstracts data analysis process, (6) The system is fully automated, yet provides many optional features for the seasoned analyst, and (7) GEMS possesses a client-server architecture.

Acknowledgements

This research was supported by NIH grants RO1 LM007948-01 and P20 LM 007613-01. The authors wish to thank Yerbolat Dosbayev for his contribution to the user interface of GEMS.

References

Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc. 2003:21-5.

Statnikov A, Aliferis CF, Tsamardinos I. Methods for Multi-category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development. Medinfo. 2004;2004:813-7.

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005 21: 631-643.