

# On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection

See-Kiong Ng<sup>1</sup>

skng@i2r.a-star.edu.sg

Soon-Heng Tan<sup>1</sup>

soonheng@i2r.a-star.edu.sg

V.S. Sundararajan<sup>1,2</sup>

sundar@i2r.a-star.edu.sg

<sup>1</sup> Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

<sup>2</sup> School of Computing, National University of Singapore, Lower Kent Ridge Road, Singapore 119260

## Abstract

As microarray technologies become routinely applied in genome laboratories for studying gene expression, it is not uncommon that experiments on identical or similar sets of genes are conducted by multiple laboratories for various functional studies of these genes. Much of such data are often available to researchers for their data analysis, either through collaborators or from online gene expression databases. It will be useful to combine data from different microarray studies to improve the microarray data mining results.

We show that the functional classification of genes from microarray data can be improved further by combining gene expression data from multiple microarray studies, even if the experimental focus or conditions for each experimental study may differ. However, blindly combining all available datasets may not always improve the analysis results—it is important to be selective of the datasets for inclusion. In our approach, we consider each dataset to be one feature, and then apply feature selection strategies to select appropriate datasets for training. With a simple hill-climbing method, we show that gene classification performances can be improved by whole-dataset feature selection.

**Keywords:** microarray, functional classification, multiple datasets, feature selection, support vector machines, multi-layer perceptrons

## 1 Introduction

Increasingly accessible microarray platforms have now allowed routine functional study of genes by microarray technologies in genome laboratories. This has resulted in the generation of many large gene expression datasets. In fact, it is not uncommon that microarray experiments on identical or similar sets of genes are repeatedly conducted by various laboratories for different functional studies of these genes. As such, multiple sets of microarray data on the same set of genes can often be collected from different laboratories and research centers, either through collaborators or from online gene expression data repositories. It will be useful if we can effectively combine these additional datasets with the data generated in one's laboratory to further improve our microarray data mining results.

Although many of the microarray experiments may have been conducted on identical sets of genes, the studies are often designed to address different scientific and experimental investigations, usually conducted under varying experimental conditions. For example, one microarray experiment may be focused on identifying new components in polyphosphate metabolism using the gene knockout method such as [11], while another similar microarray experiment on the same set of genes can be designed

to study spore morphogenesis by times series investigation, such as [3]. Intuitively, it should be beneficial to combine the two expression datasets for microarray data analysis, given that they have been conducted on the same set of genes (both cited experiments used the *Saccharomyces cerevisiae*'s genome in their investigations). On the other hand, their differences in the study objectives and experimental conditions may not warrant that combining data from these two different studies can improve the data mining results.

In this paper, we will show—in the case of functional analysis of genes by microarray data mining—our intuition that combining data from multiple experimental studies can improve data mining results is correct, even in the case where the scientific focus and experimental conditions of the individual microarray studies differ from one another. However, we will also show that blindly combining all available microarray data from different studies in a naive way does not always lead to the best microarray data mining results. The inclusion of additional data in certain combinations can worsen the data mining results, as we will see in Section 6. It is therefore important to be selective in the inclusion of datasets for data analysis. In our work, we consider the entire dataset from each study to be one feature. We then devise a whole-dataset feature selection method to decide on the appropriate microarray datasets to be combined for improved functional analysis. We use a simple hill-climbing method for whole-dataset feature selection, and show that it can better improve the data analysis results from multiple microarray datasets.

In Section 2, we describe the background of functional analysis of gene expression data. Then, in Sections 3 and 4, we report our evaluation study for investigating whether combining microarray data from multiple experimental studies will improve functional analysis results, and whether blindly combining all available datasets will lead to the best possible data mining results. Having shown that the latter is not always true, we propose a whole-dataset feature selection method in Section 5 for choosing appropriate datasets for inclusion in microarray data mining. In Section 6, we present positive results on our whole-dataset feature selection process. Finally, in Section 7, we conclude with discussion on further issues regarding functional analysis of multiple microarray datasets.

## 2 Background

Our living cell is a complicated system comprising multiple cellular pathways performing different biological functions dynamically. Through genome-wide measurements of mRNA expression levels across multiple experimental conditions, we can obtain global snapshots of the cell's genetic activities at various stages and in different conditions. We can then use these gene expression data to elucidate the functional roles of the various genes as they partake in the underlying biological pathways.

One common approach in functional analysis of gene expression data is *clustering*—organizing genes into different functional groups based on the principle that genes belonging to the same functional groups or pathways will have similar expression profiles over a range of experimental conditions. One major drawback of clustering approaches is that classification is learned directly from the expression data [2, 5] without taking advantage of the often available predefined classification information. As a result, clustering approaches can generate clusters of genes that do not correspond well to the true underlying biological pathways.

Biologists often already knew a subset of genes involved in a biological pathway of interest and wish to discover other genes that can be assigned to the same pathway. As such, the *classification* approach is more suitable than clustering for the functional classification of genes using microarray data. Unlike clustering, classification can learn to classify new genes based on predefined classes, taking advantage of the domain knowledge already possessed by the biologists. As such, supervised classification learning algorithms tend to assign pathway memberships that correspond well to the true underlying biological pathways.

Supervised machine learning algorithms such as neural networks, support vector machines, naive bayes, and decision tree methods have been shown to be useful for microarray data analysis in gene

clustering [11, 13, 17] and classification [1, 6, 7, 8, 15]. Most of the previous works have focused on the mining of microarray data from individual experimental studies. Those that have used microarray data from multiple experimental studies—such as Brown *et al.* [1] and Mateos *et al.* [8]—generally included all the available datasets unselectively in their learning procedures. We will show in this paper that blindly combining all available datasets is not guaranteed to improve classification results. To achieve the best results from multiple microarray datasets, it is important to be selective in including datasets from different microarray studies for combined data mining. In this paper, our approach is to consider each study’s dataset as one “feature”, so that feature selection approaches can then be applied to choose the appropriate experimental datasets for combined analysis.

### 3 Materials and Evaluation

As a start, we perform an evaluation study to investigate (a) whether combining data from multiple microarray studies can improve the data mining results, and (b) whether blindly combining all available microarray data will lead to the best data mining results possible.

#### 3.1 Gene Expression Datasets

For our evaluation, we use the gene expression data of *Saccharomyces cerevisiae* from six different microarray studies available from Eisen’s Lab [5] at <http://rana.lbl.gov/EisenData.htm>. These six microarray studies have been performed on the same set of genes from yeast, but with different experimental objectives and under varying experimental conditions. Table 1 below shows the major differences between the six datasets *Alp* [14], *Cdc* [14], *Elu* [14], *Ccc* [14], *Spo* [3], and *Dia* [4]. Collectively, the six datasets comprise expression vectors from a total of 80 experiments on 6,221 yeast ORFs. Out of the 6,221 genes used in the experiments, 2,550 are known yeast genes with annotated functions by MIPS (Munich Information Centre for Protein Sequences) [9].

Table 1: Gene expression datasets used in our evaluation study. Microarray data from six different gene expression studies on *Saccharomyces cerevisiae* were selected for our evaluation of gene functional classification.

Study	Experimental condition	Experimental objective	Number of experiments
<i>Alp</i>	$\alpha$ factor-based synchronization	cell cycle	18
<i>Cdc</i>	<i>Cdc15</i> -based synchronization	cell cycle	25
<i>Elu</i>	elutriation synchronization	cell cycle	14
<i>Ccc</i>	<i>Cln3</i> and <i>Clb2</i> experiments	cell cycle	3
<i>Spo</i>	nitrogen deficiency	spore morphogenesis	13
<i>Dia</i>	glucose depletion	diauxic shift	7

#### 3.2 Functional Assignments

For the known functional classification of the 2,550 annotated genes, we refer to functional assignments provided with the Eisen’s data [5] which was based on the Comprehensive Yeast Genome Database (CYGD) [9] from MIPS. The CYGD is a yeast gene annotation database that has been compiled based on extensive knowledge in the literature.

For comparison, we focus on the five different MIPS classes that both Brown *et al.* [1] and Mateos *et al.* [8] had analyzed previously. While many functional classes could be unlearnable [8], these five functional classes have been proven to be machine-learnable by several previous studies [1, 5,

8]. Biologically, they represent categories of genes expected to exhibit similar expression profiles on biological grounds—making them also challenging cases for machine classification. The five classes are shown in Table 2.

Table 2: Functional classes used for our evaluation study. Five functional classes from the MIPS *Comprehensive Yeast Genome Database* covering a total of 219 yeast genes were used for our comparative evaluation.

Function	Description	Number of genes
TCA	Tricarboxylic acid cycle	22
Resp	genes in respiratory processes	24
Ribo	ribosomal genes	129
Prot	genes of the proteasome	33
Hist	histone-related genes	11

For a more comprehensive study, we also apply our method on all the MIPS-annotated yeast genes in non-singleton functional classes (i.e., functional classes with more than one genes). Unlike previous similar studies such as the study by Mateos *et al.*, we have chosen in our analysis here to exclude genes with ambiguous functional assignments—namely, genes that belong to multiple functional classes—as we have observed that the inclusion of such genes in the training process can affect the results, causing deterioration of the classifiers learned (data not shown). Out of the 2,550 annotated yeast genes in our expression datasets, there are 1,851 genes unambiguously assigned to a total of 60 non-singleton MIPS functional classes and available for our comprehensive evaluation study.

### 3.3 Data Mining Algorithms

The use of support vector machines (SVM) and multi-layer perceptrons (MLP) in gene classification have been investigated in details by Brown *et al.* and Mateos *et al.* in their previous works in [1] and [8] respectively. Together, their results showed that SVM far outperformed other data mining algorithms (including MLP) in the functional classification of yeast genes based on the gene expression datasets described above. As such, we focus primarily on SVM in our study here.

For SVM, the classification performance is highly dependent on the settings of tuning parameters such as the regularization parameter and the kernel parameter. Brown *et al.* have considered different kernel functions in their SVM study [1] and showed that SVMs using the radial basis or a higher dimension dot product kernel (such as *D-p 3 SVM*) outperformed their contemporaries. Based on their results, we use the “*D-p 3 SVM*” kernel method as described by Brown *et al.* in our evaluation study reported in this paper.

For MLP, we use a multilayer perceptron architecture based on the one described in Mateos *et al.* [8]. Our MLP has one input layer consisting of 80 units, one hidden layer with eight units, and one output layer with five units—one for each of the functional classes in Table 2.

The programs that we use in our study are as follows. For SVMs, we use the GIST package available at <http://microarray.cpmc.columbia.edu/gist> (version 2.0.5). For MLPs, we use the neural network implementation in the WEKA package [16] available at <http://www.cs.waikato.ac.nz/~ml/weka> (version 3.3.4).

### 3.4 Evaluation Metrics

We evaluate the performance of the classifiers using three-way cross-validated experiments. The gene expression data are randomly divided into three groups: two-thirds of the data are used for training

the classifiers, the remaining third is used for testing. The procedure is then repeated for three times using different data subsets.

During testing, each classifier must produce a positive or negative class label for each test gene based only on what it has learned from the training set. The outputs are categorized as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). As a measure of the overall performance for each classification method  $M$ , we use  $S(M)$ —the “learning cost savings” as defined by Brown *et al.* in [1]. For each machine learning method  $M$ , the learning cost  $C(M)$  is defined as  $C(M) = fp(M) + 2.fn(M)$ , where  $fp(M)$  and  $fn(M)$  are the number of false positives and false negatives for method  $M$ . The learning cost savings for a method  $M$  is then defined as  $S(M) = C(null) - C(M)$ , which compare the learning cost of  $M$  with that of the “null” learning procedure which classifies all test examples as negative.

## 4 Study

To start, we show that our intuition that combining data from multiple microarray studies can improve gene functional classification performance is sensible. First, we apply both SVM and MLP gene classification algorithms using only individual datasets. Then, we apply the classification algorithms using all available datasets. We check if there is an improvement by comparing their learning cost savings. The  $S(M)$ ’s in Table 3 showed that using all the six datasets for training can sometimes beat using only individual datasets. In the case of SVM, with the exception of TCA, the “all-dataset” approach is always advantageous over the “single-dataset” approach. This shows that combining data from multiple microarray studies can improve the classification performance, even if they may have been conducted under varying experimental objectives and conditions.

Table 3: Gene classification using individual datasets versus using all available datasets from the different microarray studies on the five function classes. The dataset *Ccc* is omitted here because of its small experimental size.

Function	Learning cost savings $S(M)$											
	SVM						MLP					
	<i>Alp</i>	<i>Cdc</i>	<i>Elu</i>	<i>Spo</i>	<i>Dia</i>	all	<i>Alp</i>	<i>Cdc</i>	<i>Elu</i>	<i>Spo</i>	<i>Dia</i>	all
TCA	-360	-5	0	-157	-532	-9	-6	-6	-11	-1	-1	-2
Resp	-232	-160	-258	-348	-1319	-103	0	-4	0	0	-1	-12
Ribo	-250	69	-66	-6	-612	217	117	117	138	182	174	209
Prot	-438	-66	-367	-27	-116	25	-6	-2	-3	24	3	30
Hist	2	16	-2	11	-87	17	14	16	18	10	0	17

However, observe that in the case of MLP, the “all-dataset” approach does not always beat the “single-dataset” approach. Table 4 shows some further examples for both SVM and MLP that certain selective combinations of all the available datasets can lead to better classification results. The results indicate that blindly including all available microarray data in data analysis is not the best approach for combining multiple microarray datasets for improving data mining results.

Table 4: Examples showing that blindly combining the datasets from all the microarray studies may not always lead to the best classification performance.

Function	SVM		MLP	
	Datasets	$S(M)$	Datasets	$S(M)$
Resp	all	-103	all	-12
	$Cdc + Elu + Spo$	-10	$Cdc$	0
Prot	all	25	all	30
	$Alp + Ccc + Spo$	39	$Spo + Dia$	40

## 5 Method

The results in our initial evaluation study reported above suggests that for improved data mining results, additional microarray datasets should always be included in a selective manner. Here, we consider this dataset selection problem as one of feature selection by treating each dataset as one single feature. As such, traditional feature selection methods cannot be immediately applied. They typically consider each experiment—instead of the entire set of experiments from a study—as a feature.

To evaluate our whole-dataset feature selection approach, we devise a simple hill-climbing (greedy) method for choosing which datasets to learn from during the training phase. As a hill-climbing approach, the effect of adding a candidate dataset is tested and added one at a time until no further improvement in learning occurs.

Let  $D_{start}$  be a starting microarray dataset that is to be analyzed with a classification algorithm  $M$ . Typically,  $D_{start}$  would be a new microarray dataset generated by one’s own laboratory. We want to maximize the performance of  $M$  on this dataset by combining it with additional datasets from other studies in the data analysis. In the case where there are no specific start sets, we set  $D_{start} = \emptyset$ .

Let  $\Phi_{additional} = \{D_1, \dots, D_n\}$  be  $n$  additional microarray datasets conducted on the same set of genes as  $D_{start}$ . These additional datasets can be from different laboratories and experimental studies. Our objective is to search for a subset from  $\Phi_{additional}$  that can be combined with  $D_{start}$  to give the best data analysis results by  $M$ :

**Step 1:** Normalize the expression vectors in  $D_{start}, D_1, \dots, D_n$  to be of real values between 0 and 1.

**Step 2:** Let  $\mathcal{S}_{M,\Psi}$  be the  $S(M)$  score of applying  $M$  on the datasets in  $\Psi \subseteq \Phi_{all}$ , where  $\Phi_{all} = \Phi_{additional} \cup \{D_{start}\}$ . Set  $\mathcal{D}_{best}^{(0)} := D_{start}$ .

**Step 3:** In the  $i$ -th iteration, let  $\mathcal{D}_{best}^{(i)} \in \Phi_{all} - \{\mathcal{D}_{best}^{(k)} \mid 0 \leq k \leq i-1\}$  such that

$$\mathcal{S}_{best}^{(i)} := \max_{D_j \in \Phi_{all} - \{\mathcal{D}_{best}^{(k)} \mid 0 \leq k \leq i-1\}} \mathcal{S}_{M, \{\mathcal{D}_{best}^{(k)} \mid 0 \leq k \leq i-1\} \cup \{D_j\}}$$

**Step 4:** Halt the iteration process in Step 3 if  $i > n$  or  $\mathcal{S}_{best}^{(i)} \leq \mathcal{S}_{best}^{(i-1)}$ .

Upon termination,  $\mathcal{D}_{best}^{(1)}, \dots, \mathcal{D}_{best}^{(i-1)}$  will be a selection of additional microarray datasets that can be combined with  $D_{start}$  to produce better classification performance than that from just using  $D_{start}$  alone. In the next section, we report results from our comparative evaluation study showing cases in which  $\mathcal{S}_{best}^{(i-1)} \geq \mathcal{S}_{M, \Phi_{all}}$ , confirming that our whole-dataset feature selection approach can perform better than the “all-dataset” approach.

## 6 Results

We have applied the simple hill-climbing whole-dataset feature selection described in the previous section for the selective combination of multiple microarray datasets in yeast gene functional classification. We compare our results with the previous studies by Brown *et al.* [1] using SVM and Mateos *et al.* [8] using MLP. Both groups have used the “all-dataset” approach for the functional classification of genes on the same yeast gene expression datasets. We also compare our whole-dataset feature selection approach with traditional feature selection methods that treat each of the 80 experiments in the datasets as individual features, using such feature evaluation metrics as *Fisher criterion scores* and *standard t-tests* to evaluate individual features or experiments.

Table 5: Comparison of error rates for SVM with various feature selection methods. The SVM method used here is the *D-p 3 SVM* method as described in Brown *et al.* [1]. *Fisher* and *t – test* are two feature selection methods for selecting each of the 80 experiments as individual features, using the Fisher criterion score and standard t-test as the feature evaluation metric respectively. DATASET<sub>all</sub> denotes the naive approach of blindly combining all available datasets for analysis. DATASET<sub>hill</sub> denotes the method of whole-dataset feature selection by hill-climbing as described in Section 5.

Class	Method <i>M</i>	Data	FP	FN	TP	TN	<i>S(M)</i>
TCA	SVM + DATASET <sub>all</sub>	all	39	7	15	2489	-9
	SVM + <i>Fisher</i>	8/80 exp.	36	22	0	2492	-36
	SVM + <i>t-test</i>	8/80 exp.	37	22	0	2491	-37
	SVM + DATASET <sub>hill</sub>	<i>Alp + Elu + Spo</i>	10	13	9	2518	8
Resp	SVM + DATASET <sub>all</sub>	all	123	14	10	2403	-103
	SVM + <i>Fisher</i>	8/80 exp.	698	23	1	1828	-696
	SVM + <i>t-test</i>	8/80 exp.	926	22	2	1600	-922
	SVM + DATASET <sub>hill</sub>	<i>Cdc + Elu + Spo</i>	14	22	2	2512	-10
Ribo	SVM + DATASET <sub>all</sub>	all	29	6	123	2392	217
	SVM + <i>Fisher</i>	8/80 exp.	105	129	0	2316	-105
	SVM + <i>t-test</i>	8/80 exp.	98	129	0	2323	-98
	SVM + DATASET <sub>hill</sub>	<i>Cdc + Spo + Dia</i>	15	8	121	2406	227
Prot	SVM + DATASET <sub>all</sub>	all	33	4	29	2484	25
	SVM + <i>Fisher</i>	8/80 exp.	5	33	0	2512	-5
	SVM + <i>t-test</i>	8/80 exp.	2	31	2	2515	2
	SVM + DATASET <sub>hill</sub>	<i>Cdc + Ccc + Spo</i>	7	10	23	2510	39
Hist	SVM + DATASET <sub>all</sub>	all	1	2	9	2538	17
	SVM + <i>Fisher</i>	8/80 exp.	0	11	0	2539	0
	SVM + <i>t-test</i>	8/80 exp.	0	11	0	2539	0
	SVM + DATASET <sub>hill</sub>	<i>Cdc + Spo</i>	0	2	9	2539	18

Table 5 shows the detailed results of using various dataset and feature selection methods to improve classification by SVM. Our whole-dataset feature selection beats the naive “all-dataset” approach used by the previous groups. Traditional feature selection methods that selects each of the 80 experiments as individual features using standard metrics such as *Fisher* and *t – test* did not improve but actually worsened the classification performance. Unlike our whole-dataset feature selection approach, these conventional feature selection methods did not consider the dependency in the experiments within a study—for example, five out of the six microarray studies were time-series experiments. The results of

our hill-climbing whole-dataset feature selection show that by considering whole datasets as individual “features”, it becomes advantageous to perform feature selection.

The overall results for both SVM and MLP are shown in Table 6. Our results show that our simple hill-climbing approach for whole-dataset feature selection is better in improving classification performance than the use of the best individual dataset or all available datasets for the learning procedure. In a more comprehensive study, we apply our whole-dataset feature selection method with SVM to classify the 1,851 MIPS-annotated genes unambiguously assigned to 60 non-singleton functional classes. The results—as shown in Figure 1—ascertained that our hill-climbing whole-dataset feature selection approach almost always outperforms the blind “all-dataset” approach, as well as conventional feature selection methods such as Fisher and t-test.

Table 6: Comparison of gene classification performance with SVM and MLP using different dataset selection schemes.

Class	Selection method	SVM		MLP	
		$S(M)$	Dataset(s)	$S(M)$	Dataset(s)
TCA	INDIVIDUAL <sub>best</sub>	-1	<i>Elu</i>	0	<i>Ccc</i>
	DATASET <sub>all</sub>	-9	all	-2	all
	DATASET <sub>hill</sub>	8	<i>Alp + Elu + Spo</i>	0	<i>Ccc</i>
Resp	INDIVIDUAL <sub>best</sub>	-171	<i>Cdc</i>	0	<i>Alp, Elu, or Spo</i>
	DATASET <sub>all</sub>	-103	all	-12	all
	DATASET <sub>hill</sub>	-10	<i>Cdc + Elu + Spo</i>	0	<i>Alp</i>
Ribo	INDIVIDUAL <sub>best</sub>	103	<i>Cdc</i>	182	<i>Spo</i>
	DATASET <sub>all</sub>	217	all	209	all
	DATASET <sub>hill</sub>	227	<i>Cdc + Spo + Dia</i>	200	<i>Elu + Spo</i>
Prot	INDIVIDUAL <sub>best</sub>	-28	<i>Spo</i>	24	<i>Spo</i>
	DATASET <sub>all</sub>	25	all	30	all
	DATASET <sub>hill</sub>	39	<i>Cdc + Ccc + Spo</i>	40	<i>Spo + Dia</i>
Hist	INDIVIDUAL <sub>best</sub>	16	<i>Cdc</i>	18	<i>Elu</i>
	DATASET <sub>all</sub>	17	all	17	all
	DATASET <sub>hill</sub>	18	<i>Cdc + Spo</i>	18	<i>Elu</i>

## 7 Conclusion

Microarray technology has certainly revolutionized the experimental study of functional relationships among genes. Successful functional analysis of the experimental gene expression data can lead to information useful for the elucidation of molecular mechanisms underlying various diseases [10, 12]. It is therefore important to improve on the functional analysis of experimental gene expression data. Our work addresses this need, enhancing data analysis performance in the functional classification of genes by combining multiple microarray studies, using data resources that are increasingly common to scientists.

We have shown in this paper that multiple microarray studies can be combined together for improved microarray data analysis. Even if the various experimental studies may have been conducted under different conditions or for different objectives, including additional microarray datasets from other studies during data mining generally leads to improved performance. However, we have also

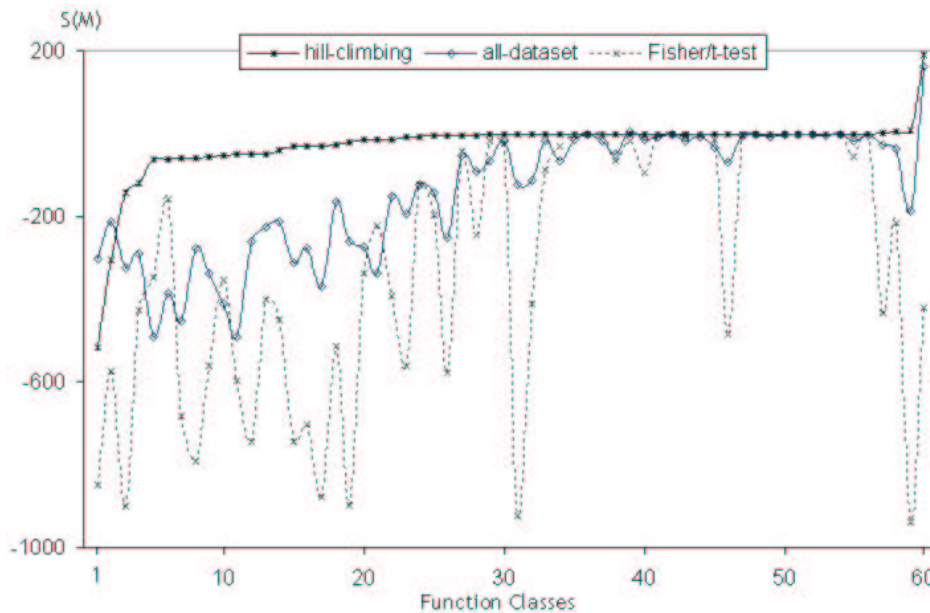


Figure 1: Classification performance of 60 unambiguous non-singleton MIPS functional classes using SVM with three different feature selection schemes:  $\text{DATASET}_{hill}$  (“hill-climbing”),  $\text{DATASET}_{all}$  (“all-dataset”), and the better of Fisher or t-test (“Fisher/t-test”).

shown that naively combining all available datasets is not the best approach. As such, we have devised a simple hill-climbing selection process for deciding which of the available datasets to be included in the combined data analysis for improved performance. We have shown that our simple hill-climbing feature selection method generally performed better than blindly combining all datasets for analysis. For further work, we will investigate the use of more sophisticated whole-dataset feature selection algorithms that can lead to optimal data analysis performance.

The learning task of functional classification of genes from whole-genome microarray data is not an easy one. One major problem is the imbalance in the number of positive and negative training examples: each functional class often contains very few members relative to the total number of genes in the datasets. Furthermore, many of the negative examples are weakly labeled as such, or mislabeled, as illustrated by Mateos *et al.* [8] and others. Such classification noise existing in the large proportion of the negative training examples can easily outweigh the small number of positive examples, making it difficult for machine learning. As a result, some researchers have found that only  $\sim 10\%$  of the gene functional classes are learnable [8]. To combat this problem, researchers have attempted refining the machine learning algorithms. For example, Brown *et al.* modified the kernel values for their support vector machines [1]. In our work, we showed that the strategy of selectively combining additional datasets from multiple microarray studies can also improve the learning rate. Our whole-dataset feature selection approach can be an alternative to combat this machine learning problem, in addition to improving the machine learning algorithms themselves.

## References

- [1] Brown, M.P., Grundy, W.N., Lin, D., *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA*, 97(1):262–267, 2000.
- [2] Cheng, Y. and Church, G.M., Biclustering of expression data, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:93–103, 2000.

- [3] Chu, S., DeRisi, J., Eisen, M., *et al.*, The transcriptional program of sporulation in budding yeast, *Science*, 282(5389):699–705, 1998.
- [4] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278(5338):680–686, 1997.
- [5] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.
- [6] Hvidsten, T.R., Komorowski, J., Sandvik, A.K., and Laegreid, A., Predicting gene function from gene expressions and ontologies, *Pac. Symp. Biocomput.*, 299–310, 2001.
- [7] Li, J., Liu, H., Downing, J.R., *et al.*, Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients, *Bioinformatics*, 19(1):71–78, 2003.
- [8] Mateos, A., Dopazo, J., Jansen, R., *et al.*, Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons, *Genome. Res.*, 12(11):1703–1715, 2002.
- [9] Mewes, H.W., Frishman, D., Guldener, U., *et al.*, Mips: a database for genomes and protein sequences, *Nucleic Acids Res.*, 30(1):31–34, 2002.
- [10] Miller, L.D., Long, P.M., Wong, L., *et al.*, Optimal gene expression analysis by microarrays, *Cancer Cell*, 2(5):353–361, 2002.
- [11] Ogawa, N., DeRisi, J., and Brown, P.O., New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis, *Mol. Biol. Cell*, 11(12):4309–4321, 2000.
- [12] Schulze, A. and Downward, J., Navigating gene expression using microarrays—a technology review, *Nat. Cell Biol.*, 3(8):E190–E195, 2001.
- [13] Shannon, W., Culverhouse, R., and Duncan, J., Analyzing microarray data using cluster analysis, *Pharmacogenomics*, 4(1):41–52, 2003.
- [14] Spellman, P.T., Sherlock, G., Zhang, M.Q., *et al.*, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, 9(12):3273–3297, 1998.
- [15] Theilhaber, J., Connolly, T., Roman-Roman, S., *et al.*, Finding genes in the c2c12 osteogenic pathway by k-nearest-neighbor classification of expression data, *Genome Res.*, 12(1):165–176, 2002.
- [16] Witten, I. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.
- [17] Yoshimoto, H., Saltsman, K., Gasch, A.P., *et al.*, Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in *Saccharomyces cerevisiae*, *J. Biol. Chem.*, 277(34):31079–31088, 2002.