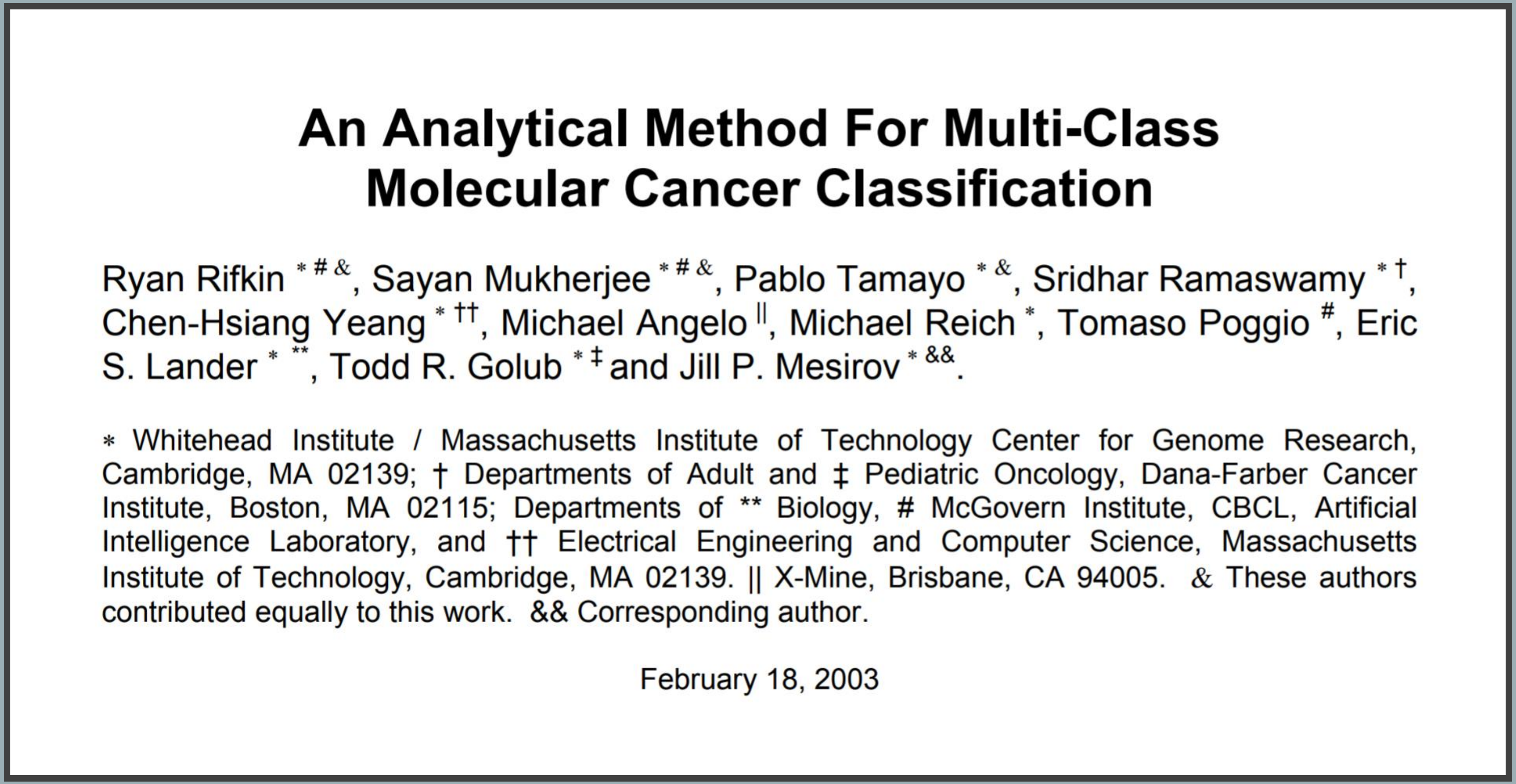


An Analytical Method For Multi-Class Molecular Cancer Classification

Ryan Rifkin [*#&], Sayan Mukherjee [*#&], Pablo Tamayo [*&], Sridhar Ramaswamy [*†], Chen-Hsiang Yeang [*††], Michael Angelo [||], Michael Reich [*], Tomaso Poggio [#], Eric S. Lander [* **], Todd R. Golub [*‡] and Jill P. Mesirov [*&&].

∗ Whitehead Institute / Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139; † Departments of Adult and ‡ Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115; Departments of ** Biology, # McGovern Institute, CBCL, Artificial Intelligence Laboratory, and †† Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139. || X-Mine, Brisbane, CA 94005. & These authors contributed equally to this work. && Corresponding author.

February 18, 2003

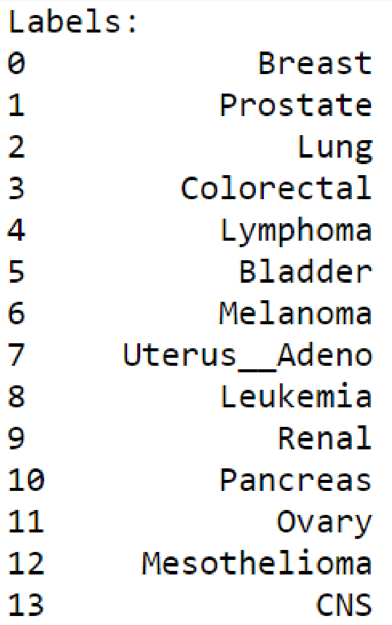
Machine Learning in Bioinformatics (NAIL107) Project

May 15, 2019

Richard Eliáš, Kateřina Břicháčková

Introduction

- Cancer treatment by molecular methods
- Multi-class classification of 14 tumor types
- DNA microarray gene expression data
- Combine multiple binary Support Vector Machine classifiers

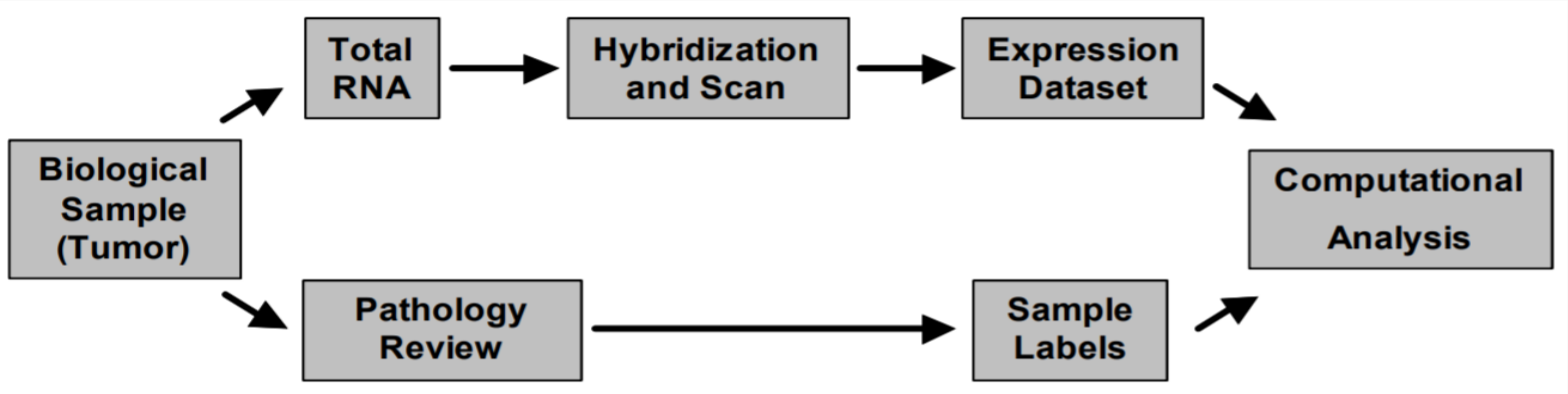


```
Labels:
0            Breast
1          Prostate
2              Lung
3        Colorectal
4          Lymphoma
5           Bladder
6          Melanoma
7     Uterus__Adeno
8          Leukemia
9             Renal
10         Pancreas
11            Ovary
12     Mesothelioma
13              CNS
```

Challenges

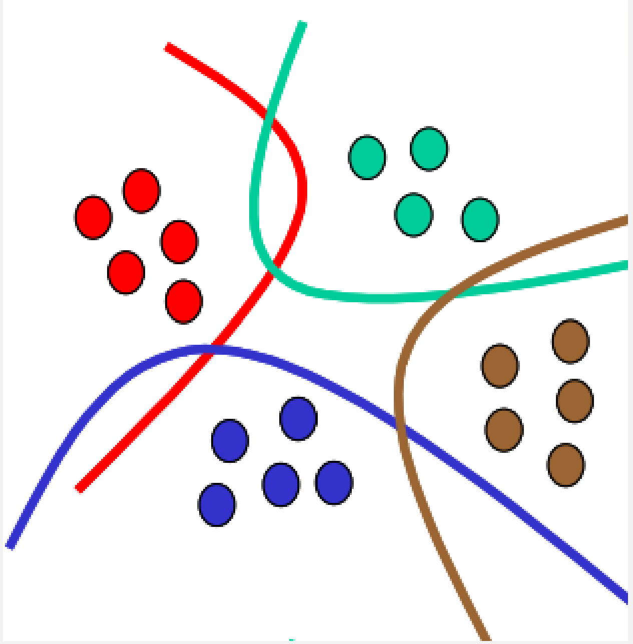
- Multi-class classification – more complex
  - Greater number of separation boundaries
  - Error rates can be higher
  - Random prediction:
    - Binary problem – accuracy 50%
    - K classes – accuracy 1/K

- Large dimensionality – 16063 genes
- Small number of samples – 198
- Solid tumors – harder sample preparation
- Uncertainty in the original labeling
- Measurement noise



Multi-class supervised classification

- Decomposition → set of binary problems
- Combine the binary classifiers
- 2 common approaches:
  - One-versus-all (OVA)
  - All-pairs (AP)
- AP approach should be theoretically more accurate
  - But very problem dependent…

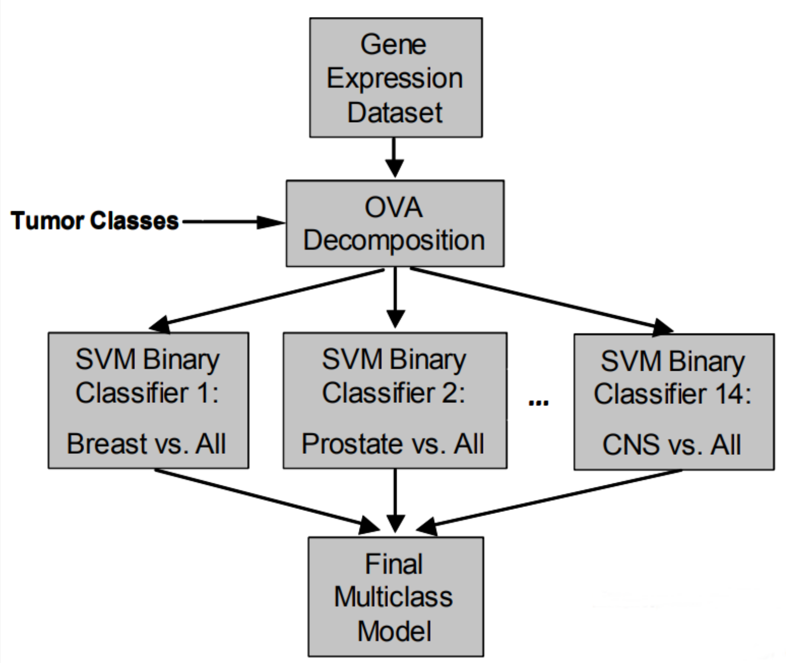


Methodology

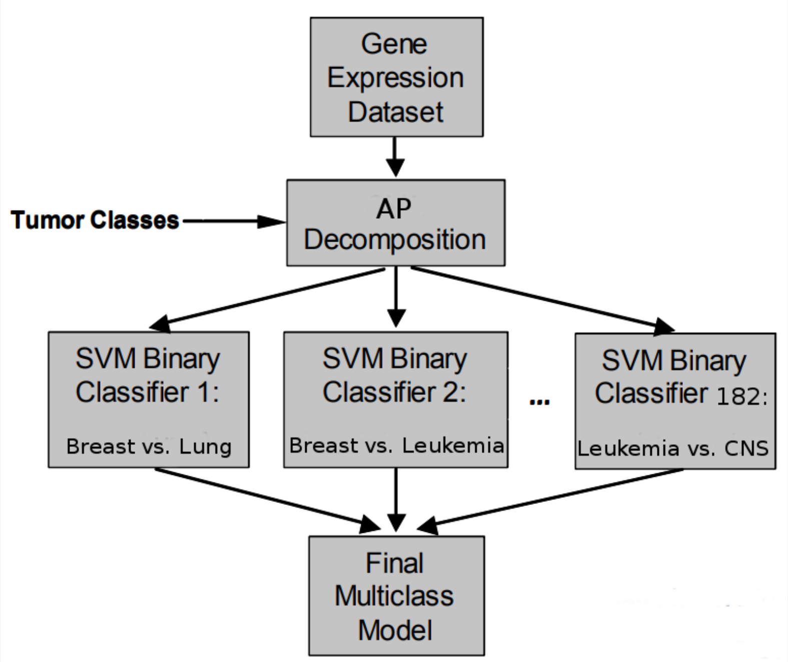
- **OVA** approach → 14 binary SVM classifiers
- Winning class:

$$\mathit{class} = \arg\max_{i=1..K} f_i$$

- Linear kernel
- No feature selection



Gene Expression Dataset
Tumor Classes
OVA Decomposition
SVM Binary Classifier 1: Breast vs. All
SVM Binary Classifier 2: Prostate vs. All
...
SVM Binary Classifier 14: CNS vs. All
Final Multiclass Model



Gene Expression Dataset
Tumor Classes
AP Decomposition
SVM Binary Classifier 1: Breast vs. Lung
SVM Binary Classifier 2: Breast vs. Leukemia
...
SVM Binary Classifier 182: Leukemia vs. CNS
Final Multiclass Model

- **AP** approach → 182 binary SVM classifiers
- Winning class:

$$\mathit{class} = \arg\max_{i=1..K} \left[\sum_{j=1}^{K} f_{ij} \right]$$

- Linear kernel
- No feature selection

Results

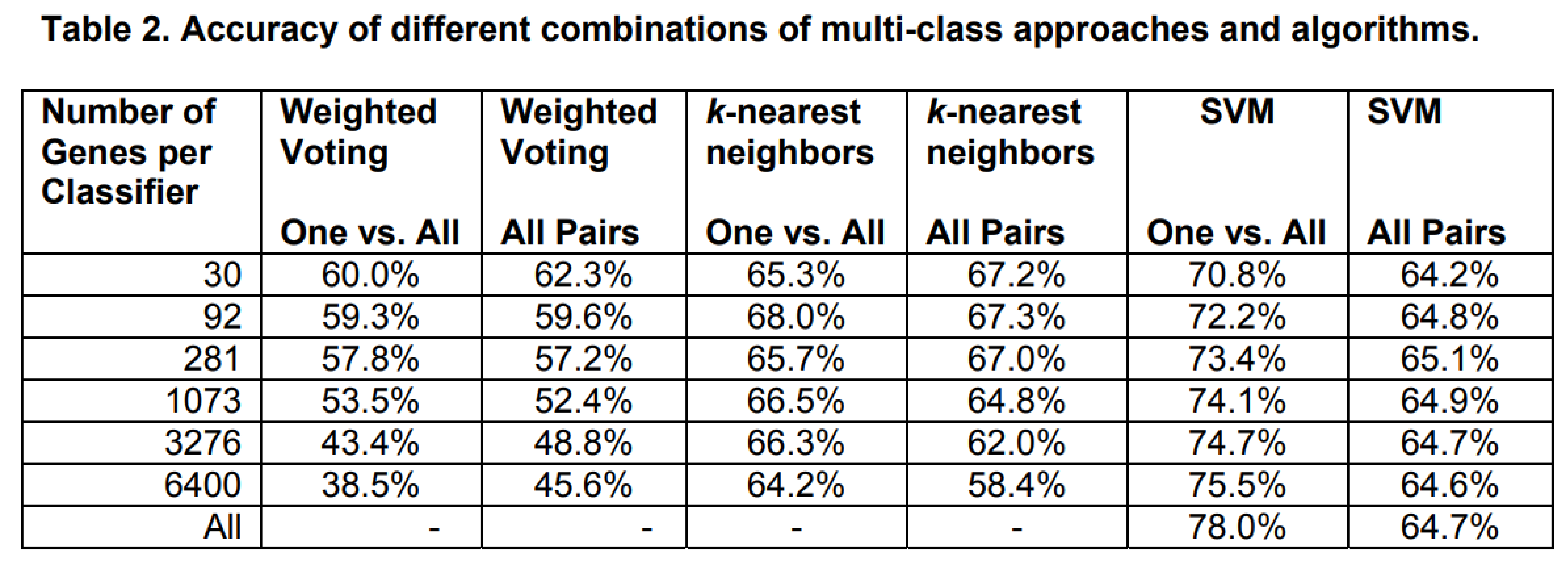


Table 2. Accuracy of different combinations of multi-class approaches and algorithms.

Number of Genes per Classifier	Weighted Voting One vs. All	Weighted Voting All Pairs	*k*-nearest neighbors One vs. All	*k*-nearest neighbors All Pairs	SVM One vs. All	SVM All Pairs
30	60.0%	62.3%	65.3%	67.2%	70.8%	64.2%
92	59.3%	59.6%	68.0%	67.3%	72.2%	64.8%
281	57.8%	57.2%	65.7%	67.0%	73.4%	65.1%
1073	53.5%	52.4%	66.5%	64.8%	74.1%	64.9%
3276	43.4%	48.8%	66.3%	62.0%	74.7%	64.7%
6400	38.5%	45.6%	64.2%	58.4%	75.5%	64.6%
All	-	-	-	-	78.0%	64.7%