massachusetts institute of technology — artificial intelligence laboratory

Multiclass Classification of SRBCTs

Gene Yeo and Tomaso Poggio

AI Memo 2001-018 August 25, 2001
CBCL Memo 206

Abstract

A novel approach to multiclass tumor classification using Artificial Neural Networks (ANNs) was introduced in a recent paper [1]. The method successfully classified and diagnosed small, round blue cell tumors (SRBCTs) of childhood into four distinct categories, neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), using cDNA gene expression profiles of samples that included both tumor biopsy material and cell lines. We report that using an approach similar to the one reported by Yeang et al [2], i.e. multiclass classification by combining outputs of binary classifiers, we achieved equal accuracy with much fewer features. We report the performances of 3 binary classifiers (k-nearest neighbors (kNN), weighted-voting (WV), and support vector machines (SVM)) with 3 feature selection techniques (Golub's Signal to Noise (SN) ratios [3], Fisher scores (FSc) and Mukherjee's SVM feature selection (SVMFS))[4].

1 Introduction

There currently exists no single biological or chemical test that can precisely distinguish small, round blue cell tumors of childhood (SRBCTs) into their subclasses, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS) [1]. A recent paper by Khan et al. reports that using gene expression profiles obtained from cDNA microarrays of samples which included both tumor tissue as well as cell lines, artificial neural networks (ANNs) can accurately distinguish the tumor sub-types using 96 top genes obtained from Principal Component Analysis from the more than 6000 genes of which the expression was measured. In order to identify candidate targets for therapy, it is of course important to identify a small subset of genes and yet retain high classification accuracy. Using an approach similar to that of Yeang et al [2], we have performed multiclass classification using 3 binary classifiers (k-nearest neighbors (kNN), weighted-voting (WV), and linear support vector machines (SVM)) in a one-versus-rest fashion with 3 feature selection techniques (Golub's Signal to Noise (SN) ratios [3], Fisher scores (FSc) and Mukherjee's SVM feature selection (SVMFS))[4]. With all of these techniques we have obtained accuracy equal to Khan et al. with much fewer genes (features).

2 Dataset

Khan et al. filtered the 6567 cDNA gene expression profiles by requiring a minimal intensity of expression, which reduced the number of genes to 2308. Khan et al. then generated 3750 (linear!) ANN models by minimizing the summed squared error using three-fold cross-validation. The genes were ranked by their significance for the classification using the 3750 previously calibrated models. Using the top 96 genes, Khan et al. achieved 0 error on their training data set of 63 samples. A set of 25 test samples, which included 6 EWS, 3 BL,6 NB,5 RMS and 5 non-SRBCTs (consisting of 2 normal muscle tissues and 3 cell lines including an undifferentiated sarcoma, osteosarcoma and a prostate carcinoma) were correctly classified (the 5 non-SRBCTs were excluded by a 95% percentile distance) by the ANNs. We downloaded the data set (http://www.nhgri.nih.gov/DIR/Microarray/Supplement/) consisting of 64 training samples (23 EWS; 8 Burkitt lympohmas (BL), a subset of NHL; 12 NB and 21 RMS samples) that belonged to a total of 4 classes and 25 test samples (6 EWS, 3 BL,6 NB,5 RMS and 5 non-SRBCTs (consisting of 2 normal muscle tissues and 3 cell lines including an undifferentiated sarcoma, osteosarcoma and a prostate carcinoma). Each sample was a feature vector of 2308 natural log-normalized gene expression values.

3 One-versus-rest Multiclass

In solving multiclass problems using binary classifiers combined in a one-versus-rest fashion, each classifier trained on one class versus the rest of the classes makes a prediction about a given test sample. A sample's predicted label is the class for which that classifier returned a positive label, and for which the rest of the classifiers returned negative labels. Sample predictions can be rejected in these two cases: (1) If each of the 4 binary classifiers does not predict that the sample is in its respective classes. (2) If conflicting predictions arise i.e. 2 or more binary classifiers predict that the sample belongs in their respective classes. Results with rejection are indicated in the tables below by “w.r.”.

In the tables below, (signed) refers to using the signed SVM labels to determine the class prediction (hard errors [2]), and (max) refers to assigning the class label for which the distance from the margin in the positive direction (i.e. the direction in which that class resides rather than the “rest”) is maximal. This approach has been used to solve other multiclass pattern recognition problems [7]. We can also determine the confidence of the prediction using the magnitudes of the SVM outputs. Figure 1 shows examples of rejections and low confidence samples.

4 Feature Selection

Features were selected by three methods, namely, Golub's Signal to Noise (S2N) ratios [3], Fisher scores (FSc), which is similar to S2N ratios and Murkerjee et al's SVM feature selection technique (SVMFS))[4]. Binary classifiers (WV, SVM and kNN) were combined in a one-versus-rest fashion to perform multiclass classification. We performed leave-one-out cross validation on the training data set, using top genes ranked by the various feature selection methods. In the tables below, the top X genes means X genes discriminated one class from the rest of the classes, and because there are four classes, the total number of genes used are 4 times X, or fewer, as there are overlaps in some of the gene subsets. For example, 5 S2N genes (in the table below) means using top 5 genes from Table 8, top 5 genes from Table 9, top 5 genes from Table 10 and top 5 genes from Table 11. The results for the 20 SRBCT test samples are presented below. The bracketed numbers are the Leave-One-Out Cross Validation (LOOCV) error for the 64 training samples.

As a reminder, Golub et al's Signal to Noise Ratio (S2N) used to rank genes (features) corresponds to the following statistic:

$$P(j) = \left| \frac{\mu_1(j) - \mu_{-1}(j)}{\sigma_1(j) + \sigma_{-1}(j)} \right|, \tag{1}$$

where j is the gene index, μ_1 is the mean of class 1 for gene j, μ_{-1} is the mean of class -1 for gene j, σ_1 is the standard deviation of class 1 for gene j, and σ_{-1} is the standard deviation of class -1 for gene j.

A related coefficient has been used in Pavlidis et al. [6], which we term Fisher scores (FSc) as it's similar to Fisher's discriminant criterion [5]:

$$P(j) = \left| \frac{(\mu_1(j) - \mu_{-1}(j))^2}{\sigma_1(j)^2 + \sigma_{-1}(j)^2} \right|, \tag{2}$$

Mukherjee et al. formulated a more direct SVM feature selection algorithm. During the QP minimization the input space is rescaled such that the margin in feature space increases subject to the constraint that the volume feature space remains constant[4]. In particular, a diagonal matrix A of scaling factors $a_1, a_2 ... a_n$ is incorporated into the kernel:

$$K_A(\mathbf{x}, \mathbf{y}) = K(A\mathbf{x}, A\mathbf{y}) \tag{3}$$

and training is performed by the following iterative steps: 1) Optimize the α's for a fixed A and 2) Optimize A for fixed α's by gradient descent. The genes with the largest scaling factors are selected as important features.

5 Classification Results

Referring to Table 1, SVM (max) achieves perfect performance (on the 20 SRBCT test samples) and made 1 LOOCV error using all 2308 genes (Figure 1), compared to 2 test errors and 4 LOOCV errors made by kNN and 9 test errors and 19 LOOCV error made by WV. The sample misclassified during LOOCV by SVM (max) is EWS-T13 [1]. This sample occurs frequently as a LOOCV error, which agrees with Khan et al's analysis, who opined that EWS-T13 cannot be confidently diagnosed (as it fell outside the expected 95th percentile distance of a perfect vote). The sample that is misclassified by SVM (signed) is RMS-T9 which occurs the next most frequently. Khan et al's analysis reported training with 63 samples, and excluded this sample [1].

kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
2 (4)	2 (4)	9 (19)	4 (4)	0 (2)	0 (1)	0 (1)

Table 1: Test errors (20 samples) using all genes. LOOCV errors (64 samples) in brackets. (signed) refers to signed output as predicted class, (max) refers to maximal output as predicted class. w.r. stands for "with rejections".

FSc	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	0 (2)	0 (1)	3 (2)	0 (0)	0 (2)	0 (0)	**0 (0)**
60	2 (3)	0 (0)	4 (2)	0 (0)	1 (2)	0 (0)	**0 (0)**
20	2 (3)	1 (0)	2 (2)	1 (0)	2 (3)	1 (0)	**1 (0)**
10	3 (2)	1 (0)	1 (2)	1 (0)	4 (3)	1 (0)	**1 (0)**
5	4 (2)	0 (0)	6 (4)	0 (0)	4 (4)	0 (0)	**1 (0)**
S2N	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	0 (2)	0 (1)	2 (0)	0(0)	0 (2)	0 (0)	**0 (0)**
60	0 (2)	0 (0)	1 (0)	0(0)	1 (3)	0 (0)	**0 (0)**
20	2 (2)	0 (0)	1 (1)	0(0)	3 (2)	0 (0)	**1 (0)**
10	1 (2)	1 (0)	1 (2)	1(0)	2 (3)	1 (0)	**1 (0)**
5	2 (2)	1 (0)	2 (4)	0(0)	2 (2)	0 (0)	**0 (0)**
SVMFS	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	1 (1)	0 (1)	**0 (0)**	0 (0)	0 (1)	0 (0)	**0 (0)**
60	**0 (0)**	0 (0)	1 (0)	0 (0)	0 (1)	0 (0)	**0 (0)**
20	2 (2)	1 (0)	2 (1)	1 (0)	**1 (0)**	1 (0)	**1 (0)**
10	3 (3)	1 (1)	4 (2)	1 (0)	1 (2)	1 (0)	**1 (0)**
5	6 (5)	1 (0)	7 (6)	1 (0)	7 (4)	1 (0)	**1 (1)**

Table 2: Test errors (20 samples). LOOCV errors (64 samples) in brackets. (signed) refers to signed output as predicted class, (max) refers to maximal output as predicted class. w.r. stands for "with rejections". Left column is top number of genes ranked by FSc, S2N or SVMFS used to train one classifier (multiply by 4 to get max possible number of distinct genes).

Table 2 presents the errors for the 20 SRBCT test samples and the LOOCV errors for the 64 training samples in brackets, achieved by the different classification methods, for different feature selection methods, and number of features used. The numbers in **bold** are the best errors in each row for the errors achieved without rejections of samples (as described above). It is evident that SVM (max) achieves better results. Rejection of samples reduces the errors, in most cases, to 1 or even 0. Also, using SVMFS reduces errors for kNN, WV and SVM (signed) when using larger numbers of features (20, 60, 100 per class). However, SVMS performed poorly for smaller sets of genes (5,10). SVMFS distributes the importance of features to get better classification accuracy but by using more features.

Table 3 presents the predicted classes of the 5 non-SRBCT samples in the test set, retaining the sample names as used in the Khan et al's paper [1]. A value of "0" in the table means the non-SRBCT test sample was rejected (by the conditions above). Test 9 and 13 are 2 normal muscle tissues and the rest were from cell lines: an undifferentiated sarcoma (Test 5), osteosarcoma (Test 3) and prostate carcinoma (Test 11). Test 9, 11, 5, 13 and 3 were predicted by Khan et al's ANNs to be classes 4, 1, 3, 4 and 4 respectively (1,2,3,4 referring to EWS, BL, NB and RMS). The best performance with regard to the non-SRBCTs was achieved by SVM with the top 5 genes per class (20 in total) selected by SVMFS, which predicted that all 5 non-SRBCT samples did not belong in any of the 4 classes. Using top 5 genes per class (20 in total) selected by S2N or FSc resulted in 1 mistake out of the 5 non-SRBCT samples made by SVM, Test 13 (predicted to be class 4), which agrees with the ANN prediction [1]. The one test sample that is misclassified by almost all methods is Test 20, which agrees with Khan et al's [1] analysis. Figure 2 depicts the typical low confidences associated with both Test 13 and Test 20. In addition, it may be biologically significant that Test 9 and 13 (skeletal muscle tissues) are misclassified often as belonging to subclass EWS.

Genes	Sample	FSc-kNN	FSc-WV	FSc-SVM	S2N-kNN	S2N-WV	S2N-SVM	SVMFS-kNN	SVMFS-WV	SVMFS-SVM
100	Test 9	4	4	4	4	4	4	4	4	4
100	Test 11	0	0	0	0	0	0	0	0	0
100	Test 5	0	0	0	0	0	0	0	0	0
100	Test 13	4	4	4	4	4	4	4	4	4
100	Test 3	0	0	0	4	0	0	0	0	0
60	Test 9	0	4	4	0	4	0	4	4	4
60	Test 11	0	0	0	0	0	0	0	0	0
60	Test 5	0	0	0	0	0	0	1	0	0
60	Test 13	4	4	4	4	4	4	4	4	4
60	Test 3	1	0	0	1	0	0	0	0	0
20	Test 9	0	0	0	0	0	0	4	0	4
20	Test 11	0	0	0	0	0	0	0	0	0
20	Test 5	0	0	0	0	0	0	0	0	0
20	Test 13	0	0	0	0	0	0	4	4	4
20	Test 3	0	0	0	0	0	0	0	0	0
10	Test 9	0	0	0	0	0	0	0	4	4
10	Test 11	0	0	0	0	0	0	0	0	0
10	Test 5	0	0	0	0	0	0	0	0	0
10	Test 13	0	4	0	0	4	4	4	0	0
10	Test 3	0	0	0	0	0	0	0	0	0
5	Test 9	0	0	0	0	0	0	4	4	0
5	Test 11	0	0	0	0	0	0	0	0	0
5	Test 5	0	0	0	0	0	0	0	0	0
5	Test 13	4	4	4	4	4	4	0	0	0
5	Test 3	0	0	0	0	0	0	1	0	0

Table 3: Predicted labels for 5 non-SRBCT Test samples. (1,2,3,4 refers to predicted as EWS, BL, NB and RMS, 0 refers to rejection).

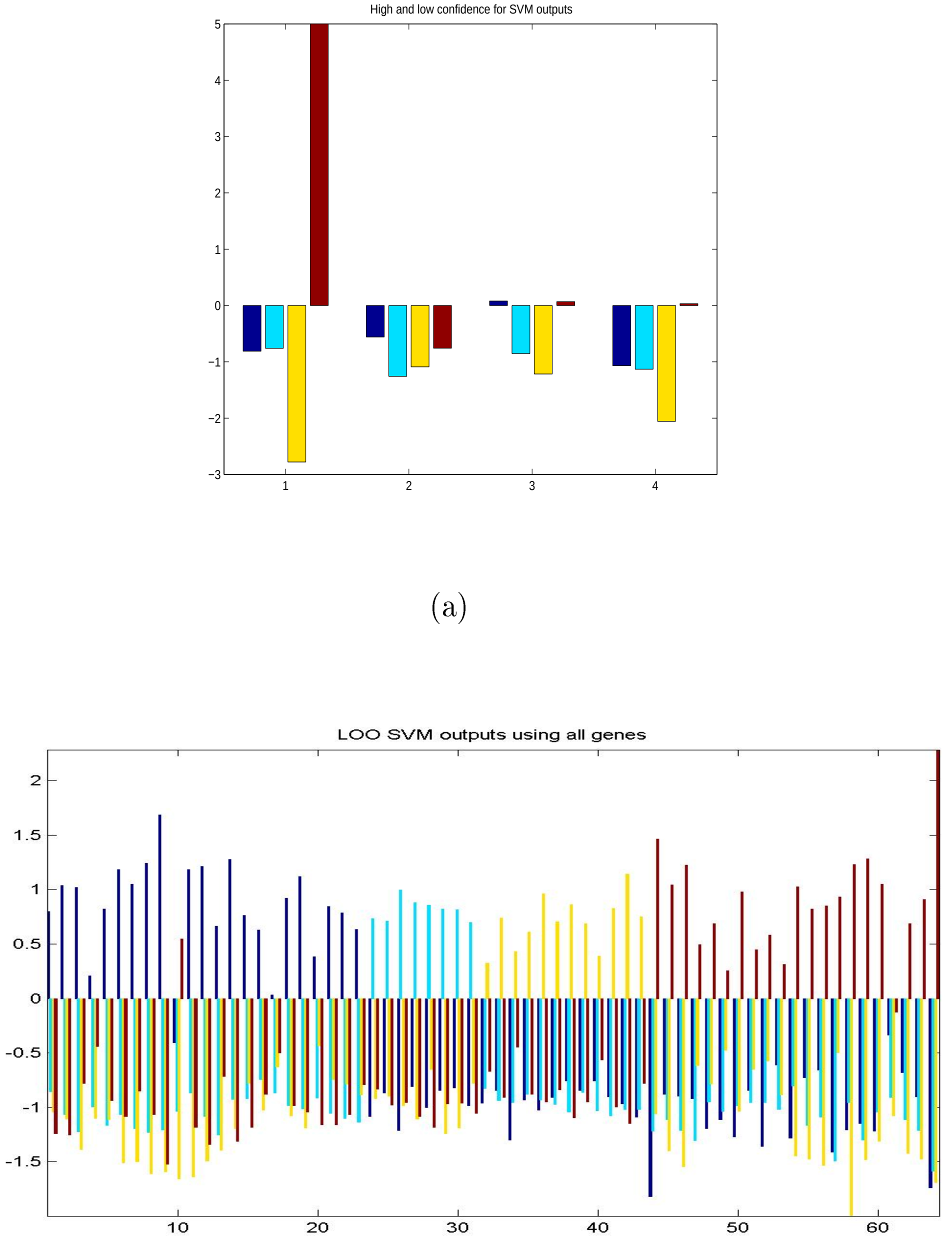
In summary, SVM(max) is able to achieve 0 test error and 0 LOOCV error, but made 1 mistake (with rejections) in the non-SRBCTs, using a total of 20 genes (5 genes per class) ranked by S2N (refer to Figure 3). Since the genes involved may be biologically relevant, then we suggest that these 20 genes should be investigated biologically. The list of genes ranked by the different feature selection schemes are in the appendix.

6 Comparing Ranking of Genes

Table 16 and Table 17 compare the top 10 genes (per class) ranked by S2N and SVMFS (FSc gives similar features to S2N) with the rank of the genes in Khan et al's analysis [1]. 22 of the 40 genes ranked by S2N and 20 of the 40 genes ranked by SVMFS overlap with Khan et al's top 96 genes.

7 Conclusion

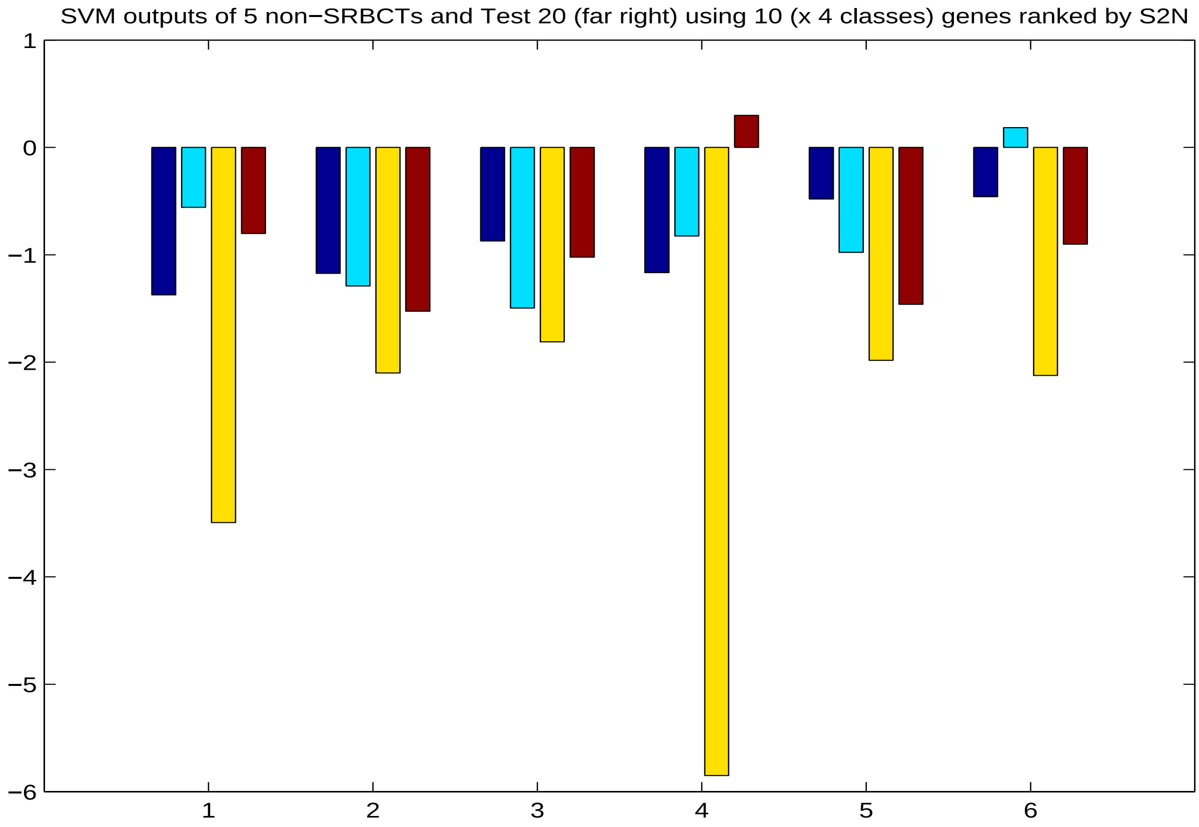
Small round, blue cell tumors can be easily classified into their classes by classical methods for classification, such as Weighted voting, kNN, and linear SVM classifiers (a special case extension of classical regularization approach). Khan et al. used a one-layer ANN, which is in effect a very simple linear discriminant method (not regularized and thus just minimizing the empirical error), motivating us to use other classical methods better suited for classification to assess accuracy and gene selection. The methods used in this paper are generally comparable in terms of performance, regardless of the feature selection algorithm (Signal to Noise, Fisher scores and SVM feature selection (SVMFS). SVMFS improves accuracy for different classification methods when relatively large numbers of features are used, but fails when very small numbers are used. It may be best to consider the genes selected by SVMFS to discriminate samples when large feature spaces are available, and the genes selected by, say, S2N when it is of pharmaceutical interest to pick few genes. Using the maximal SVM outputs gives the best accuracy (1 low confidence error for Test 13 (non-SRBCT)) with just 20 genes (in total) ranked by their Signal to Noise ratios. Rejection of samples in this multiclass scheme is useful for reducing errors, and may indicate biological peculiarities in the rejected samples. Lastly, the top genes (in appendix) are candidates for further histochemical and biological validation. Notice that it cannot be expected in general that successful classification of tumor types may be achieved using a small number of genes (P. Tamayo, personal communication). When, however, this is possible as in the case discussed here, it is likely to be biologically significant.



(a)

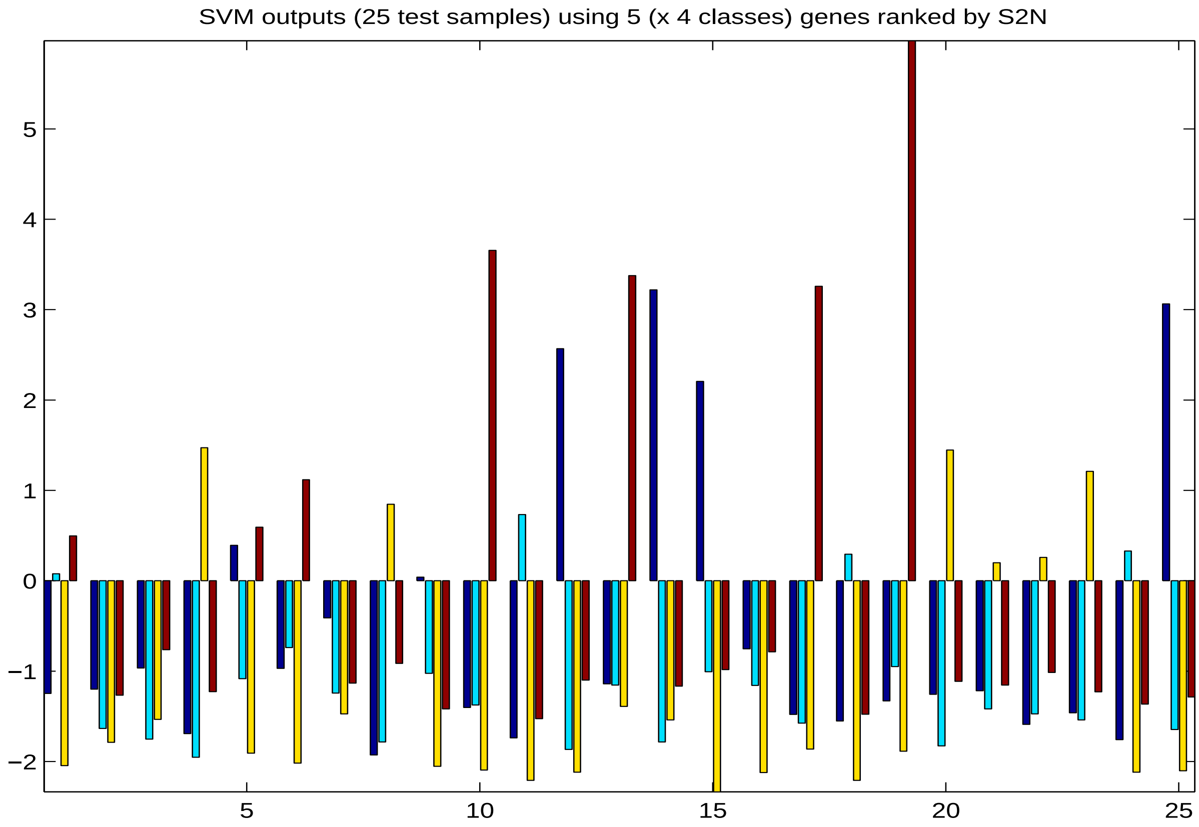
(b)

Figure 1: (a) The first sample (from the left) has high confidence, the second and third are low confidence and also rejectable, and the fourth is a low confidence (non-rejectable) sample. The four bars (per sample) are the outputs of the 4 linear SVMs. (b) SVM LOOCV outputs using all genes (64 samples).

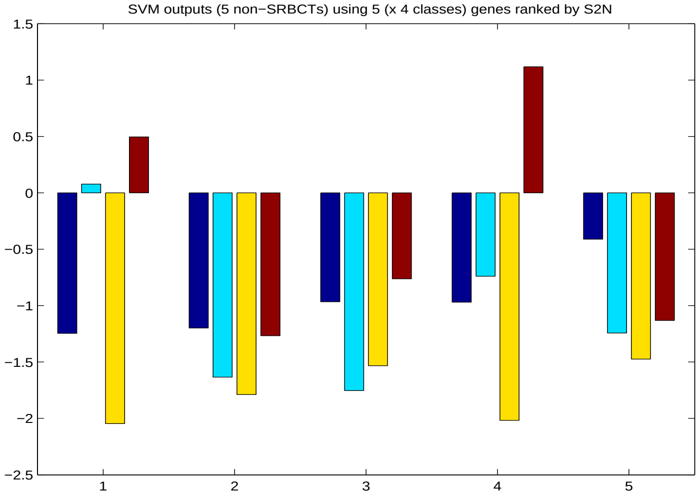


(a)

Figure 2: (a) SVM outputs of 5 non-SRBCTs (9,11,5,13,3 from left) and Test 20 (far right) using 10 (x 4 classes) genes ranked by S2N. Note the typical low confidences of Test 13 and Test 20.



(a)



(b)

Figure 3: (a) SVM outputs (S2N selected top 5 (x 4 classes) genes for all 25 test samples. (b) 5 non-SRBCTs (Test 9, 11, 5, 13, 3 from left to right) from (a). All rejectable except Test 13 (misclassifed as RMS).

A Ranked Genes by SVMFS

rank	Image Id.	Gene Description
1	139957	ESTs
2	754649	chromosome 14 open reading frame 3
3	302933	nucleolin
4	897690	tumor rejection antigen (gp96) 1
5	50887	ESTs
6	296448	insulin-like growth factor 2 (somatomedin A)
7	866702	”protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)”
8	244618	ESTs
9	379708	
10	811000	”lectin, galactoside-binding, soluble, 3 binding protein (galectin 6 binding protein)”
11	755578	”solute carrier family 7 (cationic amino acid transporter, y+ system), member 5”
12	194384	basic transcription factor 3
13	627939	cysteine and glycine-rich protein 3 (cardiac LIM protein)
14	1435862	”antigen identified by monoclonal antibodies 12E7, F21 and O13”
15	814260	follicular lymphoma variant translocation 1
16	842861	heterogeneous nuclear ribonucleoprotein R
17	34357	”actin, beta”
18	172751	”amyloid beta (A4) precursor protein-binding, family A, member 1 (X11)”
19	42076	TRK-fused gene (NOTE: non-standard symbol and name)
20	781097	”neurotrophic tyrosine kinase, receptor-related 1”

Table 4: Discriminated Class 1 (EWS) from the rest

rank	Image Id.	Gene Description
1	868304	”actin, alpha 2, smooth muscle, aorta”
2	840942	”major histocompatibility complex, class II, DP beta 1”
3	233721	insulin-like growth factor binding protein 2 (36kD)
4	745343	”regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)”
5	47475	”Homo sapiens inducible protein mRNA, complete cds”
6	80109	”major histocompatibility complex, class II, DQ alpha 1”
7	1461138	”H4 histone family, member G”
8	241412	E74-like factor 1 (ets domain transcription factor)
9	344134	immunoglobulin lambda-like polypeptide 3
10	1416782	”creatine kinase, brain”
11	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
12	841620	dihydropyrimidinase-like 2
13	755145	villin 2 (ezrin)
14	626502	”actin related protein 2/3 complex, subunit 1B (41 kD)”
15	208718	annexin A1
16	1493527	asparagine synthetase
17	45544	transgelin 2
18	183337	”major histocompatibility complex, class II, DM alpha”
19	814526	ESTs
20	486110	profilin 2

Table 5: Discriminated Class 2 (BL) from the rest

rank	Image Id.	Gene Description
1	629896	microtubule-associated protein 1B
2	812105	transmembrane protein
3	878652	postmeiotic segregation increased 2-like 12
4	810057	cold shock domain protein A
5	44563	growth associated protein 43
6	82225	secreted frizzled-related protein 1
7	784224	fibroblast growth factor receptor 4
8	135688	GATA-binding protein 2
9	325182	”cadherin 2, N-cadherin (neuronal)”
10	365826	growth arrest-specific 1
11	34355	”calmodulin 2 (phosphorylase kinase, delta)”
12	377461	”caveolin 1, caveolae protein, 22kD”
13	1474174	”matrix metalloproteinase 2 (gelatinase A, 72kD gelatinase, 72kD type IV collagenase)”
14	755239	methyltransferase-like 1
15	950574	”H3 histone, family 3B (H3.3B)”
16	1435862	”antigen identified by monoclonal antibodies 12E7, F21 and O13”
17	244637	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 45620
18	841620	dihydropyrimidinase-like 2
19	812965	v-myc avian myelocytomatosis viral oncogene homolog
20	45544	transgelin 2

Table 6: Discriminated Class 3 (NB) from the rest

rank	Image Id.	Gene Description
1	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
2	784224	fibroblast growth factor receptor 4
3	296448	insulin-like growth factor 2 (somatomedin A)
4	840708	”superoxide dismutase 2, mitochondrial”
5	882522	argininosuccinate synthetase
6	302933	nucleolin
7	244618	ESTs
8	878798	beta-2-microglobulin
9	810512	thrombospondin 1
10	52076	olfactomedinrelated ER localized protein
11	629896	microtubule-associated protein 1B
12	45544	transgelin 2
13	377461	”caveolin 1, caveolae protein, 22kD”
14	413633	
15	878652	postmeiotic segregation increased 2-like 12
16	33826	mitogen-activated protein kinase kinase 1
17	814260	follicular lymphoma variant translocation 1
18	788107	amphiphysin-like
19	839552	nuclear receptor coactivator 1
20	809603	”ESTs, Weakly similar to cDNA EST EMBL:M89154 comes from this gene [C.elegans]”

Table 7: Discriminated Class 4 (RMS) from the rest

B Ranked Genes by S2N

rank	Image Id.	Gene Description
1	770394	”Fc fragment of IgG, receptor, transporter, alpha”
2	377461	”caveolin 1, caveolae protein, 22kD”
3	814260	follicular lymphoma variant translocation 1
4	1435862	”antigen identified by monoclonal antibodies 12E7, F21 and O13”
5	295985	ESTs
6	866702	”protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)”
7	491565	”Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2”
8	1471841	”ATPase, Na+/K+ transporting, alpha 1 polypeptide”
9	52076	olfactomedinrelated ER localized protein
10	841641	cyclin D1 (PRAD1: parathyroid adenomatosis 1)
11	43733	glycogenin 2
12	214572	ESTs
13	713922	glutathione S-transferase M1
14	308497	KIAA0467 protein
15	1470048	”lymphocyte antigen 6 complex, locus E”
16	139957	ESTs
17	1473131	”transducin-like enhancer of split 2, homolog of Drosophila E(sp1)”
18	770868	NGFI-A binding protein 2 (ERG1 binding protein 2)
19	357031	”tumor necrosis factor, alpha-induced protein 6”
20	842820	inducible poly(A)-binding protein

Table 8: Discriminated Class 1 (EWS) from the rest

rank	Image Id.	Gene Description
1	236282	Wiskott-Aldrich syndrome (ecezema-thrombocytopenia)
2	745019	EH domain containing 1
3	183337	”major histocompatibility complex, class II, DM alpha”
4	504791	glutathione S-transferase A4
5	814526	ESTs
6	21652	”catenin (cadherin-associated protein), alpha 1 (102kD)”
7	624360	”proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)”
8	47475	”Homo sapiens inducible protein mRNA, complete cds”
9	813742	PTK7 protein tyrosine kinase 7
10	897788	”protein tyrosine phosphatase, receptor type, F”
11	897164	”catenin (cadherin-associated protein), alpha 1 (102kD)”
12	1416782	”creatine kinase, brain”
13	490772	small nuclear ribonucleoprotein polypeptide A’
14	1469292	pim-2 oncogene
15	344134	immunoglobulin lambda-like polypeptide 3
16	785793	”capping protein (actin filament) muscle Z-line, alpha 1”
17	241412	E74-like factor 1 (ets domain transcription factor)
18	68977	”proteasome (prosome, macropain) subunit, beta type, 10”
19	204545	ESTs
20	868304	”actin, alpha 2, smooth muscle, aorta”

Table 9: Discriminated Class 2 (BL) from the rest

rank	Image Id.	Gene Description
1	812105	transmembrane protein
2	786084	chromobox homolog 1 (Drosophila HP1 beta)
3	134748	glycine cleavage system protein H (aminomethyl carrier)
4	325182	"cadherin 2, N-cadherin (neuronal)"
5	486110	profilin 2
6	629896	microtubule-associated protein 1B
7	810057	cold shock domain protein A
8	81518	apelin; peptide ligand for APJ receptor
9	756401	Ras homolog enriched in brain 2
10	244637	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 45620
11	383188	recoverin
12	810864	"ESTs, Highly similar to CGI-48 protein [H.sapiens]"
13	544664	matrin 3
14	789376	thioredoxin reductase 1
15	308231	"Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c"
16	823886	"Smooth muscle myosin heavy chain isoform SMemb [human, umbilical cord, fetal aorta, mRNA Partial, 971 nt]"
17	823775	"guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3"
18	377048	"Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c"
19	878652	postmeiotic segregation increased 2-like 12
20	220096	

Table 10: Discriminated Class 3 (NB) from the rest

rank	Image Id.	Gene Description
1	784224	fibroblast growth factor receptor 4
2	796258	"sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)"
3	244618	ESTs
4	769716	neurofibromin 2 (bilateral acoustic neuroma)
5	789253	presenilin 2 (Alzheimer disease 4)
6	839552	nuclear receptor coactivator 1
7	143306	lymphocyte-specific protein 1
8	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
9	296448	insulin-like growth factor 2 (somatomedin A)
10	142134	ESTs
11	898219	mesoderm specific transcript (mouse) homolog
12	25725	farnesyl-diphosphate farnesyltransferase 1
13	813841	"plasminogen activator, tissue"
14	298062	"troponin T2, cardiac"
15	42558	glycine amidinotransferase (L-arginine:glycine amidinotransferase)
16	79022	FBJ murine osteosarcoma viral oncogene homolog B
17	246035	ESTs
18	859359	quinone oxidoreductase homolog
19	814444	"cofactor required for Sp1 transcriptional activation, subunit 9 (33kD)"
20	128054	ESTs

Table 11: Discriminated Class 4 (RMS) from the rest

C Ranked Genes by FSc

rank	Image Id.	Gene Description
1	770394	”Fc fragment of IgG, receptor, transporter, alpha”
2	1435862	”antigen identified by monoclonal antibodies 12E7, F21 and O13”
3	377461	”caveolin 1, caveolae protein, 22kD”
4	814260	follicular lymphoma variant translocation 1
5	491565	”Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2”
6	295985	ESTs
7	866702	”protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)”
8	1471841	”ATPase, Na+/K+ transporting, alpha 1 polypeptide”
9	841641	cyclin D1 (PRAD1: parathyroid adenomatosis 1)
10	713922	glutathione S-transferase M1
11	308497	KIAA0467 protein
12	52076	olfactomedinrelated ER localized protein
13	770868	NGFI-A binding protein 2 (ERG1 binding protein 2)
14	139957	ESTs
15	1470048	”lymphocyte antigen 6 complex, locus E”
16	842820	inducible poly(A)-binding protein
17	1473131	”transducin-like enhancer of split 2, homolog of Drosophila E(sp1)”
18	214572	ESTs
19	1323448	cysteine-rich protein 1 (intestinal)
20	345232	”lymphotoxin alpha (TNF superfamily, member 1)”

Table 12: Discriminated Class 1 (EWS) from the rest

rank	Image Id.	Gene Description
1	236282	Wiskott-Aldrich syndrome (ecezema-thrombocytopenia)
2	745019	EH domain containing 1
3	814526	ESTs
4	183337	”major histocompatibility complex, class II, DM alpha”
5	624360	”proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)”
6	47475	”Homo sapiens inducible protein mRNA, complete cds”
7	490772	small nuclear ribonucleoprotein polypeptide A’
8	785793	”capping protein (actin filament) muscle Z-line, alpha 1”
9	344134	immunoglobulin lambda-like polypeptide 3
10	868304	”actin, alpha 2, smooth muscle, aorta”
11	701751	cut (Drosophila)-like 1 (CCAAT displacement protein)
12	840942	”major histocompatibility complex, class II, DP beta 1”
13	21652	”catenin (cadherin-associated protein), alpha 1 (102kD)”
14	813742	PTK7 protein tyrosine kinase 7
15	68977	”proteasome (prosome, macropain) subunit, beta type, 10”
16	504791	glutathione S-transferase A4
17	897164	”catenin (cadherin-associated protein), alpha 1 (102kD)”
18	1469292	pim-2 oncogene
19	297392	metallothionein 1L
20	855487	N-acylsphingosine amidohydrolase (acid ceramidase)

Table 13: Discriminated Class 2 (BL) from the rest

rank	Image Id.	Gene Description
1	812105	transmembrane protein
2	786084	chromobox homolog 1 (Drosophila HP1 beta)
3	134748	glycine cleavage system protein H (aminomethyl carrier)
4	486110	profilin 2
5	81518	apelin; peptide ligand for APJ receptor
6	629896	microtubule-associated protein 1B
7	756401	Ras homolog enriched in brain 2
8	244637	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 45620
9	810057	cold shock domain protein A
10	325182	”cadherin 2, N-cadherin (neuronal)”
11	383188	recoverin
12	810864	”ESTs, Highly similar to CGI-48 protein [H.sapiens]”
13	544664	matrin 3
14	789376	thioredoxin reductase 1
15	823775	”guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3”
16	823886	”Smooth muscle myosin heavy chain isoform SMemb [human, umbilical cord, fetal aorta, mRNA Partial, 971 nt]”
17	878652	postmeiotic segregation increased 2-like 12
18	377048	”Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c”
19	308231	”Homo sapiens incomplete cDNA for a mutated allele of a myosin class I, myh-1c”
20	811161	ATP-binding cassette 50 (TNF-alpha stimulated)

Table 14: Discriminated Class 3 (NB) from the rest

rank	Image Id.	Gene Description
1	784224	fibroblast growth factor receptor 4
2	796258	”sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)”
3	142134	ESTs
4	143306	lymphocyte-specific protein 1
5	769716	neurofibromin 2 (bilateral acoustic neuroma)
6	789253	presenilin 2 (Alzheimer disease 4)
7	296448	insulin-like growth factor 2 (somatomedin A)
8	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
9	839552	nuclear receptor coactivator 1
10	244618	ESTs
11	79022	FBJ murine osteosarcoma viral oncogene homolog B
12	813841	”plasminogen activator, tissue”
13	898219	mesoderm specific transcript (mouse) homolog
14	859359	quinone oxidoreductase homolog
15	42558	glycine amidinotransferase (L-arginine:glycine amidinotransferase)
16	68950	cyclin E1
17	770059	heparan sulfate proteoglycan 2 (perlecan)
18	299737	Homo sapiens clone 24411 mRNA sequence
19	814444	”cofactor required for Sp1 transcriptional activation, subunit 9 (33kD)”
20	25725	farnesyl-diphosphate farnesyltransferase 1

Table 15: Discriminated Class 4 (RMS) from the rest

D Comparison of top S2N genes with Khan et al's 96 top genes

S2N Class	rank	Khan Rank	Khan Class	Image Id.	Gene Description
1	1	6	1(3)	770394	"Fc fragment of IgG, receptor, transporter, alpha"
1	2	18	1(6)	377461	"caveolin 1, caveolae protein, 22kD"
1	3	75	1(9)	814260	follicular lymphoma variant translocation 1
1	4	73	1(14)	1435862	"antigen identified by monoclonal antibodies 12E7, F21 and O13"
1	5	10	not 1(1)	295985	ESTs
1	6	15	1(2)	866702	"PTP, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)"
1	7			491565	"Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2"
1	8			1471841	"ATPase, Na+/K+ transporting, alpha 1 polypeptide"
1	9	19	1(7)	52076	olfactomedinrelated ER localized protein
1	10	3	1(11)/3(118)	841641	cyclin D1 (PRAD1: parathyroid adenomatosis 1)
2	1			236282	Wiskott-Aldrich syndrome (ecezema-thrombocytopenia)
2	2			745019	EH domain containing 1
2	3	23	2(8)	183337	"major histocompatibility complex, class II, DM alpha"
2	4	59	not 2(24)	504791	glutathione S-transferase A4
2	5	78	2(105)/4(198)	814526	ESTs
2	6	55	not 2(15)	21652	"catenin (cadherin-associated protein), alpha 1 (102kD)"
2	7			624360	"proteasome(prosome, macropain)subunit, beta type, 8(large multifunctional protease 7)"
2	8			47475	"Homo sapiens inducible protein mRNA, complete cds"
2	9			813742	PTK7 protein tyrosine kinase 7
2	10	66	not 2(20)	897788	"protein tyrosine phosphatase, receptor type, F"
3	1	22	3(2)	812105	transmembrane protein
3	2			786084	chromobox homolog 1 (Drosophila HP1 beta)
3	3			134748	glycine cleavage system protein H (aminomethyl carrier)
3	4	72	3(5)	325182	"cadherin 2, N-cadherin (neuronal)"
3	5	54	3(31)	486110	profilin 2
3	6	11	3(1)	629896	microtubule-associated protein 1B
3	7			810057	cold shock domain protein A
3	8			81518	apelin; peptide ligand for APJ receptor
3	9			756401	Ras homolog enriched in brain 2
3	10			244637	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 45620
4	1	68	4(5)	784224	fibroblast growth factor receptor 4
4	2	89	4(10)	796258	"sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)"
4	3	7	4(3)	244618	ESTs
4	4			769716	neurofibromin 2 (bilateral acoustic neuroma)
4	5			789253	presenilin 2 (Alzheimer disease 4)
4	6			839552	nuclear receptor coactivator 1
4	7			143306	lymphocyte-specific protein 1
4	8	2	4(2)	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
4	9	1	4(1)	296448	insulin-like growth factor 2 (somatomedin A)
4	10			142134	ESTs

Table 16: Top 10 S2N genes discriminating each class and the corresponding Khan et al's overall rank (top 96) and the category in which they are highly expressed; brackets indicate the rank of the gene for that category [1].

E Comparison of top SVMFS genes with Khan et al’s 96 top genes

SVMFS Class	rank	Khan Rank	Khan Class	Image Id.	Gene Description
1	1			139957	ESTs
1	2			754649	chromosome 14 open reading frame 3
1	3			302933	nucleolin
1	4			897690	tumor rejection antigen (gp96) 1
1	5			50887	ESTs
1	6			296448	insulin-like growth factor 2 (somatomedin A)
1	7	15	1(2)	866702	”protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)”
1	8	7	4(3)	244618	ESTs
1	9			379708	
1	10			811000	”lectin, galactoside-binding, soluble, 3 binding protein (galectin 6 binding protein)”
2	1	83	2(71)	868304	”actin, alpha 2, smooth muscle, aorta”
2	2	12	2(12)	840942	”major histocompatibility complex, class II, DP beta 1”
2	3	8	not 2(1)	233721	insulin-like growth factor binding protein 2 (36kD)
2	4	57	1(166)/2(153)	745343	”regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)”
2	5			47475	”Homo sapiens inducible protein mRNA, complete cds”
2	6	13	2(3)	80109	”major histocompatibility complex, class II, DQ alpha 1”
2	7			1461138	”H4 histone family, member G”
2	8	58	2(27)	241412	E74-like factor 1 (ets domain transcription factor)
2	9			344134	immunoglobulin lambda-like polypeptide 3
2	10	36	not 2(4)	1416782	”creatine kinase, brain”
3	1	11	3(1)	629896	microtubule-associated protein 1B
3	2	22	3(2)	812105	transmembrane protein
3	3			878652	postmeiotic segregation increased 2-like 12
3	4			810057	cold shock domain protein A
3	5	31	3(3)	44563	growth associated protein 43
3	6	30	3(17)	82225	secreted frizzled-related protein 1
3	7	68	4(5)	784224	fibroblast growth factor receptor 4
3	8	47	3(37)	135688	GATA-binding protein 2
3	9	72	3(5)	325182	”cadherin 2, N-cadherin (neuronal)”
3	10	4	1(25)/4(69)	365826	growth arrest-specific 1
4	1	2	4(2)	207274	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
4	2	68	4(5)	784224	fibroblast growth factor receptor 4
4	3	1	4(1)	296448	insulin-like growth factor 2 (somatomedin A)
4	4			840708	”superoxide dismutase 2, mitochondrial”
4	5			882522	argininosuccinate synthetase
4	6			302933	nucleolin
4	7	7	4(3)	244618	ESTs
4	8			878798	beta-2-microglobulin
4	9			810512	thrombospondin 1
4	10	19	1(7)	52076	olfactomedinrelated ER localized protein

Table 17: Top 10 SVMFS genes discriminating each class and the corresponding Khan et al’s overall Rank (top 96) and the category in which they are highly expressed; brackets indicate the rank of the gene for that category [1].

References

[1] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson and P. Meltzer *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Nature Medicine, **7**, Number 6, 673-679, June 2001

[2] C. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov and T. Golub *Molecular classification of multiple tumor types* , Bioinformatics. **17**, Suppl. 1 ISMB 2001 S316

[3] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caliguiri, C. Bloomfield and E. Lander *Molecular classification of cancer: Class discovery and Class prediction by Gene expression Monitoring*, Science **286**, 531-537, 1999

[4] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov and T. Poggio *Support Vector Machine Classification of Microarray Data*, MIT AI MEMO No. 1677, CBCL Paper No. 182, 1998

[5] Duda R. et al. *Pattern Classification and Scene Analysis*, Wiley 1973

[6] Pavlidis P. et al. *Gene Functional Analysis from Heterogeneous Data*, submitted

[7] Blanz V. et al. *Comparison of view-based object recognition algorithms using realistic 3D models*, Artificial Neural Networks -ICANN'96 251-256, Berlin 1996. Springer Lecture Notes in Computer Science **1112**