Local Learning Based Feature Selection for High Dimensional Data Analysis (Supplementary Data)

Yijun Sun[†,*], Sinisa Todorovic[‡], Steve Goodison[§]
[†]Interdisciplinary Center for Biotechnology Research
University of Florida, Gainesville, FL 32610
[‡]School of EECS, Oregon State University, Corvallis, OR 97331
[§]Cancer Research Institute, M. D. Anderson Cancer Center, Orlando, FL 32827

1 Using the Proposed Algorithm for Classification Purposes

The proposed algorithm can be easily extended for classification purposes. Given a training data set, the algorithm first learns a feature weight vector, then computes the averaged distances between a test sample and the training samples with the positive and negative class labels, respectively, and assigns the test sample to the class with a smaller distance. We perform some experiments on the UCI datasets contaminated by varying numbers of irrelevant features ranging from 0 to 10000. The kernel width and regularization parameter are estimated through ten-fold cross validation using the training data. The classification errors averaged over 10 runs and the standard deviations are reported in Table 1. We observe that the performance of our algorithm is largely insensitive to a growing number of irrelevant features. For comparison, the classification errors of SVM, KNN and C4.5 performed on the UCI datasets containing 5000 irrelevant features are presented in Table 2. The latter three algorithms clearly suffer from the curse of dimensionality.

2 Experiments Using Simba and Gflip

We perform some experiments to compare our algorithm with the well-known Simba and Gflip algorithms [3]. Both algorithms are also based on local learning. Our work is, in part, motivated by the Simba algorithm. Compared to RELIEF, Simba re-evaluates the

[*]Please address all correspondence to: Dr. Yijun Sun, Interdisciplinary Center for Biotechnology Research, University of Florida, P.O. Box 103622, Gainesville, FL 32610-3622, USA. E-mail: sunyijun@biotech.ufl.edu.

Table 1: Classification errors and standard deviations (%) obtained by using our algorithm performed on the seven UCI datasets containing a varying number of irrelevant features, ranging from 0 to 10000. The classification errors are nearly insensitive to the growing number of features.

	Number of Irrelevant Features					
	0	100	500	1000	5000	10000
diabetes	24.8(1.8)	24.9(1.8)	24.0(2.0)	24.3(1.7)	25.3(2.3)	24.8(2.1)
heart	17.5(4.7)	17.5(4.6)	17.6(4.5)	17.6(4.5)	17.8(4.5)	17.7(4.2)
splice	10.8(1.4)	10.8(1.5)	10.7(1.3)	10.8(1.2)	10.6(1.7)	10.5(0.9)
thyroid	6.1(1.8)	6.0(1.7)	6.1(1.8)	6.1(2.0)	6.2(1.8)	6.4(1.5)
waveform	13.8(0.8)	13.9(0.9)	13.8(0.9)	13.9(0.8)	14.7(1.0)	14.9(1.0)
banana	11.2(0.6)	11.2(0.6)	11.2(0.6)	11.2(0.7)	11.2(0.6)	11.3(0.6)
twonorm	4.6(0.5)	4.6(0.6)	4.6(0.6)	4.6(0.5)	4.6(0.6)	4.6(0.6)

Table 2: Classification errors and standard deviations (%) obtained by using SVM, KNN, C4.5 and our algorithm performed on the seven UCI datasets containing 5000 irrelevant features. The three competing algorithms clearly suffer from the curse of dimensionality.

	Methods			
	SVM	KNN	C4.5	Our Algorithm
diabetes	34.3(1.9)	36.7(2.0)	36.2(3.7)	**25.3(2.3)**
heart	35.4(3.4)	34.8(4.5)	35.8(4.1)	**17.8(4.5)**
splice	40.5(1.7)	40.3(1.9)	14.0(1.9)	**10.6(1.7)**
thyroid	30.1(3.9)	32.3(3.7)	12.0(3.7)	**6.2(1.8)**
waveform	32.9(0.2)	31.0(0.9)	23.9(1.5)	**14.7(1.0)**
twonorm	35.0(1.0)	36.4(1.7)	28.4(2.4)	**4.6(0.6)**
banana	32.9(0.2)	49.2(1.7)	40.6(11.5)	**11.2(0.6)**

distances according to learned weight vectors, and thus is superior to RELIEF. One major problem with Simba and Gflip, however, is their implementation. The objective function optimized by Simba/Gflip is characterized by many local minima. This problem is mitigated in Simba/Gflip by restarting the algorithms from several different starting points. Nevertheless, the reach of a global optimal solution is not guaranteed. The codes of Simba and Gflip are downloaded from [3]. The non-linear sigmoid activation function is used. The number of starting points is set to 5, while the default values of the original codes are 5 and 1, respectively. Also, we set the number of passes of the training data to be 5, the default value of which is 1. All other parameters use their default values. The computational complexity of Gflip and Simba are $\mathcal{O}(N^2J^2)$ and $\mathcal{O}(N^2J)$, respectively. Here, J is the number of features and N is the number of samples. When $J \gg N$, Gflip is computationally much more expensive than Simba. Due to computational reasons, Gflip is run on the UCI and spiral data containing only 500 irrelevant features. The feature weights learned by Gflip and Simba are plotted in Figs. 1 and 2, respectively. For comparison, the feature weights learned by Simba performed on the UCI data with only 100 irrelevant features are also plotted in Fig. 3. Gflip can only provide information of whether a feature is selected (1) or not selected (0). From the figures, we can see that Gflip performs much worse than Simba, while Simba performs very well when the number of irrelevant features is small, but may fail completely when the number of irrelevant features become excessively large (for example, *banana*, *spiral* and *diabetes*). One possible explanation is that the chance for Simba to be stuck into local minima is increased dramatically with the increased number of features. In contrast, our algorithm is not sensitive to the number of features (see Fig. 4). The CPU times of Gfilp and Simba are reported in Table 3. As expected, Gflip is computationally much more intensive than Simba and our algorithm. The computational complexity of Simba is greater than ours, partially due to the fact that Simba restarts the algorithm from five different initial points to alleviate the local-minima problem.

3 Experiments Using AMS

AMS [1], along with RFE [2], is among the first to perform feature selection directly in the SVM formulation. The basic idea of AMS to automatically tune the scaling parameters of a kernel by minimizing some generalization error bounds. The code is downloaded from [1]. The default settings of the algorithm are used, and the span bound is minimized. Due to computational reasons, AMS is only applied to the UCI and spiral data with only 1000 irrelevant features. The learned feature weights are plotted in Fig. 5. AMS performs very well to identify the useful features (except for *spiral*), but leads to many false positives. Moreover, AMS is computationally much more expensive than both Simba and ours (see

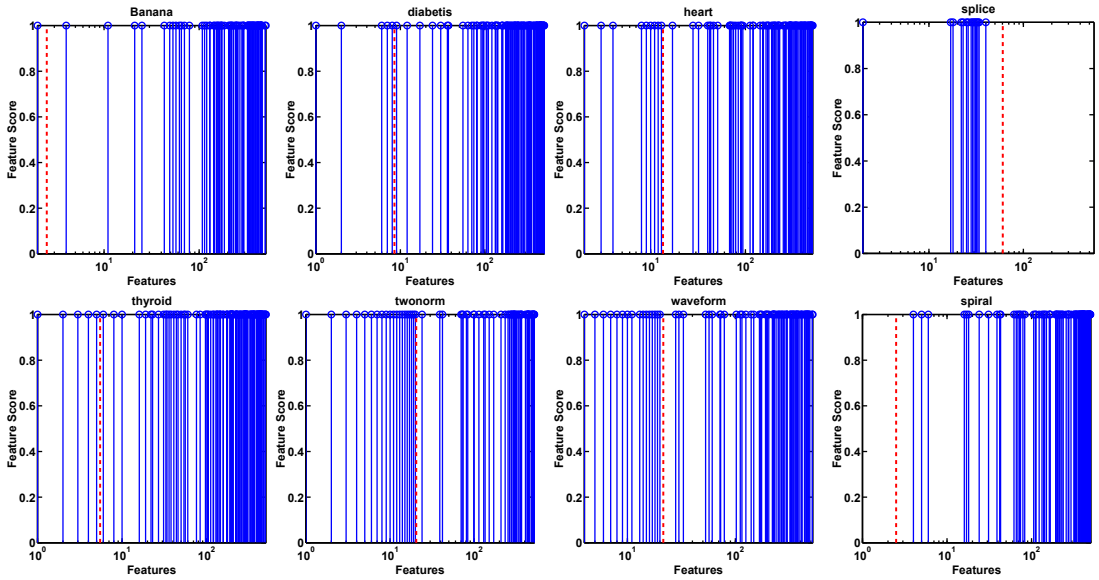

Figure 1: Feature weights learned by Gflip in one sample trial of the spiral and seven UCI datasets with 500 irrelevant features. The red dashed line indicates the number of original features. The weights plotted on the left side of the dashed line are associated with the original features, while those on the right, with the additional 500 irrelevant features.

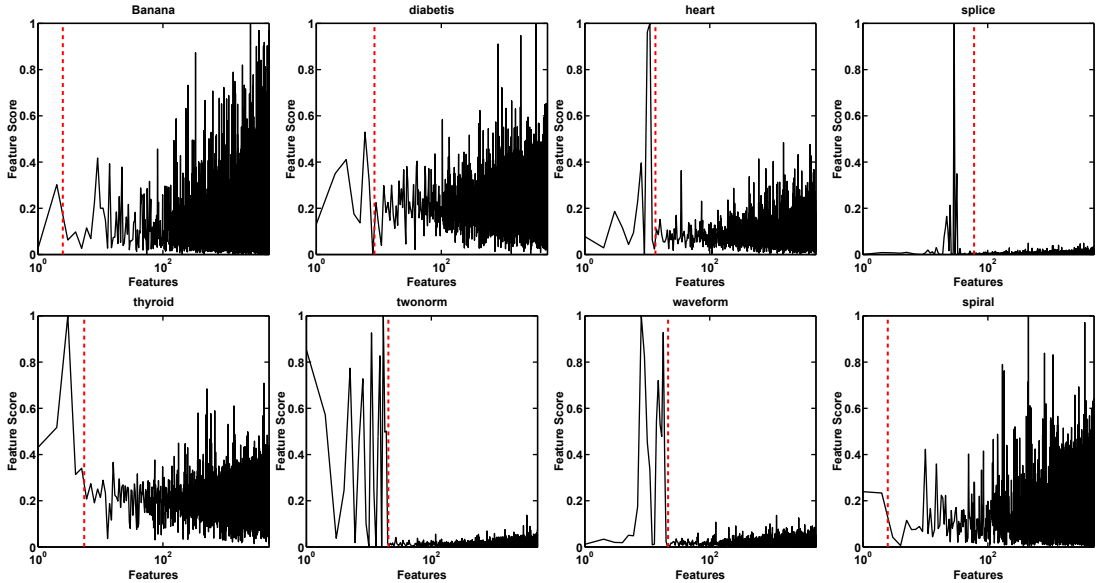

Figure 2: Feature weights learned by Simba in one sample trial of the spiral and seven UCI datasets with 5000 irrelevant features.

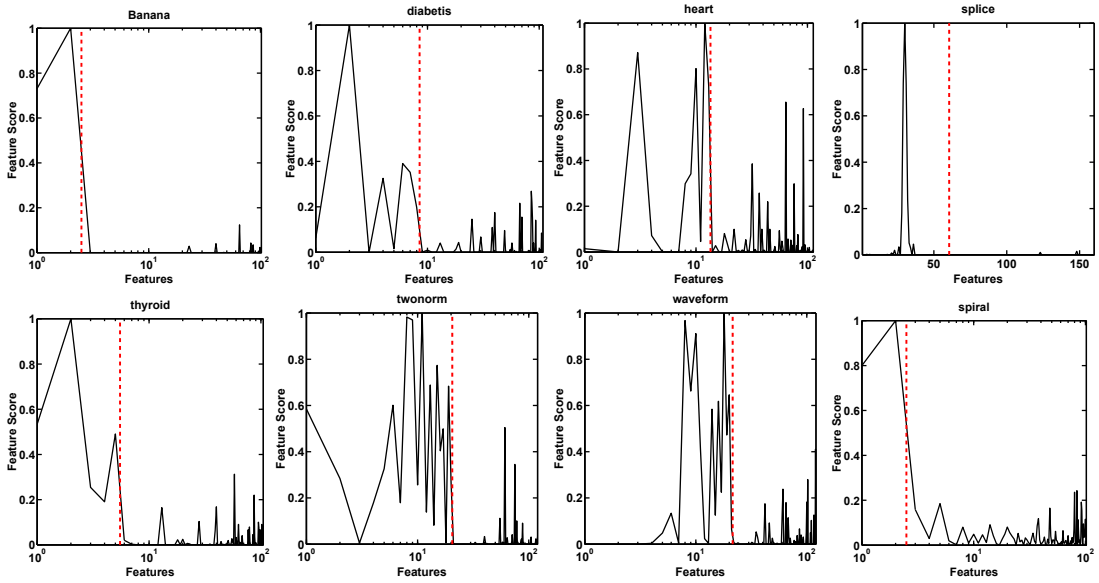

Figure 3: Feature weights learned by Simba in one sample trial of the spiral and seven UCI datasets with 100 irrelevant features.

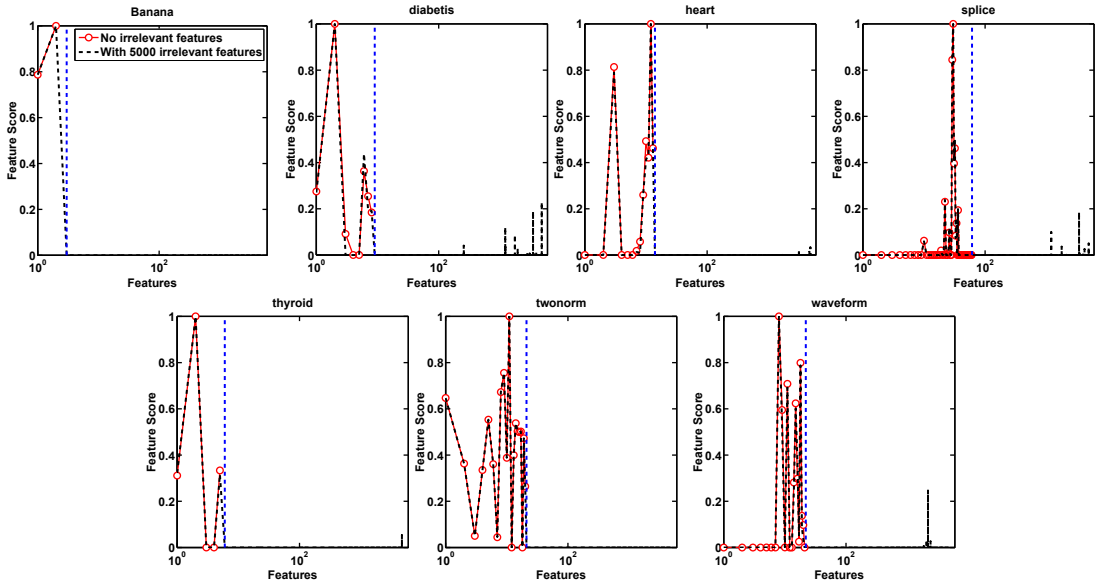

Figure 4: Feature weights learned by our algorithm in one sample trial of seven UCI datasets with and without 5000 irrelevant features. The performance of our algorithm is nearly insensitive to the presence of 5000 irrelevant features. The results empirically confirm that (1) our algorithm is a fixed-point method; and (2) the algorithm has a logarithmical sample complexity. The figure is better viewed electronically.

Table 3). (The computational complexity of both Simba and our algorithm is linear with respect to the number of features.)

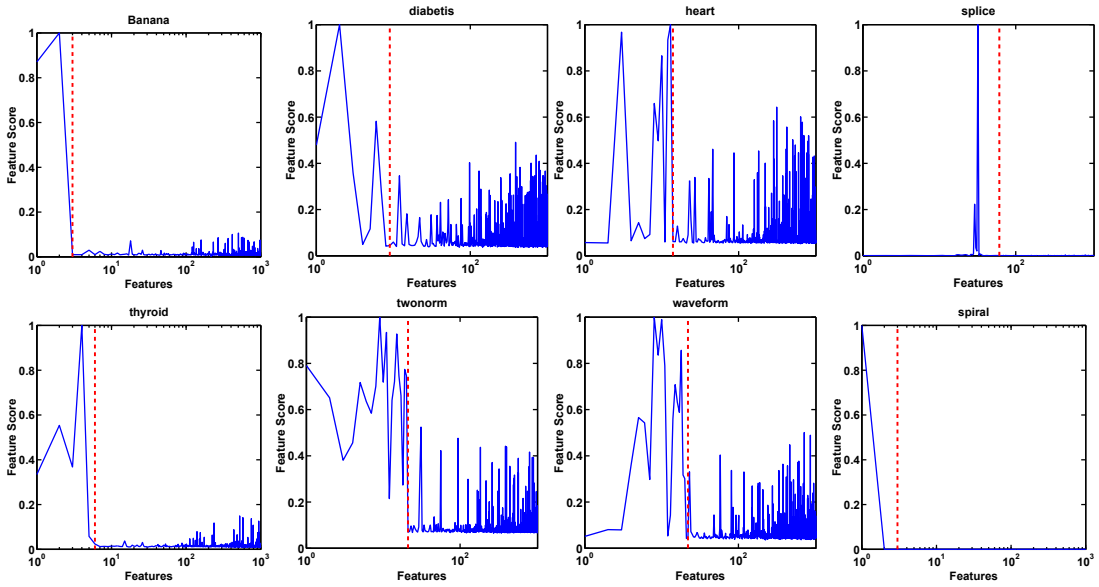

Figure 5: Feature weights learned by AMS in one sample trial of the spiral and seven UCI datasets with 1000 irrelevant features.

4 More Experimental Results on Breast Cancer Study

We perform an experiment on the breast cancer data that shows that the choice of the kernel width is not critical, and our algorithm yields nearly identical prediction performance for a wide range of sigma values (Fig. 6).

5 Prostate Cancer Study

We conduct a computational study to investigate whether a genetic based model can outperform a clinical nomogram[1] for predicting the recurrence of prostate cancer after radical prostatectomy, and the combination of nomogram and genetic information can lead to an improved prognostic performance (i.e., using nomogram prediction scores as a feature along with microarray data). The gene expression and clinical data used in the study are provided by the senior author Dr. William Gerald of [5]. The gene expression data was built from tissue samples obtained from 79 patients with clinically localized prostate cancer treated by

[1]Nomogram is a commonly used clinical tool for predicting prostate cancer recurrence after radical prostatectomy. See, for example, [4].

Table 3: CPU times (in seconds) of four algorithms performed on the spiral and seven UCI datasets. The number of irrelevant features is indicated in the parentheses. The computational complexity of both Simba and our algorithm is linear with respect to the number of features.

	Methods			
	AMS (1000)	Simba (5000)	Gflip (500)	Our Algorithm (5000)
spiral	5672	516	246	139
diabetes	3335	507	271	267
heart	477	79	31	73
splice	6475	2263	1326	330
thyroid	353	60	24	13
waveform	1429	389	205	157
twonorm	1120	373	204	162
banana	2345	413	198	97

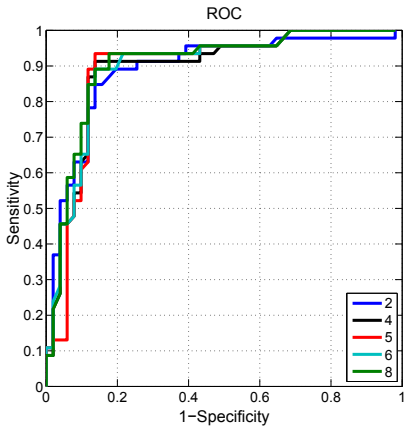

Figure 6: ROC curves of the breast cancer prognostic systems using the gene signatures identified by using different sigma values, ranging from 2 to 8. All prognostic systems perform very similarly.

radical prostatectomy at MSKCC between 1993 and 1999. Thirty-nine cases had disease recurrence as classified by 3 consecutive increases in the serum level of prostate specific antigen after radical prostatectomy, and forty samples were classified as non-recurrent samples by virtue of maintaining an undetectable prostate specific antigen (< 0.05 ng/mL) for at least 5 years after radical prostatectomy. No patient received any neo-adjuvant or adjuvant therapy before documented disease recurrence. The complete clinical characteristics of the 79 primary tumors are listed in [5].

The experimental procedures are exactly the same as those in the breast cancer study presented in the main text. Fig. 7 presents the ROC curves comparing the prediction performance of the nomogram, genetic and hybrid (combination of nomogram and genetic) models constructed by using our algorithm, and the ROC curves obtained by using SVM-RFE, norm-1 regularized logistical regression and AMS. Our algorithm outperforms the three competing algorithms. However, the results do suggest that using advanced computational algorithms to combine both nomogram and genetic information can indeed improve the prognosis performance of prostate cancer. The clinical implications of the results and the biological significance of the identified genes are discussed elsewhere.

References

[1] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 131-159, 2002.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[3] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection - theory and algorithms," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 43–50.

[4] M. L. Blute, E. J. Bergstralh, A. Iocca, B. Scherer, and H. Zincke, "Use of gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy," *J. Urol.*, vol. 165, no. 1, pp. 119–125, 2001.

[5] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald, "Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy," *Cancer*, vol. 104, no. 2, pp. 290–298, 2005.

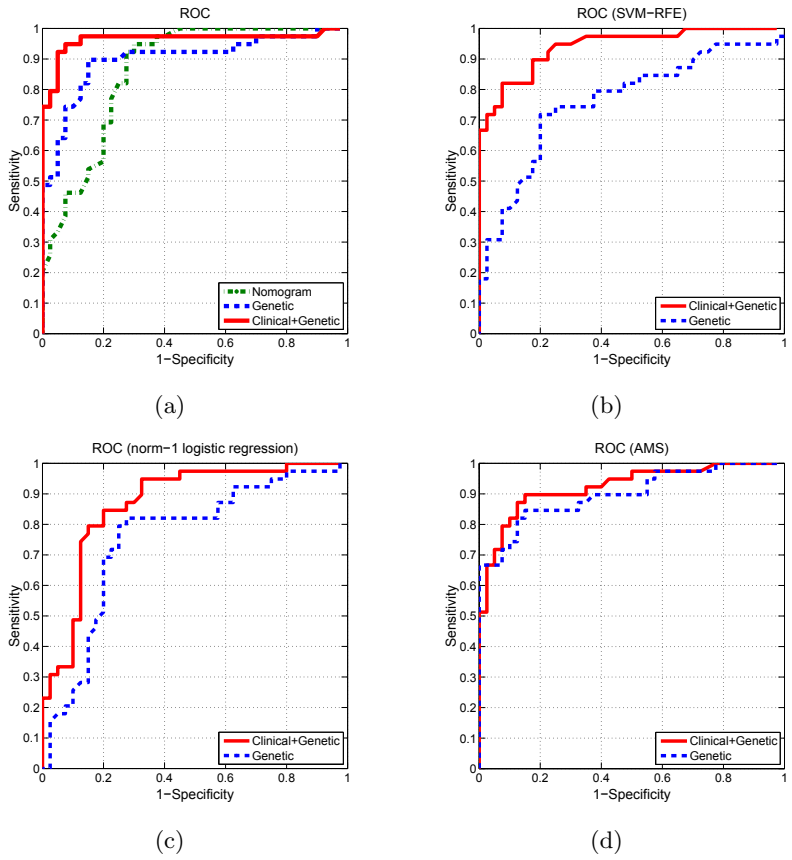

Figure 7: (a) Receiver operating characteristic (ROC) plot comparing the prediction performance of the nomogram, genetic and hybrid (combination of nomogram and genetic) models constructed by using our algorithm. (b-d) ROC curves obtained by using SVM-RFE, norm-1 regularized logistical regression and AMS.