Correction

Correction: Multiclass classification of microarray data with repeated measurements: application to cancer

Ka Yee Yeung and Roger E. Bumgarner

Address: Department of Microbiology, Box 358070, University of Washington, Seattle, WA 98195, USA.

Correspondence: Ka Yee Yeung. Email: kayee@u.washington.edu

Published: 3 January 2006

Genome ***Biology*** 2005, **6:**405 (doi:10.1186/gb-2005-6-13-405)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/13/405

After the publication of this work [1], we discovered programming errors in our software implementation of the proposed error-weighted, uncorrelated shrunken centroid (EWUSC) algorithm and the uncorrelated shrunken centroid (USC) algorithm. We have corrected these errors, and the updated results are summarized in the revised Table 6.

On the NCI 60 data, both Figure 1 in [1] and the revised Figure 1 showed that USC generally produces higher prediction accuracy than the 'shrunken centroid' algorithm (SC) [2] using the same number of relevant genes. Using the revised software implementation, USC requires fewer (2,116 instead of 2,315 as reported in [1]) genes to achieve 72% accuracy. The number of genes required by SC to achieve the same prediction accuracy remains the same (3,998).

Figure 2 shows the results of applying EWUSC to the training set, four-fold cross-validation data, and test set of the multiple tumor data over a range of shrinkage thresholds (Δ) and correlation thresholds (ρ_0). The revised Figure 2 shows the same general trend as Figure 2 in [1]: the percentage of errors is reduced when $\rho_0 < 1$ over most values of Δ on the training set, cross-validation data and test set; Figure 2d shows that the number of relevant genes is drastically reduced when genes with correlation threshold above 0.9 are removed. The values of the optimal shrinkage thresholds (Δ) determined from the cross-validation results have changed using the revised implementation. Specifically, the optimal shrinkage threshold values (Δ) for both EWUSC and USC are reduced to 4.8 and 4 respectively (see revised Table 6). The numbers of relevant genes selected by EWUSC and USC are reduced and the resulting prediction accuracy for both USC and SC are also reduced in the revised results. In the case of using the global optimal parameters when Δ = 0, the EWUSC in the revised implementation selected slightly fewer genes (1,622 instead of 1,626) at the expense of slightly lower prediction accuracy (74% instead of 78%). Figure 4 compares the prediction accuracy on the test set of the multiple tumor data using the EWUSC and USC algorithms at the estimated optimal correlation threshold (ρ_0 = 0.8), the SC algorithm and the Support Vector Machine (SVM). The general observations previously reported in [1] still hold with the revised Figure 4. First, USC produces higher prediction accuracy than SC using the same number of relevant genes. Second, EWUSC generally produces higher prediction accuracy than USC using the same number of relevant genes. In fact, the performance of EWUSC is stronger than previously reported in [1] when the number of genes is small.

Figure 5 shows the comparison of prediction accuracy of EWUSC, USC, and SC on the breast cancer data. With the

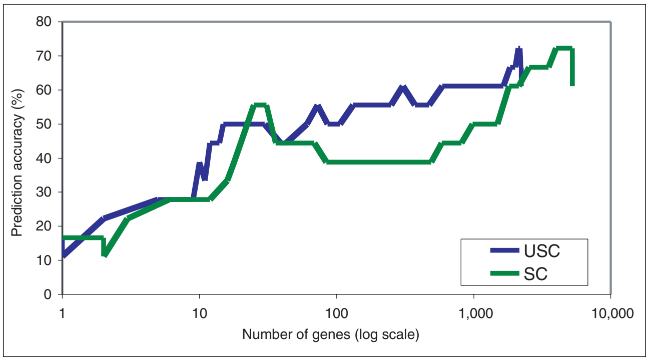

Figure 1
A corrected figure showing the comparison of prediction accuracy of USC and SC on the NCI 60 data. The percentage of prediction accuracy is plotted against the number of relevant genes using the USC algorithm at ρ_0 = 0.6 and the SC algorithm (USC at ρ_0 = 1.0). The horizontal axis is shown on a log scale. Because no independent test set is available for this data, we randomly divided the samples in each class into roughly three parts multiple times, such that a third of the samples are reserved as a test set. Thus the training set consists of 43 samples and the test set of 18 samples. The graph represents typical results over these multiple random runs.

Table 6

Summary of prediction accuracy results

Data	Parameters	**EWUSC**	USC	SC	Published results
NCI 60 data*	ρ_0	NA	0.6	1.0	NA
	Δ	NA	0.6	**0.9**	NA
	# relevant genes	NA	***2,116*** **(2315)**	3,998	200
	Prediction accuracy	NA	72%	72%	~ 40-60% [4]
Multiple tumor data (estimated optimal parameters) †	ρ_0	0.8	0.8	1.0	NA
	Δ	**4.8 (5.6)**	**4 (5.6)**	8.8	NA
	# relevant genes	***241*** **(680)**	***356*** **(735)**	3902	All genes
	Prediction accuracy	93%	**82%** (85%)	**63%** (78%)	78% [5]
Multiple tumor data (global optimal parameters) ‡	ρ_0	0.9	0.9	1.0	NA
	Δ	0	0	0.4	NA
	# relevant genes	***1,622*** **(1626)**	1634	7129	All genes
	Prediction accuracy	**74% (78%)**	74%	**59%** (74%)	78% [5]
Breast cancer data	ρ_0	**0.6 (0.7)**	0.6	1.0	NA
	Δ	**0.80**	0.55 (1.15)	0.5 (1.1)	NA
	# relevant genes	***189*** **(271)**	**1,114 (82)**	**3,193** (187)	70
	Prediction accuracy	**84% (89%)**	***84*****% (79%)**	84%	89% [6]

Results different from those previously reported are highlighted in bold. Previous results are in brackets. Results improved over previously reported are highlighted in italic, while results worse than previously reported are underlined. The optimal parameters (ρ_0 and Δ), number of relevant genes chosen, and prediction accuracy for the NCI 60 data, multiple tumor data and breast cancer data are summarized here. Both EWUSC (error-weighted, uncorrelated shrunken centroid) and USC (uncorrelated shrunken centroid) were motivated by SC (shrunken centroid) [2]. Both EWUSC and USC take advantage of interdependence between genes by removing highly correlated relevant genes. EWUSC makes use of error estimates or variability over repeated measurements. SC [2] is equivalent to USC at $\rho_0 = 1$. The optimal parameters (Δ, ρ_0) for EWUSC are estimated from the cross-validation results of EWUSC, while the optimal parameters (Δ, ρ_0) for USC are independently estimated from the cross-validation results of USC. *Since no repeated measurements or error estimates are available, EWUSC is not applicable to the NCI 60 data. In addition, there is no separate test set available for the NCI 60 data, typical results of random partitions of the original 61 samples into training and test sets are shown. †The prediction accuracy and number of relevant genes are produced using optimal parameters (Δ, ρ_0) estimated by visual observation of 'bends' in the random cross-validation curves. ‡The prediction accuracy and number of relevant genes are produced using global optimal parameters, that is (Δ, ρ_0) that produces the minimum average numbers of cross-validation errors over all Δ and all ρ_0.

revised implementation, the optimal correlation threshold (ρ_0) is changed from 0.7 in [1] to 0.6 (see revised Table 6). The observation reported in [1] that EWUSC produces higher prediction accuracy on the test set than USC and SC when the number of relevant genes is small still holds. The numbers of relevant genes selected by USC and SC are significantly larger with the revised implementation (see revised Table 6).

The major conclusions and observations in the original manuscript [1] remain valid with the revised implementation. Our EWUSC and USC algorithms represent improvements over the SC algorithm. In general, fewer genes are required to produce comparable prediction accuracy. On the multiple tumor data, our EWUSC and USC algorithms produce higher prediction accuracy using fewer relevant genes compared to published results. The revised software implementation is available on our web site [3]. Note: the revised version (1.0) of the software was placed on the web site on May 9, 2005.

References

1. Yeung KY, Bumgarner RE: **Multiclass classification of microarray data with repeated measurements: application to cancer**. *Genome Biol* 2003, **4:**R83.
2. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA* 2002, **99:**6567-6572.
3. **Supplementary Web Site: Multiclass classification of microarray data with repeated measurements: application to cancer.** [http://www.expression.washington.edu/publications/kayee/shrunken_centroid]
4. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *J Am Stat Assoc* 2002, **97:**77-87.
5. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al.*: **Multiclass cancer diagnosis using tumor gene expression signatures**. *Proc Natl Acad Sci USA* 2001, **98:**15149-15154.
6. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al.*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415:**530-536.

(Continues on the next page)

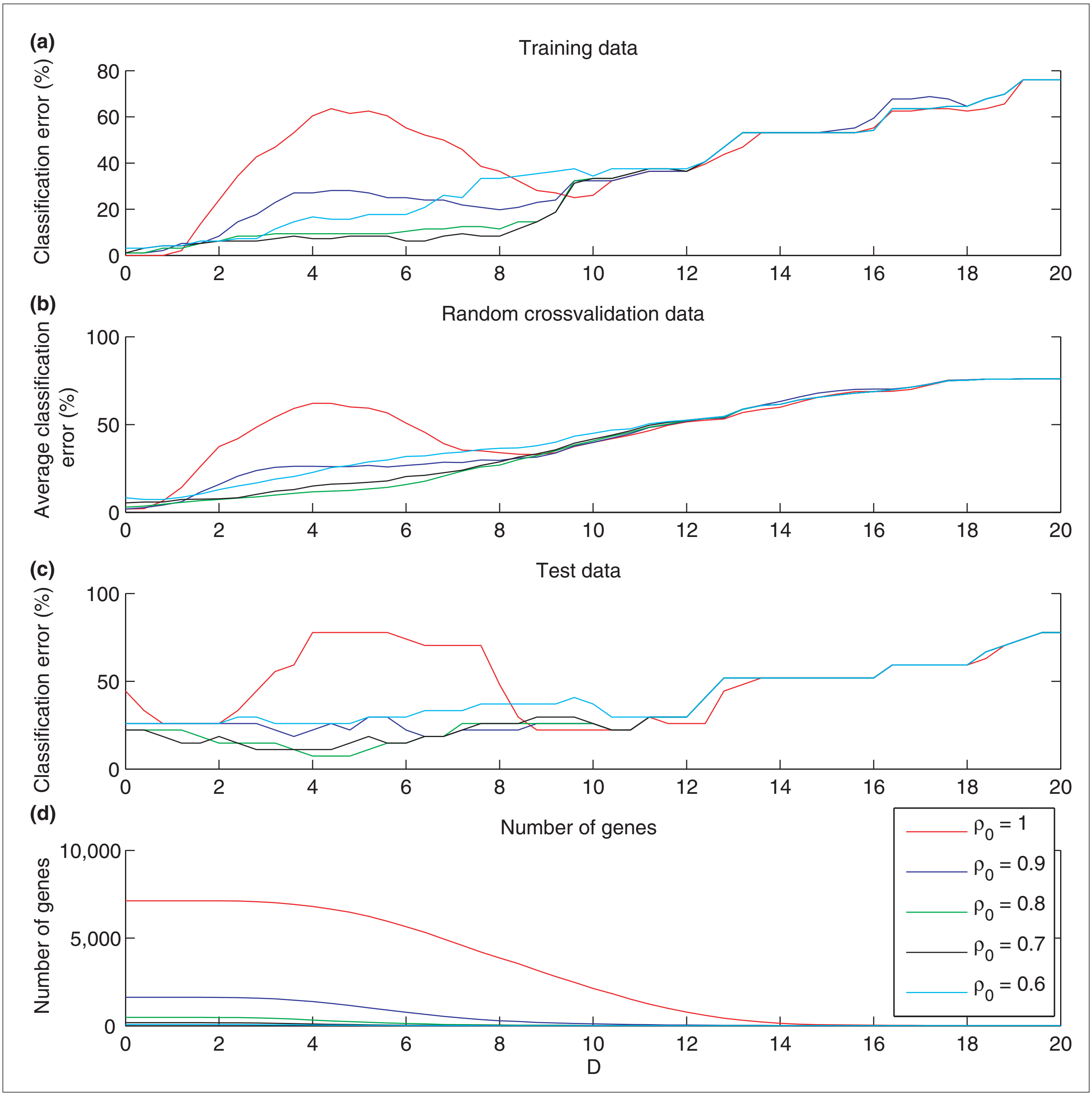

Figure 2

A corrected figure showing the prediction accuracy on the multiple tumor data using the EWUSC algorithm over the range of Δ from 0 to 20. The percentage of classification errors is plotted against Δ on **(a)** the full training set (96 samples) and **(c)** the test set (27 samples). In **(b)** the average percentage of errors is plotted against Δ on the cross-validation data over five random runs of fourfold cross-validation. In **(d)**, the number of relevant genes is plotted against Δ. Different colors are used to specify different correlation thresholds (ρ_0 = 0.6, 0.7, 0.8, 0.9 or 1). Optimal parameters are inferred from the cross-validation data in (b).

(Continues on the next page)

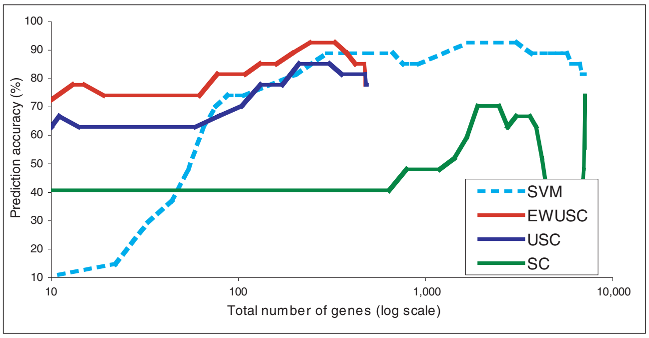

Figure 4

A corrected figure showing the comparison of prediction accuracy of EWUSC ($\rho_0 = 0.8$), USC ($\rho_0 = 0.8$), SVM and SC algorithms on the multiple tumor data. The horizontal axis shows the total number of distinct genes selected over all binary SVM classifiers on a log scale. Some results are not available on the full range of the total number of genes. For example, the maximum numbers of selected genes for EWUSC and USC are roughly 1,000. The reported prediction accuracy is 78% [5] using all 16,000 available genes on the full data. The EWUSC algorithm achieves 85% prediction accuracy with only 77 genes. With 241 genes, EWUSC produces 93% prediction accuracy.

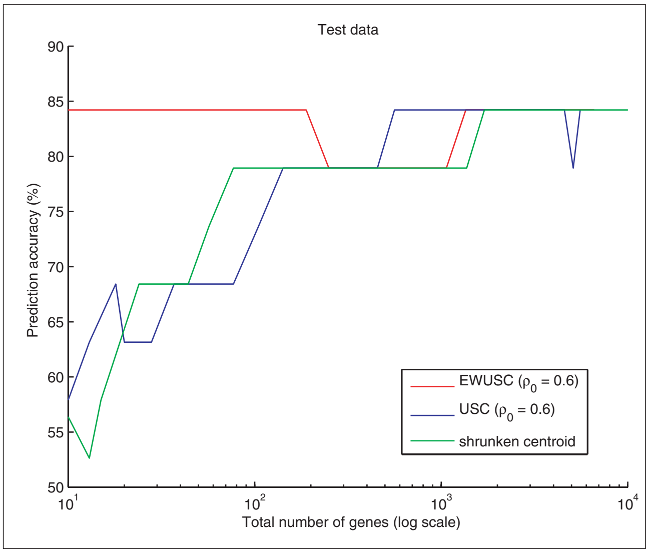

Figure 5

A corrected figure showing the comparison of prediction accuracy of EWUSC, USC and SC on the breast cancer data. The percentage of prediction accuracy is plotted against the number of relevant genes using the EWUSC algorithm at $\rho_0 = 0.6$, the USC algorithm at $\rho_0 = 0.6$ and the SC algorithm (USC at $\rho_0 = 1.0$). Note that the horizontal axis is shown on a log scale.