Biomarker and Classifier Selection in Diverse Genetic Datasets

Edward Hemphill, James Lindsay, Chih Lee, Craig Nelson and Ion Măndoiu

Computer Science and Engineering
Molecular and Cell Biology
University of Connecticut
{james.lindsay,chihlee,ion}@engr.uconn.edu
{edward.hemphill,craig.nelson}@uconn.edu

Abstract. Biomarker panels are increasingly important clinical tools for the classification of diseased tissue samples and have been more recently been used for characterizing differentiating stem cell cultures. In order to facilitate high sample throughput biomarker panels are limited to a finite number of hand-picked genes deemed to be of significance by the researcher. However, without statistical support that the most informative biomarkers have been selected, biomarker panels can be subject to extensive sampling bias that can result in misclassification and wasted resources. Moreover, the accurate mapping of marker profiles to discrete classes is not always straightforward. Here we present a pipeline for the rational design and interpretation of biomarker panels from underlying biological databases.

1 Introduction

Over the past decade advances in genetic technologies have enabled the high-throughput screening of millions of biomarkers for low-cost. These biomarkers can be anything from phenotypic observations, protein, gene, SNP, CNV microarrays to a plethora of sequencing based measurements such as RNA-seq and ChIP-seq. Biomarker panels can be used to characterize biological entities and activities based on an underlying knowledge base. Typically researchers attempt to reduce the dimensionality of their data in a process known as *feature* selection and then create predictive *models* based on only a few biomarkers. There have been numerous publications describing approaches for both tasks in genetic and medicinal applications. Unfortunately many of these publications demonstrate methodology in highly specified context, e.g. disease classification using gene-expression microarray data.

Each biomarker technology provides a unique type of data with specific properties. Gene-expression microarrays measure the same 20,000 (or so) genes in every experiment and therefore produce dense, continuous and moderately high-dimension data. However *in situ hybridization* experiments or literature search may produce sparse, binary low-dimensional data. Understanding which feature selection algorithm, predictive models and parameter combination that works

best in a particular scenario can be a daunting, non-intuitive task. A recent publication [8] performed an in-depth comparison of many common feature selection and classification algorithms on 4 Affymetrix HG-U133A gene expression arrays. The paper was very insightful however it did not interrogate dataset specific effects on the various combinations of feature selection and classifiers.
The work presents a simple method for the design of biomarker panels combined with statistical interpretation of data that will move the field of biomarker analysis from manual curation to computationally supported rational design.

2 Methods

Feature Selection Algorithms

We have selected four feature selection algorithms to compare. One of the most common is Support Vector Machine (SVM)-recursive feature elimination (RFE) [9] which utilizes an SVM classifier to choose features that improve the predictive power in a greedy fashion. Another method computes the ANOVA F-values for each feature, selecting the desired number of features with the best F-values. The two final methods are tree based algorithms, Random Forests [2] and Extra Trees [7].

Predictive Models

There are numerous approaches to building predictive models for multi-class classification scenarios. We chose eight classification algorithms representing four different areas. The first three are simple distance based methods; correlation, cosine, and K-Nearest Neighbors (KNN). The next two methods are SVM [5] and Decision Tree [1]. The last set of algorithms is ensemble based, which combine the predictions from multiple models. These final three ensemble methods are Random Forests [2], Extra Trees [7], and Gradient Boosting [4].

Cross Validation

Nested cross-validation [15] is a well established technique for parametrizing and choosing the best predictive model. We have chosen to include the feature selection part of the pipeline within each fold of the cross validation because it is not clear which approach will work best with a given dataset. Additionally we run the entire cross-validation pipeline for different number of features. This enables us to determine what effect dimensionality has on the predictive models accuracy.

Datasets

We elected to use hematopoietic cell types to evaluate the pipeline as extensive characterization using microarrays has been performed on these cell types over the years. Two datasets, training and test, were used in this study.The training dataset was used for cross validation to select the best models. The test dataset is an independent dataset to validate the selected models. We chose the training dataset from an experiment [12] that performed microarray analysis on 38 hematopoietic cell types on the same microarray platform (Affymetrix). The training dataset was limited to 15 cell types to match the 15 cell types found in the dataset used for the independent validation. This resulted in a dataset of 82 samples with approximately 4-7 samples per cell type. The independent dataset [6, 3, 10, 11, 13, 14, 16] consists of 70 samples of 15 cell types with about 3-7 samples per cell type analyzed across multiple platforms, Affymetrix microarray and Illumina bead array. The 15 hematopoietic cell types consisted of primarily terminally differentiated cell types.

3 Experimental Results

3.1 Complete Gene Expression Microarray

The first scenario we benchmark our pipeline on is a typical one, namely the data is gene expression microarrays and the goal is multi-class classification. As described above in the methods section the training data consists of 15 cell types and 4-7 samples per cell types. A 3-fold stratified cross-validation was used in the nested cross validation to ensure each cell type had at least one sample in each fold. The area under the ROC curve (AUC) for each combination of feature selection algorithm and classifier was calculated for each feature size. The sizes considered are 2, 8, 16, 32, 64, 96, 128, 256, and 384, representing feature set sizes often used in biological experiments. The calculated AUCs for individual combinations are shown in Table 1. Figure 1 shows, for each feature size, the AUC achieved by the best combination in cross-validation. The best model of each feature size was evaluated on the test dataset. Figure 1 shows the results. The actual model (feature selection algorithm and classifier combination) selected for each feature size is shown in Table 2.

We observe that RFE with a distance-based classifier (usually correlation) provides the best results. There are a couple of exceptions at 256 and 384 markers, where Random Forest and Anova F-value respectively provide the best model when coupled with the Correlation classifier. At the other end of the feature set sizes (2, 8, and 16) classifiers KNN, Cosine, and Extra Trees provide the best results combined with the RFE method.

As expected, for cross-validation the AUC score increases as feature size increases, but at around 32 features the AUC levels off for the remaining feature sizes. While not as distinct, this same general trend is found in the external validation plot as well. Although, the external validation plot usually have an overall lower AUC score for each feature size, it is lower than expected. This may be due to the inclusion of samples profiles by two different platforms and will be further investigated.

Table 1. AUCs for all the feature selection algorithm and classifier combinations for each feature size.

Feature Selection	Classifier	2	8	16	32	64	96	128	256	384
Extra Trees	Extra Trees	0.721	0.858	0.870	0.870	0.875	0.878	0.892	0.874	0.872
	Random Forrest	0.688	0.768	0.797	0.811	0.832	0.818	0.789	0.809	0.792
	Correlation	0.574	0.861	0.880	0.892	0.894	0.897	0.900	0.895	0.895
	Cosine	0.722	0.866	0.889	0.897	0.898	0.893	0.896	0.893	0.891
	Decision Tree	0.605	0.668	0.642	0.649	0.632	0.628	0.627	0.638	0.690
	Gradient Boosting	0.594	0.660	0.678	0.688	0.741	0.746	0.753	0.772	0.783
	KNN	0.724	0.851	0.864	0.851	0.869	0.859	0.868	0.868	0.874
	SVM	0.622	0.638	0.639	0.632	0.633	0.629	0.631	0.630	0.628
Random Forest	Extra Trees	0.738	0.857	0.885	0.881	0.889	0.875	0.881	0.875	0.881
	Random Forrest	0.713	0.782	0.791	0.827	0.815	0.810	0.783	0.813	0.806
	Correlation	0.621	0.879	0.886	0.891	0.892	0.895	0.900	0.899	0.894
	Cosine	0.760	0.882	0.884	0.889	0.892	0.891	0.900	0.895	0.892
	Decision Tree	0.604	0.612	0.634	0.638	0.648	0.639	0.615	0.633	0.632
	Gradient Boosting	0.597	0.689	0.651	0.690	0.770	0.752	0.756	0.788	0.805
	KNN	0.747	0.833	0.841	0.864	0.860	0.872	0.867	0.871	0.874
	SVM	0.598	0.638	0.626	0.622	0.628	0.626	0.632	0.627	0.623
Anova F-value	Extra Trees	0.748	0.800	0.846	0.853	0.853	0.852	0.866	0.872	0.860
	Random Forrest	0.742	0.752	0.746	0.793	0.778	0.797	0.823	0.785	0.828
	Correlation	0.612	0.812	0.861	0.877	0.871	0.883	0.887	0.890	0.898
	Cosine	0.728	0.814	0.854	0.882	0.870	0.880	0.880	0.890	0.890
	Decision Tree	0.626	0.639	0.613	0.613	0.643	0.641	0.661	0.614	0.647
	Gradient Boosting	0.631	0.655	0.765	0.740	0.734	0.704	0.705	0.781	0.792
	KNN	0.722	0.747	0.801	0.818	0.842	0.831	0.836	0.855	0.858
	SVM	0.687	0.630	0.632	0.647	0.652	0.653	0.655	0.653	0.627
RFE	Extra Trees	0.792	0.881	0.894	0.884	0.893	0.899	0.890	0.883	0.885
	Random Forrest	0.750	0.822	0.816	0.815	0.812	0.821	0.835	0.812	0.805
	Correlation	0.627	0.890	0.892	0.898	0.905	0.904	0.902	0.893	0.891
	Cosine	0.693	0.890	0.892	0.898	0.901	0.901	0.895	0.896	0.894
	Decision Tree	0.600	0.690	0.701	0.659	0.640	0.652	0.660	0.645	0.628
	Gradient Boosting	0.643	0.736	0.735	0.751	0.765	0.788	0.770	0.774	0.770
	KNN	0.808	0.860	0.866	0.869	0.858	0.859	0.858	0.868	0.868
	SVM	0.692	0.666	0.648	0.638	0.631	0.631	0.628	0.623	0.616

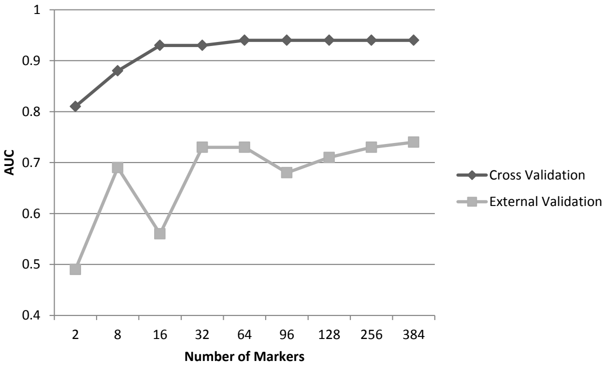


Fig. 1. Results of cross-validation on the training dataset and external validation of the best model on the test dataset using complete microarray data.

Table 2. The best feature selection and classifier combination in terms of AUC for each feature size.

Feature Size	Feature Selection	Classifier
2	RFE	KNN
8	RFE	Cosine
16	RFE	Extra Trees
32	RFE	Correlation
64	RFE	Correlation
96	RFE	Correlation
128	RFE	Correlation
256	Random Forest	Correlation
384	Anova F-value	Correlation

3.2 Simulated Sparse

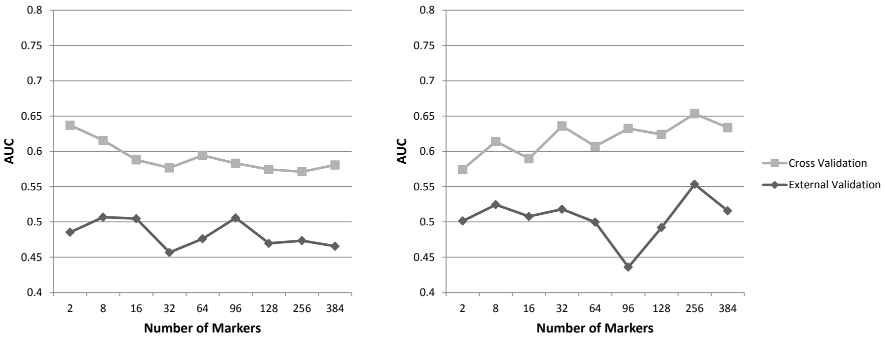
The second test we ran is another scenario encountered by biologists. There are numerous cell or tissue types with gene expression annotated by a small number of genes due to the experimental techniques employed such as QRT-PCR, in-situ hybridization and northern blots, etc. This type of data is considered sparse as compared to cell types annotated with micorarrays which can profile the entire transcriptome. An additional factor to consider is, the genes annotated in one cell or tissue type may not be the same ones across all the cell/tissues types being examined.

To sample an expression matrix of a particular coverage (30% and 50%), we considered the coverage of a marker. That is, the fraction of cell types having known expression statuses for the marker. We assumed that the coverage of a marker follows a Beta distribution. For each marker, the coverage was sampled from the Beta distribution and the samples having known expression statuses are randomly chosen to achieve the desired coverage. The complete data in [12] was converted to simulated sparse data. This was done for coverages of 30% and 50%. The same experimental approach described in the Complete Gene Expression Microarray section was taken. Three simulations for each coverage (30% and 50%) was performed and the average AUC was taken. For cross-validation, the best AUC for each model (feature selection algorithm and classifier combination) is plotted for the feature sizes mentioned above. Next, the best models were evaluated on the independent dataset. The results are shown in Figure 2.

As expected the AUC scores are not as high as the previous experiment, which provided complete expression profile, due to the missing data points. One interesting observation in this experiment is with the 30% coverage. In general, you expect as the number of features selected increases the overall AUC should increase as well. This does not occur with the data set at 30% coverage. Instead, the AUC decreases as the number of features is selected, although it is not a dramatic drop. This is most likely due to the situation mentioned above, where the features are not found consistently across the cell types resulting in the feature selection algorithms to select non-informative features. This effect is not found in the data set with 50% coverage probably because selected features are more likely to be annotated across more samples, therefore being more informative. Further examination is needed, using a broader range of coverages and looking at the actual markers selected to see how they are distributed across the samples, to determine the true pattern.

3.3 Conclusions

In this work, we developed a pipeline for biomarker panel design and unknown sample prediction. As the algorithms are often dataset-dependent, this pipeline allows easy discovery of the best feature selection and classification algorithms. The pipeline revealed that, for the 15-cell type dataset, the classifier plays a much important role than the feature selection algorithm. It further showed that simple classifiers using cosine or correlation outperform sophisticated ones such



(a) 30% (b) 50%

Fig. 2. Results of cross-validation on the training dataset and external validation of the best model on the test dataset using sparse microarray data.

as random forest and SVM. An interesting observation with the sparse data sets, although perhaps not surprising, is at a lower coverage (30%) the AUC dropped as the number of selected features increased, probably due to the incomplete annotation of features across the samples increasing the chance of any feature selected being non-informative. Further experiments need to be carried out to determine the true pattern, before any conclusive statements can be made. The pipeline will help biologists with rational biomarker panel design across a range of different data sets by exploring different combinations of feature selection and classification methods selecting the best combination for a selected set size of features with statistical backing.

Acknowledgments. This project was supported in part by awards IIS-0546457 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 2011-67016-30331 from the USDA National Institute of Food and Agriculture.

References

1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. Wadsworth, Belmont, CA (1984)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Constantinides, M.G., Picard, D., Savage, A.K., Bendelac, A.: A naive-like population of human CD1d-restricted T cells expressing intermediate levels of promyelocytic leukemia zinc finger. J. Immunol. 187(1), 309–315 (Jul 2011)
4. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5), 1189–1232 (2001)
5. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10), 906–914 (Oct 2000), `http://dx.doi.org/10.1093/bioinformatics/16.10.906`

6. Gattinoni, L., Lugli, E., Ji, Y., Pos, Z., Paulos, C.M., Quigley, M.F., Almeida, J.R., Gostick, E., Yu, Z., Carpenito, C., Wang, E., Douek, D.C., Price, D.A., June, C.H., Marincola, F.M., Roederer, M., Restifo, N.P.: A human memory T cell subset with stem cell-like properties. Nat. Med. 17(10), 1290–1297 (Oct 2011)
7. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning 63(1), 3–42 (2006)
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (Mar 2003), http://portal.acm.org/citation.cfm?id=944968
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46(1-3), 389–422 (2002), http://citeseer.ist.psu.edu/guyon02gene.html
10. Jeffrey, K.L., Brummer, T., Rolph, M.S., Liu, S.M., Callejas, N.A., Grumont, R.J., Gillieron, C., Mackay, F., Grey, S., Camps, M., Rommel, C., Gerondakis, S.D., Mackay, C.R.: Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat. Immunol. 7(3), 274–283 (Mar 2006)
11. Lindstedt, M., Lundberg, K., Borrebaeck, C.A.: Gene family clustering identifies functionally associated subsets of human in vivo blood and tonsillar dendritic cells. J. Immunol. 175(8), 4839–4846 (Oct 2005)
12. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., Frampton, G.M., Drake, A.C., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J.W., Liefeld, T., Smutko, J.S., Chen, J., Friedman, N., Young, R.A., Golub, T.R., Regev, A., Ebert, B.L.: Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144(2), 296–309 (Jan 2011), http://dx.doi.org/10.1016/j.cell.2011.01.004
13. Rossi, R.L., Rossetti, G., Wenandy, L., Curti, S., Ripamonti, A., Bonnal, R.J., Birolo, R.S., Moro, M., Crosti, M.C., Gruarin, P., Maglie, S., Marabita, F., Mascheroni, D., Parente, V., Comelli, M., Trabucchi, E., De Francesco, R., Geginat, J., Abrignani, S., Pagani, M.: Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4+ T cells by the microRNA miR-125b. Nat. Immunol. 12(8), 796–803 (Aug 2011)
14. Stirewalt, D.L., Choi, Y.E., Sharpless, N.E., Pogosova-Agadjanyan, E.L., Cronk, M.R., Yukawa, M., Larson, E.B., Wood, B.L., Appelbaum, F.R., Radich, J.P., Heimfeld, S.: Decreased IRF8 expression found in aging hematopoietic progenitor/stem cells. Leukemia 23(2), 391–393 (Feb 2009)
15. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(1), 91+ (Feb 2006), http://dx.doi.org/10.1186/1471-2105-7-91
16. Watkins, N.A., Gusnanto, A., de Bono, B., De, S., Miranda-Saavedra, D., Hardie, D.L., Angenent, W.G., Attwood, A.P., Ellis, P.D., Erber, W., Foad, N.S., Garner, S.F., Isacke, C.M., Jolley, J., Koch, K., Macaulay, I.C., Morley, S.L., Rendon, A., Rice, K.M., Taylor, N., Thijssen-Timmer, D.C., Tijssen, M.R., van der Schoot, C.E., Wernisch, L., Winzer, T., Dudbridge, F., Buckley, C.D., Langford, C.F., Teichmann, S., Gottgens, B., Ouwehand, W.H.: A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood 113(19), 1–9 (May 2009)