Disease Markers 20 (2004) 45–51
IOS Press

Statistical considerations in combining biomarkers for disease classification

Ziding Feng and Yutaka Yasui
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

1. Introduction

The recent advances in genomics (gene expression arrays and SNPs), proteomics (protein expression using mass spectrometry or antibody arrays) opened the door for biomedical researchers to combine multiple biomarkers measured using noninvasive procedures (e.g., samples from serum, urine, stool) for disease classifications (detection, diagnosis, and prognosis). This is important because for most diseases a single biomarker is not adequate in terms of classification performance, e.g., sensitivity (true positive fraction) and specificity (true negative fraction), for intended clinical use. By combining the biomarkers, we hope to achieve more accurate disease classifications and increased clinical benefits.

However, this great promise comes with great challenges. The number of ways of combining multiple biomarkers increases exponentially with the number of biomarkers. As a consequence, searching for the optimal combination of biomarkers is not only computationally demanding but also run the high risk of false positive findings. False positive findings waste precious resources used for further investigations, or potentially cause harm to patients if they are put for clinical use without firm validations.

The goal of this paper is to provide biomarker researchers and analysts who evaluate biomarkers for potential clinical use some guidance on how to combine biomarkers. The organization of the paper is as follows: Section 2 discusses the challenges due to the high dimensionality of modern genomic and proteomic data, "the curse of dimensionality", and the importance of a test set of samples in such investigations. When a test set is not practical, we propose to use Cross-Validation for model selection and a bootstrap procedure for estimating the prediction error of the selected model. The principles of the optimal approach to combine biomarkers when the joint distribution of biomarkers for each of the disease and control groups is known are described in Section 3. This optimality is usually unattainable as we often have little knowledge about these joint distributions. However, the principles provide justifications of choosing one procedure over another in the settings when we do not know the joint distributions of biomarkers. For such settings, three procedures are recommended and contrasted in Section 4: nonparametric method, logistic regression, and boosting. Section 5 discusses a more challenging situation: the identities of biomarkers are unknown, or known with some uncertainty. This is typical for the protein profile data from mass spectrometry, due to both the measurement errors in protein mass (mass/charge) and the nature of such exploratory protein-expression studies mining for disease classification profiles without first identifying proteins/peptides. Concluding remarks are given in Section 6.

2. Curse of dimensionality

The term of "Curse of dimensionality" first came from Bellman [1]. It basically says that the required number of samples to uncover truth grows exponentially fast with the dimension of data (e.g., the number of biomarkers). For example, one would feel uncomfortable to study a relationship between Y, disease status (1 for diseased, 0 for non-diseased), and X, the

ISSN 0278-0240/04/$17.00

value of a single biomarker, if only three observations are given. One would probably feel comfortable if information about Y and the values of 10 biomarkers are given on 59,000 observations. However, the data sparseness for these two settings is about the same because the data density is proportional to $\mathrm{N}^{1/p}$, where N is the number of observations and p is the number of biomarkers. With 10 biomarkers, numerous complex combinations of them exist. This leads to three important messages. First, whenever possible, simple classification rules using small numbers of biomarkers guided by subject knowledge are always preferable to more complex ones. Second, findings from such sparse data should be treated as exploratory and require further confirmations. Third, the models (classification rules) built from such high dimensional sparse data have a great risk of *overfit*, a phenomenon that a model has very low prediction error in the data used for model building but has poor performance (generalizability) on new independent data.

To avoid an overfit, we need to assess the model very carefully. The best approach is to set aside samples that are not used for model construction as a test set and evaluate the classification performance in this *test set* after the model (disease classifier) is finalized from the *training set* of samples. This test set should be used only once after the model is fixed. Repeatedly update the model based on the test set performance will render the value of the test set to a training set, and could not provide an unbiased estimate of the prediction error on new data. Also, fixing more than one final models from the training set and taking the best test-set performance among them is not a proper use of the test set: this scheme uses the test set as part of the training set.

Note that assessments of prediction errors serve two purposes: for selecting a model with the minimum estimated future prediction error, and also for estimating future prediction error of the model. The test set is useful for the latter purpose.

The former purpose could be achieved by a procedure called Cross-Validation (CV) [2]. For example, a 10-fold CV goes as follows. Split total samples into 10 equal parts, leave one part out and use the remaining 9 parts to fit a model with pre-specified complexity. Use the constructed model to predict the one part that was left out and obtain the prediction error in it. Repeat the process for each of the 10 parts in turn, a total of 10 times, and average the observed prediction errors. This average is called Cross-Validation Error. We can apply this CV process to a series of models with increasing complexities separately and choose the model that has smallest Cross-Validation Error: note that too complex models would overfit and will have high Cross-Validation Errors.

Even when we cannot have enough samples to split into training and test sets, we need to assess the future prediction error for the final chosen model. This can be achieved by using bootstrap [3]. We randomly draw, with replacement, N observations from the original N observations we have in the observed data to form a bootstrap set of N samples. For each bootstrap set of N samples, we repeat the above Cross-Validation model selection process to get a final model from that bootstrap set and use this model to predict observations in the original samples that were not selected in this bootstrap sample, and compute the prediction error among them. Repeat this process B times (at least 100) and average the prediction errors. We call this quantity Validation Prediction Error. This is our estimate of the true prediction error in the absence of a test set. The best Cross-Validation Error from a given dataset is likely to be somewhat too optimistic (smaller) than the Validation Prediction Error of that dataset, or true future prediction error, because the model selection used the same data as those used for calculating the Cross-Validation Error of the final model.

Using CV for model selection and bootstrap for model selection or validation is a widely accepted practice in statistics. However, we are unaware of using CV for model selection and bootstrap for model assessment in the same data analysis, especially tailored to combining multiple biomarkers for disease classification. We believe that using these two procedures in this way is useful for high-dimensional biomarker data analysis. This is because we need to select a model that has a good chance of truly having a low future prediction error (this is accomplished via CV), and at the same time need to estimate the future prediction error of that model unbiasedly (this is accomplished via bootstrap).

3. Combining biomarkers when the joint distribution of biomarkers is known for each of diseased and control groups

An application of Neyman-Pearson Lemma to diagnostic tests has led to remarkable insights about combining multiple tests [4–8] that we elaborate here. It says that, with a biomarker covariate vector $X = (X_1, \ldots, X_p)$ and a disease status indicator Y (1 for diseased, 0 for non-diseased), the optimal classification rule is to predict $Y = 1$ for a sample if $LR(X) > c$,

where $LR(X) = P(X|Y = 1)/P(X|Y = 0)$. $(X|Y = 1)$ means "X given $Y = 1$". LR(X) is called a likelihood ratio for $Y = 1$ vs. $Y = 0$ given X. The threshold c is determined by the requirement of sensitivity or specificity. The optimality here means that, among all classification rules, the optimal rule has the highest sensitivity for a fixed specificity, the highest specificity for a fixed sensitivity, minimize the overall misclassification probability, and minimize the expected cost regardless of the cost balance between false positive and false negative results.

An equivalent, but more practical, way of applying this optimal rule is to employ the following Risk Score rule: $RS(X) > c*$, where $RS(X) = P(Y = 1|X)$. $RS(X)$ is a monotone increasing function of LR(X) so this amounts to choose a $c*$ in $RS(X)$ for a corresponding c in $LR(X)$.

Both $LR(X)$ and $RS(X)$ require a correct probability model for the relationship between multivariate biomarkers X and the disease status Y in order to obtain the optimality. In reality, we even don't know which subset of biomarkers is related to disease, even less on how they are related. Therefore, the optimality is rarely achievable. However, understanding this principle will guide us in choosing procedures for combining multiple biomarkers. The procedures that approximate the principle are likely to have better performance than the procedures that violate this principle. Some procedures could be motivated by this principle, as we will see below.

4. Combining biomarkers when the joint distribution of biomarkers is not known for each of diseased and control groups

4.1. Nonparametric method (*Baker*)

Baker [6] proposed a nonparametric procedure to combine multiple biomarkers, motivated by the likelihood ratio principle. The idea is as follows. Take two continuous biomarkers (X_1, X_2) as an example. Divide each biomarker into n intervals and form a table with nXn cells. For cell (i, j) of the table, estimate $LR_{ij} = P(X_1 = i, X_2 = j|Y = 1)/P(X_1 = i, X_2 = j|Y = 0)$ by replacing denominator and numerator with observed proportions. By the likelihood ratio principle, the optimal rule is to choose cells that have the largest values of LR_{ij} to form a disease positivity region (i.e., $LR_{ij} > c$). The size of the region can be determined by a pre-specified sensitivity or specificity.

Because the underlying mechanism relating biomarkers to the disease ought to be orderly, while the observed LR_{ijs} may not exactly follow any pattern due to randomness in the sample, it is natural to impose some ordering over the table. Assuming that higher values of each biomarker *a priori* imply a greater probability of disease, Baker proposed three ordering methods: Unordered; Jagged Ordering; and Rectangular Ordering. Of those, the Jagged Ordering may be the most natural ordering. It requires that if cell (i, j) is in the positivity region, the higher order cells ($i + 1$, j) and (i, $j + 1$) must also be in the region. The Unordered approach uses all random noise in the data and therefore tends to overfit. The Rectangular Ordering requires the shape of the positivity region to be rectangular at lower right corner of the table. This ordering may be too simple for many applications.

The attractive feature of Baker's procedure lies in its use of the likelihood ratio principle without requiring the knowledge of the probability model $P(X|Y)$. It estimates it nonparametrically from the observed data by using data's empirical distribution, the proportions of diseased and non-diseased in the cells. It can accommodate certain types of interactions that are most likely to occur, e.g., having higher values of both biomarkers increases the probability of disease more than additive effects of the two biomarkers. We even don't need to model and estimate these interactions because the procedure automatically includes them nonparametrically.

The main drawback of this approach lies in the difficulty of handling larger number of biomarkers. When the dimension of X is high, the data are too sparse (curse of dimensionality) in a high dimensional table to allow this type of nonparametric calculation effectively. Therefore, this approach is suitable for combining a small number of biomarkers and relatively large number of observations such that likelihood ratio estimate in each cell is stable. The choice of cutoff points to form the n intervals for the table could also be improved. These issues deserve more research.

4.2. Logistic regression

Logistic regression assumes the following model: $\text{logit}(E[Y]) = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots + b_MX_M$, $Y_s \sim$ independent Bernoulli distributions with means E[Y]s, where logit(a) $= \ln(a/1 - a)$. After the model is finalized, we can predict the probability of disease using the formula: probability of disease $= 1/[1+\exp\{-(b_0+b_1X_1+b_2X_2+\ldots\ldots+b_MX_M)\}]$.

The cutoff point depends on the required sensitivity or specificity.

The major advantages of logistic regression are as follows. First, logit(E[Y]) is a monotone function of RS(Y|X) so it is optimal as long as the logistic model we use with X is monotone function of the unknown true model. This makes the logistic regression robust for model mis-specification: i.e., one could focus on correctly specifying the mean structure, $b_0+b_1X_1+b_2X_2+\ldots\ldots+b_MX_M$. Second, the model parameters b's are estimable from either retrospective (case-control) or prospective (cohort) studies except b_0 that depends on the disease prevalence. Therefore, if a good estimate of the disease prevalence in a target population is available, one can easily modify a logistic regression model, built upon a case-control study, for use in the target population. Adding interaction terms are easy in logistic regression models, although they need to be specified, in contrast to the Jagged Ordering in the nonparametric method.

The size of the final logistic regression model, expressed by either the total number, or the largest p-value, of the biomarkers in the model, should be determined by Cross-Validation described in Section 2.

4.3. Boosting

Boosting was developed in computer science during 1990s [9–11], for combining multiple "weak" classifiers into a powerful committee. For a comprehensive discussion, see Hastie, Tibshirani, and Friedman [12]. We will describe two popular boosting algorithms here, Discrete AdaBoost and Real AdaBoost. The applications of boosting for combining multiple biomarkers have been discussed in Qu [13] for Discrete AdaBoost, and Yasui [14,15] for Real AdaBoost.

4.4. Discrete AdaBoost

Discrete AdaBoost uses a classifier $f_m(x)$ that takes a value of 1 (to indicate disease) or –1 (to indicate non-disease) for a given value x of a biomarker indexed by m. Suppose we use a simple "stump" classifier that classifies a sample as disease if the biomarker value is above (or below) a certain level, or detectable if the biomarker is a binary measure. We choose the first biomarker that gives the smallest misclassification error. This forms our first classifier $f_{m=1}$. In this first classifier, we use an equal weight (weight $= 1/n$) for all n observations.

This classifier is usually a weak classifier. Now we update the weights by assigning larger weights for the observations that were misclassified by this classifier. For observation Y_i ($Y_i = 1$ for disease, -1 for non-disease), the new weight is w_i = previous weight $\times \exp\{c_m \times I(y_i f_m(x_i) = -1)\}$, where $c_m = \log(1 - \mathrm{err}_m)/\mathrm{err}_m)$, err_m is the proportion of all n samples that are misclassified by the current classifier, and I(.) is an indicator function which takes a value of 1 if the statement in parenthesis is correct, 0 otherwise. Note that if the prediction is correct, then $y_i f_m(x_i) = 1$ and $\exp\{c_m \times I(y_i f_m(x_i) = -1)\} = 1$. The new weight is then equal to the old weight. If the prediction is incorrect, then $\exp\{c_m \times I(y_i f_m(x_i) = -1)\} = \exp\{c_m\}$ is larger than one as long as c_m is positive: this always holds with a high probability if the weak classifier is better than flipping a coin. This is the first important feature of boosting: it assigns larger weights to those difficult to classify. Logistic regression uses the same weight for each observation.

We now apply our favorite classifier algorithm (e.g., a stump) again, the same way as described above, but to the weighted data. We repeat this process M times and the final classifier is

$$f(x) = \mathrm{sign}\left[\sum_{m=1}^{M} c_m f_m(x)\right]$$

a weighted sum of all M classifiers: the classifier is $f(x) = 1$ if the sign of the sum is positive and –1 otherwise. Therefore, boosting is a committee voting procedure. The same biomarker could be repeatedly selected. The iteration number M can be determined by the Cross-Validation.

4.5. Real AdaBoost

Real AdaBoost has the following differences from the Discrete AdaBoost:

1. Instead of using a discrete classifier $f_m(x)$, Real AdaBoost uses a continuous classifier that produces a class-probability estimate p_m, the predicted probability of disease based on mth biomarker. A natural choice for constructing class-probability estimates is via logistic regression with a single covariate.
2. Calculates a quantitative classifier $f_m(x) = 0.5 \log[p_m(x)/(1 - p_m(x)]$
3. Updates the weight by: new weight = previous weight $\times \exp\{-(y_i f_m(x_i)\}$, $i = 1, 2, \ldots n$, and renormalize so that the sum of weights over n samples equals to 1. The initial weight is $1/n$.

4. Repeat this process M times and the final classifier is:

$$f(x) = \text{sign}\left[\sum_{m=1}^{M} f_m(x)\right].$$

Note that in Real AdaBoost, $f_m(x)$ will be positive if $p_m(x) > 1/2$, and will increase as $p_m(x)$ increases. Therefore, in each iteration, it assigns weights to each observation not only according to whether it is correctly classified, but also the confidence of the correct specification or the extent of misclassification. In each iteration, Discrete AdaBoost assigns one weight for all correctly classified and one weight for all incorrectly classified. On the other hand, Real AdaBoost uses "confidence-rated" weights that differ across n samples and we expect it will "learn faster" and have better predicting power.

Both Discrete and Real AdaBoost will overfit If the number of iterations, M, is too large. However, the slowness of the boosting learning process makes it very resistant to overfitting. This is the main advantage of using boosting in combining multiple biomarkers. Other considerations are the easy interpretation of committee voting method. Boosting procedures discussed above are all additive models and do not handle interactions among biomarkers. It is conceptually straightforward to boosting small trees instead of stumps, or boost logistic regression models up to low order interactions, to allow interactions. However, the interpretations quickly become very difficult unless M is very small.

5. Combining multiple "biomarkers" when the marker identities are unknown or measured with error

For serum protein/peptide spectra from a time-of-flight mass spectrometry, for example, the peptide or protein identities are unknown without performing further intensive protein/peptide identification steps in laboratories. A number of studies attempted to construct classifiers for diseases using protein fingerprints [16–18]. Not only we do not know the identities of proteins/peptides corresponding to the peaks observed in mass spectra, it is also well known that the same protein peaks from the same sample may appear at different mass points in repeated runs due to technological imperfection in measuring mass accuracy, as well as intensity of each peak.

Peak identification and alignment are analytically challenging but not discussed here: see, for example [19]. Instead we discuss the statistical considerations after the peaks are identified and aligned. Analytically we can treat these peaks as "biomarkers" and the strategies discussed in Section 4 all apply. However, there are further statistical challenges we need to consider.

First, after initial findings of "fingerprints", their reproducibility must be established firmly before further confirmatory studies are conducted. If biomarkers are known such as the cases in microarray gene expression and antibody array experiments, after an initial positive finding was made for a biomarker panel for disease classification, a natural step is to confirm the finding in a larger sample, a sample from more heterogeneous population, and/or confirm the panel against specified sensitivity and specificity that are of clinical significance. Reproducibility issues are still important but are more for the consideration of assay scale up for routine clinical use. However, with only "fingerprints" at hand, the first step is to show that the fingerprints identified are robust: reproducible and accurate. Human fingerprints are robust because they are easily reproducible and accurate enough for identification purpose. Biomarkers we identify must possess similar properties for identifying the disease of interest. They must have mass and intensity measurements that are reproducible and accurate enough such that they will lead to the same classification when the diagnostic model is applied on their spectra obtained from repeated runs. Existence of the fingerprints observed in repeated assays of a single sample, for example, is not adequate to claim reproducibility and accuracy. The ability to generate another panel/model with good performance using a new set of data is not adequate, either: it has to be the same panel/model with good performance on the new data. Otherwise, the model has no clinical use.

Second, the reproducibility and accuracy have to be examined on samples from multiple studies/populations than the study/population where the initial findings were made. The objective of this step, when biomarkers are known, is to examine whether the same biomarker panel works for a more general population and fine-tune the model if necessary. With "fingerprints" data, however, the main objective is more basic: mainly to find out whether the "fingerprints" of the initial findings represent artifacts or they are likely to be real biological signals. For example, if there were differences between disease and non-disease groups in the methods for sample collection, preparation, and

storage in a study, the "fingerprints" we identify from this study would reflect these methodological differences between disease and non-disease, an artifact, not a real biological signal of the disease. The distinction between artifacts and real biological signals cannot be made by any analytical approach without data from different studies/populations.

Basically, the hurdle is higher for a biomarker panel to pass to the next validation phase if the identities of these biomarkers are unknown. There are more to prove and to rule out. That is the price to pay when the identities are unknown.

One fruitful use of the "fingerprints" approach is to use fingerprints to narrow down the region for further studying of protein/peptide identification/function. This is a very challenging problem because the high dimensionality of proteomic data often allows a large number of models to have "good" performance in a given dataset so one needs to narrow down from them to make the phase of protein/peptide identification/function study practical. Creative experimental design and analysis are needed. One proposal is to conduct a sequence of smaller experiments to narrow down the candidates sequentially. A similar approach at the analysis stage is to split the dataset into several parts and artificially create a sequence of smaller experiments/datasets. The rationale for the sequential experiment is related to "the curse of dimensionality". The sample size could not keep up with the demands due to the increase in data dimension. Having a single experiment with the sample size large enough for the high-dimension problem is not practical. On the other hand, false positive findings from the preceding experiment should fail when applied to the new samples in the next experiment; thereby we can eliminate the previous false positive findings quickly in the sequence of smaller experiments. This approach with sequential experiments would avoid where little hope is shown in multiple experiments and focus on where promising features are observed across multiple experiments. It merits further research.

6. Concluding remarks

One message we want to emphasize in conclusion is the unrealistic expectation biologists and clinicians placed on quantitative scientists to "find a magic way of combining multiple biomarkers". Although a successful exploration of high-dimensional data could be achieved soundly in a sequence of sufficient-sized experiments, using a small set of biomarker candidates to start with, based on sound biology, is an effective way of alleviating the curse of dimensionality. We need to understand the importance of biomarkers' reproducibility/accuracy and validation: these are particularly crucial when a combination of biomarkers forms a panel and/or when identities of biomarkers are unknown. We need to be careful not jumping to conclusions or claims. Biologists, clinicians, and statisticians need to work closely together and use the statistical tools as aids towards understanding biology and detecting diseases, not to bypass or replace the methodical efforts towards these goals.

References

[1] R.E. Bellman, *Adaptive Control Processes*, Princeton University Press. 1961.
[2] M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. Royal Statistical Society* **36**, 111–147.
[3] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, New York: Chapman & Hall, 1993.
[4] D.M. Green and J.A. Swets, *Signal detection theory and psychophysics*, New York: Wiley, 1966.
[5] J.P. Egan, *Signal detection theory and ROC analysis*, New York: Academic Press, 1975.
[6] S.G. Baker, Identifying combinations of cancer markers for further study as triggers of early intervention, *Biometrics* **56** (2000), 1082–1087.
[7] M. McIntosh and M.S. Pepe, Combining several screening tests: Optimality of the risk score, *Biometrics* **58** (2002), 657–664.
[8] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, New York: Oxford University Press, 2003.
[9] R. Schapire, The strength of weak learnability, *Machine Learning* **5**(2) (1990), 197–227.
[10] Y. Freund, Boosting a weak learning algorithm by majority, *Information and Computation* **121** (1995), 256–285.
[11] Y. Freund and R. Shapire, A decision-theoretic generalization of online learning and an application to boosting, *Journal of Computer and System Sciences* **55** (1997), 119–139.
[12] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, New York: Springer, 2001.
[13] Y. Qu, B. Adam, Y. Yasui, M. Ward, L. Cazares, P. Schellhammer, Z. Feng, J. Semmes and G. Wright, Boosted Decision Tree Analysis of SELDI Mass Spectral Serum Profiles Discriminates Prostate Cancer from Non-Cancer Patients, *Clinical Chemistry* **48** (2002), 1835–1843.
[14] Y. Yasui, M. Pepe, M. Thompson, B. Adam, G. Wright, Y. Qu, J. Potter, M. Winget, M. Thornquist and Z. Feng, A data-analytic strategy for protein-biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, *Biostatistics* **4** (2003), 449–463.
[15] Y. Yasui, M.S. Pepe, M. Thompson, B. Adam, G. Wright, Y. Qu, J. Potter, M. Winget, M. Thornquist and Z. Feng, A data-analytic strategy for protein-biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. (January 31, 2002), UW Biostatistics Working Paper Series. Working Paper 177.

[16] E.F. Petricoin, et al., Use of proteomic patterns in serum to identify ovarian cancer, *The Lanet* **359** (2002), 572–577.
[17] B. Adams, et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from Benign Prostate Hyperplasia and healthy men, *Cancer Research* **62** (2002), 3609–3614.
[18] A.J. Rai, et al., Proteomic approaches to tumor marker discovery, *Arch Pathol Lab Med* **126** (2002), 1518–1526.
[19] Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist and Z. Feng, An automated peak-identification / calibration procedure for high-dimensional protein measures from mass spectrometers, *Journal of Biomedicine and Biotechnology* **2003** (2003), 242–248.