Methods Supplement for "How large a training set is needed to develop a classifier for microarray data" by Kevin K. Dobbin, Yingdong Zhao and Richard Simon

Calculations for Table I

For these calculations, the following normal homoscedastic model is assumed

$$x \sim \begin{cases} Normal\left(\mu, \sigma^2 I\right) & x \in C_1 \\ Normal\left(-\mu, \sigma^2 I\right) & x \in C_2 \end{cases}$$

where x is a vector of gene expression measurements, and I is the identity matrix. The length of vector x is the number of features represented on the microarray. For these calculations it is assumed that $\mu = (\mu_1, \mu_2, ..., \mu_p)'$ with $\mu_i = \Delta$ for $i = 1, ..., m$ and $\mu_i = 0$ for $i = m+1, ..., p$. Under these assumptions, the standardized fold change for the first m genes is $\frac{2\Delta}{\sigma}$. It can be shown that the lowest possible error rate of a linear predictor is $1 - \Phi\left(\frac{\Delta}{\sigma}\sqrt{m}\right)$, which is the equation used to calculate the right column in the table.

Calculations for Tables II & III

The input parameters are the tolerance γ (which is 0.10 in Table II and 0.05 in Table III), number of genes on the arrays (fixed at 22,000 for these tables), the standardized fold change $\frac{2\Delta}{\sigma}$ given by the rows of the tables, and the prevalence in the under-represented class p_u given by the columns of the tables.

Implementation of the algorithm required the development of a new method for estimating the average probability of correct classification given a random sample from a population with n_{0s} from one class and n_{0B} from the other class (with $n_{0s} < n_{0B}$), and also given a fixed effect size $\frac{2\Delta}{\sigma}$ and number of informative genes m and that one will optimize over the gene selection significance level α as described in Dobbin and Simon (2007). The total training set size is $n_0 = n_{0s} + n_{0B}$. This new function is denoted $PCC_m(n_{0s}, n_{0B})$, where the dependence on $\frac{2\Delta}{\sigma}$ is omitted to simplify the notation. The value of $PCC_m(n_{0s}, n_{0B})$ is obtained by searching over different values of α to find one

that maximizes the expected accuracy. We can represent this notationally as $PCC_m(n_{0s}, n_{0B}) = \max_{0<\alpha<1} PCC_{\alpha,m}(n_{0s}, n_{0B})$. In order to calculate $PCC_{\alpha,m}(n_{0s}, n_{0B})$, one needs an estimate of the power for detecting the fold change for an informative gene. We use the method of Dupont and Plummer (1990) with the ratio of control to experimental subjects $\frac{n_{0s}}{n_{0B}}$. To ensure internal consistency in the computations, the ratio $p_{u,n_{0s},n_{0B}} = \frac{n_{0s}}{n_{0s} + n_{0B}}$ is used in the internal search loop as the estimate of the population proportion from the under-represented class.

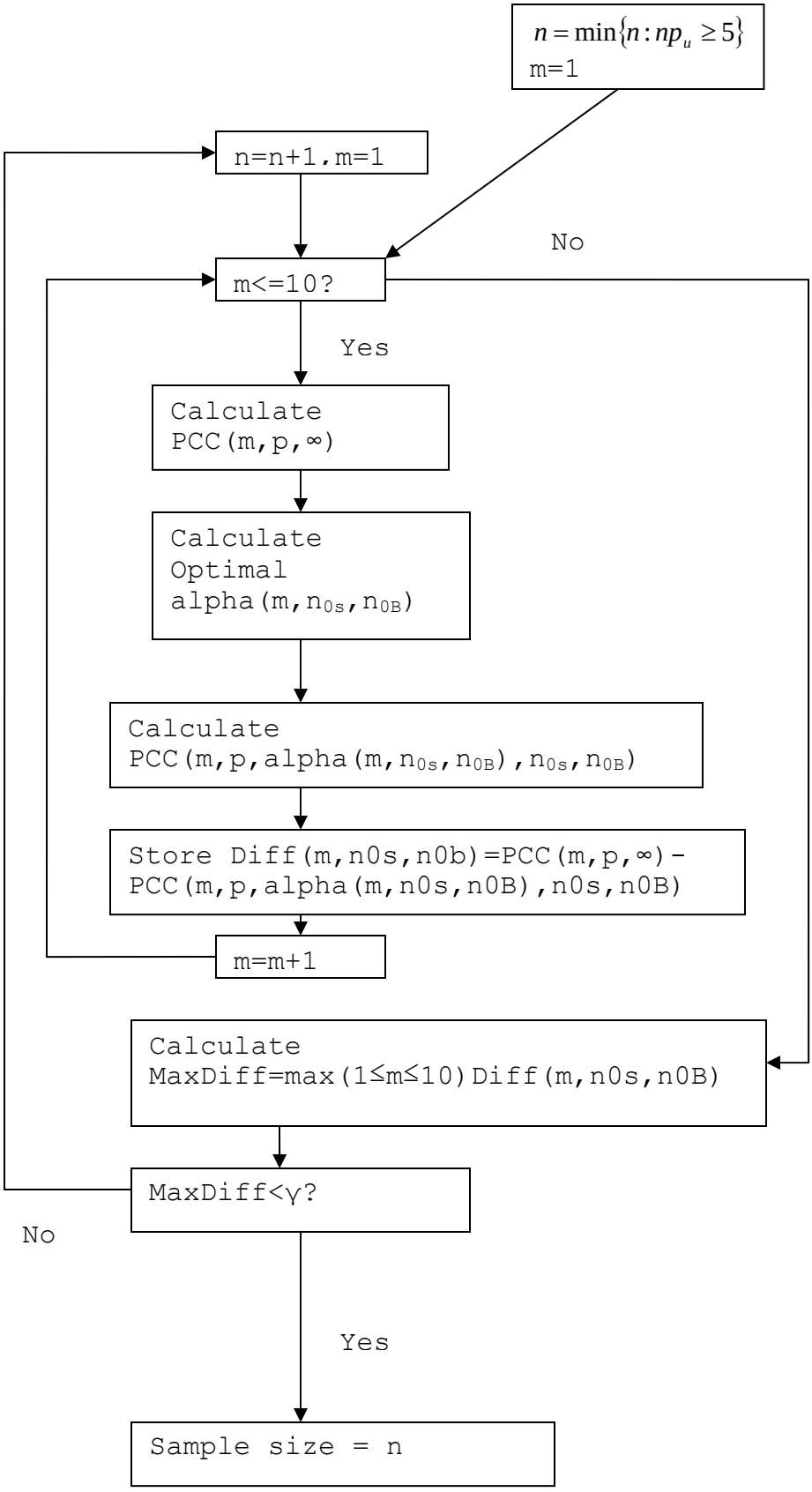
Algorithm

1. Set $n_0 = \min\{n : np_u \geq 5\}$.
2. Calculate $n_{0s} = round(n_0 \cdot p_u)$ and $n_{0B} = n_0 - n_{0s}$, the expected sample sizes in from each class.
3. For $m = 1,2,...,10$, informative genes, calculate the optimal accuracy using the formula $p_{u,n_{0s},n_{0B}} \Phi\left\{\frac{m(\Delta^2/\sigma^2) - 0.5\log\left[(1/p_{u,n_{0s},n_{0B}}) - 1\right]}{\sqrt{m\Delta^2/\sigma^2}}\right\} + (1 - p_{u,n_{0s},n_{0B}})\Phi\left\{\frac{m(\Delta^2/\sigma^2) + 0.5\log\left[(1/p_{u,n_{0s},n_{0B}}) - 1\right]}{\sqrt{m\Delta^2/\sigma^2}}\right\}$. Call these $PCC_{m,p_{u,n0s,n0B}}(\infty)$.
4. Calculate the optimal cutoff values for gene selection associated with each m. Call these $\alpha_1,...,\alpha_{10}$. Here α_m is the α that minimizes the function $PCC_{m,p_{u,n0s,n0B}}(\infty) - PCC_m(n_{0s}, n_{0B})$. Each α_m is obtained via a golden section search on the unit interval. In particular, the objective function whose maximum is obtained at each step of the search is $p_{u,n0s,n0B}\Phi\left\{\frac{1}{\sigma\sqrt{\lambda}}\frac{\Delta \cdot m \cdot DPPow(\Delta,\sigma,n_{0s},n_{0B},\alpha) - 0.5\log\left[(1/p_{u,n0s,n0B}) - 1\right]}{\sqrt{mDPPow(\Delta,\sigma,n_{0s},n_{0B},\alpha) + (p-m)\alpha}}\right\} + (1 - p_{u,n0s,n0B})\Phi\left\{\frac{1}{\sigma\sqrt{\lambda}}\frac{\Delta \cdot m \cdot DPPow(\Delta,\sigma,n_{0s},n_{0B},\alpha) + 0.5\log\left[(1/p_{u,n0s,n0B}) - 1\right]}{\sqrt{mDPPow(\Delta,\sigma,n_{0s},n_{0B},\alpha) + (p-m)\alpha}}\right\}$ where DPPow is the power function for a t-test estimated using the method of Dupont and Plummer (1990).
5. Estimate the worst-case-scenario difference between the optimal accuracy and the expected accuracy difference based on this sample size by $WCAccDiff = Max_{1\leq m\leq 10}\left\{PCC_{m,p_{u,n0s,n0B}}(\infty) - PCC_{m,p_{u,n0s,n0B}}(\alpha_m, n_{0s}, n_{0B})\right\}$. Here we are exploiting the fact that the worst-case-scenario m is usually 1,

and is extremely unlikely to be greater than 10, in order to streamline calculation.

6. If $WCAccDiff < \gamma$, then exit, otherwise let $n_0 = n_0 + 1$ and return to step 2.

A schematic of the algorithm is presented below.


$n = \min\{n : np_u \geq 5\}$
m=1
n=n+1.m=1
No
m<=10?
Yes
Calculate
PCC(m,p,∞)
Calculate
Optimal
alpha(m,n0s,n0B)
Calculate
PCC(m,p,alpha(m,n0s,n0B),n0s,n0B)
Store Diff(m,n0s,n0b)=PCC(m,p,∞)-
PCC(m,p,alpha(m,n0s,n0B),n0s,n0B)
m=m+1
Calculate
MaxDiff=max(1≤m≤10)Diff(m,n0s,n0B)
MaxDiff<γ?
No
Yes
Sample size = n

Calculations for Table IV

The algorithm described in the calculations for Tables II and III is applied with the adjustments described in the caption to Table IV.

Calculations for Table V

Entries in this table have prevalence set at 50% and the number of genes on each array set at 22,000. Columns represent the overall sample size used for the training set. Rows represent the largest absolute value of the pooled-variance t-test statistic observed on the training set (denoted below *maxT*), i.e., the t-test statistic associated with the most significant (smallest p-value) gene. Note that here we assume there is no missing data for any of the genes, so that each t-test is based on samples of size $n/2$ from each class.

Using these parameter inputs, the maximum standardized fold change size in the training set is estimated by the formula $\hat{F} = 0.80 \bullet \max T \bullet \sqrt{\frac{4}{n}}$, with motivation described in the caption to Table IV. Hence $\hat{F}$ is the estimate of $\frac{2\Delta}{\sigma}$. Using this estimate, the tolerances given in the table are estimated by $\max_{1 \le m \le 10} \left\{ PCC_{m,0.5}(\infty) - PCC_{m,0.5}(n) \right\}$.

Calculations for Figure 1

The web-based interface program (implementing the methodology described in the calculations for Tables II and III) was used to calculate individual points on the lines and cubic spline interpolation was used to connect points associated with the same sample size.

REFERENCES

Dobbin KK, and Simon RM. (2007) Sample size planning for developing classifiers using high-dimensional DNA microarray data. Biostatistics, 8: 101-17.

Dupont WD, and Plummer WD Jr. (1990) Power and sample size calculations. A review and computer program. Controlled clinical trials, 11: 116-28.