J.Natn.Sci.Foundation Sri Lanka 2016 44 (1): 3-9
DOI: http://dx.doi.org/10.4038/jnsfsr.v44i1.7976

SHORT REVIEW

High-dimensional data

Dhammika Amaratunga[1]* and Javier Cabrera[2]
[1]*Johnson & Johnson Pharmaceutical Research and Development, Raritan, New Jersey, USA.*
[2]*Department of Statistics and Biostatistics, School of Arts and Sciences, Rutgers University, New Jersey, USA.*

Revised: 28 September 2015; Accepted: 28 October 2015

Abstract: This paper provides a brief introduction to high-dimensional data, a form of 'Big Data', and gives an overview of several data analysis concepts and techniques that could be used to explore and analyse such data. An example that involves genomics data from several Sri Lankan and United Kingdom oral cancer patients is used to illustrate the methods.

Keywords: Big Data, biplot, conditional t, false discovery rate, lasso, visualisation.

INTRODUCTION

'Big Data' is a major phenomenon at the present time. It is expected that if properly managed, developed and controlled, Big Data will drive progress in a number of different spheres including medicine, science (all sciences), government, society and business. The phenomenon is powered by an exponential increase in the ability to collect, store, curate, share, and process massive amounts of data and also by technical developments in the sciences. Consequently, Big Data has become a focal point of interest in Statistics, Computer Science and Applied Mathematics.

High-dimensional data could arguably be regarded as a particular form of Big Data. High-dimensional data are data that can be organised as a pxn matrix X, where n is the number of samples in the study and p is the number of features or variables (or dimensions) being studied, and n and p are such that p is at least an order of magnitude larger than n (i.e., $n << p$). It is this latter attribute that distinguishes this type of data, sometimes referred to as 'small n, large p' data, from standard statistical multivariate data, which are similar except that there $n > p$. In this sense, it is also different from what is normally thought of as Big Data as there again $n > p$, but with n extremely large.

As with Big Data in general, technological advances have led to high-dimensional data becoming prevalent in many areas of research, particularly genomics and imaging. As an example, the expression levels (i.e., the activity levels) of $p = 8793$ genes in biopsied oral mucosal tissues were recorded for $n = 48$ oral squamous cell carcinoma (OSCC) patients, 27 of Sri Lankan (SL) origin and 21 of the United Kingdom (UK) origin. The objective of this was to study the differences in expression patterns across these two populations, since it is known that such cancers from Sri Lanka, which are associated with betel quid chewing, are phenotypically distinct from those from the United Kingdom, which are predominantly caused by smoking and alcohol consumption. But the genomic basis of these differences is largely unknown (Saeed *et al*., 2015). This data will be used for illustrative purposes in the rest of the paper. Table 1 shows a small subset of the data.

The *de facto* dimensionality of the data is, of course, p. Conventional data analysis methodologies are either not directly applicable or are unlikely to be effective when dealing with this sort of data as the sample size is considerably less than p. Methods developed for regular Big Data also could have a tendency to overfit since $n < p$; i.e., they are likely to find spurious patterns in the data. In fact, at first sight, the high dimensionality would seem to present an insurmountable problem, not just for conventional methods, but for any method. However, the essential belief when analysing a high-dimensional dataset is that it contains patterns of value, which reside in much lower (say k) dimensional subspaces, where not only $k < p$ but also $k < n$; in fact, ideally $k = 1$ or 2. In fact, neither the low-dimensional representations, nor

* Corresponding author (damaratung@yahoo.com)

Table 1: Small subset of the OSCC data, showing the data for 12 genes (G1 to G12) for three Sri Lankan samples (S1 to S3) and three United Kingdom samples (U1 to U3)

	U1	U2	U3	S1	S2	S3
G1	8.95	8.75	9.11	9.38	9.90	9.81
G2	4.93	4.86	4.83	4.81	4.75	4.73
G3	7.01	6.21	6.45	6.12	6.04	6.41
G4	8.68	8.24	8.52	8.11	8.20	8.71
G5	5.40	5.11	5.31	5.08	5.09	5.08
G6	6.62	6.95	6.44	6.74	6.76	6.75
G7	5.85	5.77	5.73	5.57	5.44	5.75
G8	5.64	5.24	5.40	5.26	5.15	5.26
G9	4.97	4.77	4.79	4.70	4.77	4.82
G10	7.40	6.74	6.84	7.14	7.46	7.85
G11	7.45	7.42	7.44	7.27	7.34	7.61
G12	7.34	6.79	7.21	6.81	6.67	6.95

even k, may be unique. In other words, it is possible that there are several low-dimensional aspects of the data involving different subsets of the p initial features, which carry some sort of meaningful information. Finding them (and distinguishing them from spurious patterns) is the major challenge.

In the remainder of the paper, we will outline some methods that have been applied successfully for analysing high-dimensional data in the genomics arena. The number of techniques that have been proposed is quite substantial and it is not possible in a short review to outline all of them; therefore, only a few selected methods that we and our colleagues have found consistently useful will be presented and the focus will be more on the underlying concepts rather than on the details of the procedures as those can generally be found either online or in the literature. [Amaratunga and Cabrera (2004) and its second edition, Amaratunga *et al.* (2014), are book length treatments of this topic].

VISUALISATION

It is useful to first assess what type of signal may be present in the data, including checking to see whether the data exhibit any signal at all and also to see whether the data contain any anomalies such as outliers (i.e., unusual observations). One way of doing this is by constructing a biplot or a spectral map.

A biplot is a two-dimensional rendering of the data made possible by the singular value decomposition, whereby any matrix X can be decomposed as $X = UDV'$, where U is pxn and column orthogonal, V is nxn and orthogonal, and D is nxn and diagonal. If we retain only the two largest values in the diagonal of D (call it D_2) and subset U and V accordingly (call them U_2 and V_2), we can approximate X by $X_2 = A_2B_2$ ' where $A_2=U_2D_2^{1/2}$ and $B_2 = V_2D_2^{1/2}$. The two columns of A_2, when plotted against each other, will offer a two-dimensional rendering of the p features. Analogously, the two columns of B_2, when plotted against each other, will offer a two-dimensional rendering of the n samples. The two plots can be shown in a single diagram called a biplot (Gabriel, 1971), so that not only the individual characteristics of the p features and the n samples, but also possible associations between the two can be assessed.

A spectral map is a special biplot, which is constructed after certain modifications have been made to X such as adjusting for size and scale differences among the features (for details, see Wouters *et al.,* 2003). This enhances the display for microarray data.

Figure 1 is a spectral map of the OSCC data. It can be observed that the two sets of samples, to a large extent, separate along the direction of the x axis, with the SL samples (labelled S1 to S27) lying mostly on the right and the UK samples (labelled U1 to U21) lying mostly on the left. The separation is not perfect and there is some variability, but this is to be expected with this type of data. Figure 1 shows that there is adequate separation between the samples and that it is reasonable to expect to pick up a good signal with this data.

It is worth noting that this was an 'unsupervised' analysis in the sense that the construction of the spectral map did not make use of the information that the samples came from two distinct populations. It is simple to tweak it to be 'supervised', in which case the spectral map shown in the Figure 2 results. How the tweaking was done will be described in the next section and it can be seen that doing so visibly improved the separation between the two sets of samples. It is also possible that even better separation exists in higher dimensions, but this is difficult to visualise, although it can be investigated analytically. In any case, the separation we have seen so far indicates that there is adequate separation in the data, prompting further analysis.

INDIVIDUAL FEATURE ANALYSIS

When analysing grouped high-dimensional data, it is useful to next carry out a supervised analysis of each of the p features individually. This will give an indication as to which features are driving the separation between groups. This can be done using conventional statistical hypothesis testing techniques. For example, for the OSCC data, since there are two groups of samples, Student's t tests can be used. However, the sample size is often limited in these

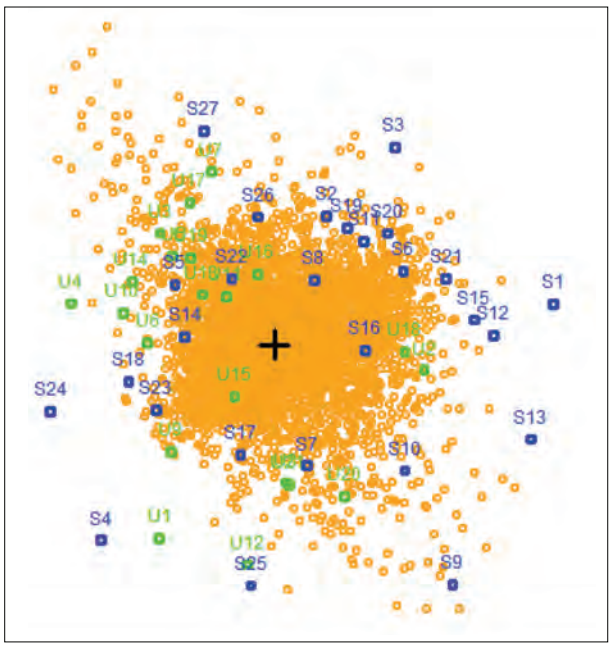

Figure 1: Spectral map of the OSCC data. The 27 blue squares represent the Sri Lankan samples, 21 green squares represent the United Kingdom samples, and the 8793 gold circles represent the genes.

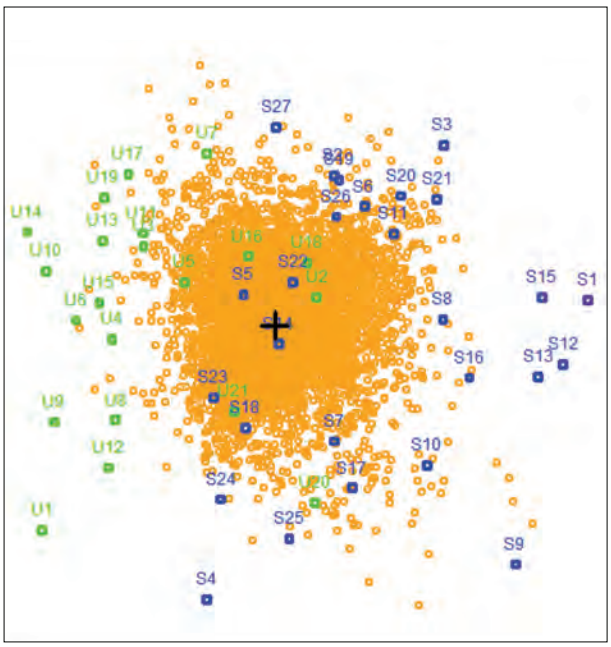

Figure 2: Supervised spectral map of the OSCC data. The 27 blue squares represent the Sri Lankan samples, 21 green squares represent the United Kingdom samples, and the 8793 gold circles represent the genes.

studies, which in turn reduces the power of these tests. Hence, it is often useful to 'borrow strength' across features to improve the sensitivity of the entire procedure. This can be done by setting some additional structures, such as setting a distribution structure for the variances. An outline of how this could be done for the two-group case follows.

Let the data be denoted as $\{X_{gij}\}$, where g ($g = 1,\ldots,p$) indexes the features, j ($j = 1,2$) indexes the groups, and i ($i = 1,\ldots, n_j$) indexes the samples (note: $n_1 + n_2 = n$). The standard t test assumes that X_{gij} is normally distributed with mean μ_{gj} and variance σ_g^2. Under these assumptions, the t test uses t test statistics $\{T_g\}$ to test whether $\mu_{g1} = \mu_{g2}$ for each feature, g.

A number of suggestions have been made as to how to borrow strength across the p features. Generally, they assume that the $\{\sigma_g^2\}$ collectively follow some distribution F_σ. Most proposed approaches are parametric in nature and assume a distributional form for F_σ, such as an inverse gamma distribution. The most widely used such method is limma (Smyth, 2004). This is a hybrid classical-Bayes approach in which a posterior variance estimate is substituted into the classical t statistic in place of the usual sample variance, giving rise to a moderated t statistic T_g^*. Like with the conventional t test, the null distribution of T_g^* can be adequately approximated by a t distribution, but with different degrees of freedom.

A semiparametric approach for borrowing strength, which is less dependent on distributional assumptions is Conditional t or Ct (Amaratunga & Cabrera, 2009). In this approach, even the normality assumption of the t test is dropped and it is assumed that X_{gij} follows an unknown distribution F. Now both F and F_σ are unspecified distributions and a resampling scheme along the lines of the bootstrap (Efron, 1981) is used to approximate them. They are then used to generate critical envelopes, $t_\alpha(s_g)$ (instead of constant critical values as in the conventional t test) for several different values of α; here s_g is the pooled standard error of the gth feature and α is the significance level of the test. For each feature, g, a p value, p_g, can be assigned by identifying the smallest value of α that results in significance for that feature.

For the OSCC data, both limma and Ct declared over 1000 features as significant at the traditional 5 % level; i.e., over 1000 features had p values less than 0.05; these can be considered to be the initial 'discoveries' by this analysis.

Clearly, however, when testing so many features, the likelihood is very high that there will be a large number of 'false discoveries' among the discoveries, i.e., differences that are declared significant even though they are not real. The traditional approach to this multiple testing problem has been to control the probability of at least one false discovery (called the Familywise Error

Rate or FWER), the Bonferroni procedure being the most common such fix. However, if this is done with a large number of features, there will be a substantial loss of statistical power and many true discoveries may go undetected. The False Discovery Rate (FDR) is a more recent alternative (Benjamini & Hochberg, 1995) that seeks to better address this problem.

This approach attempts to control the FDR, defined as:

FDR = Expected proportion of false discoveries among the set of discoveries,

rather than the FWER. There are multiple approaches to control the FDR, some common ones are by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001) and Storey (2002).

Operationally, FDR control is often done by converting the observed p values to FDR adjusted p values or q values. The FDR adjusted p value (also referred to as a q value) for feature g is the smallest FDR value for which the null hypothesis can be rejected for that feature and all others with smaller p values (Benjamini & Yekutieli, 2001; Storey, 2002).

For the OSCC data, we applied Storey's (2002) method to the Ct analysis results; 482 genes had $q < 0.01$ and 83 genes had $q < 0.001$ (note that unlike significance levels there is no value such as 5 % that is universally used; instead it depends on the situation and stringency required). Figure 2 was generated by assigning the 482 genes with $q < 0.01$ ten times the weights of the others in the spectral map; this resulted in a much better separation of the groups as we saw earlier. Figure 3 shows a heatmap of the 83 genes with $q < 0.001$ in all the 48 samples; samples 1 to 21 are UK and samples 22 to 48 are SL. The heatmap is merely a colour coded version of the 83×48 submatrix of X. The rows however have been reordered so that the rows that are similar to each other are placed close to each other; the same for the columns. This enables us to look for patterns in the data. The separation into two groups can be clearly perceived; the UK samples appear to the left and the SL samples appear to the right. Rather unexpectedly, there also appears to be a subset of SL samples (samples 22, 29, 30, 31, 33, 34, 36, 37) that is somewhat different from the others; this subset seems to separate better from the UK samples than the other SL samples. In an actual setting, the data analyst and the biologist would at this stage examine these findings, perhaps using published annotations of the genes, to try and interpret the findings biologically (Raghavan *et al.*, 2006).

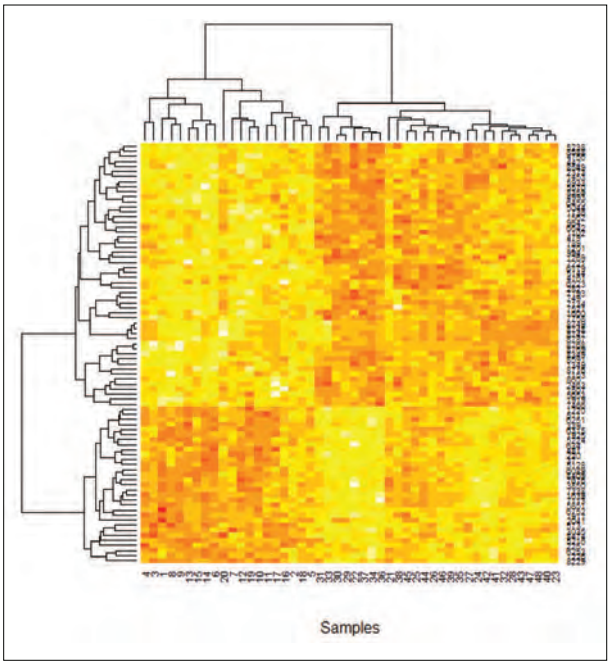

Figure 3: Heatmap of the top 83 genes of the OSCC data as determined by the individual feature analysis

ANALYSIS OF COMBINATIONS OF FEATURES

Next, it is useful to study combinations of features. Supervised classification (or discriminant analysis) refers to a class of techniques whose objective is to seek a combination of features that is able to discriminate between the groups of samples with reasonable accuracy. Again, there are a number of possible approaches. We shall now describe a method based on fitting a linear model for the case where the number of groups is two, as with the OSCC data.

Let Y_i indicate the group of the ith sample. Since there are only two groups, it can be assumed to be a binary random variable with a Bernoulli distribution:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

where π_i is the probability of sample i belonging to Group 1 and $1-\pi_i$ is the probability of sample i belonging to Group 2.

The values of the features for the ith sample are x_i, the ith column of X. The logistic regression model postulates that π_i is associated with x_i *via* the equation:

$$\log(\pi_i/(1-\pi_i)) = \beta' x_i,$$

where β is a *p*-vector of coefficients. In conventional logistic regression, these coefficients are estimated by maximising the log likelihood:

$$l(\beta) = \Sigma[Y_i \log(\pi_i) + (1- Y_i) \log(1-\pi_i)]$$

Note that here π_i depends on the features x_i and the coefficients β through the equation above, so that there is no closed form solution and the optimisation must be done algorithmically.

When $n < p$, there is insufficient data to estimate the full model. One way to overcome this problem is to maximise $l(\beta)$ under the penalty constraint $\Sigma|\beta_j| < h$, or, equivalently, to minimise

$$S(\beta) = - l(\beta) + \lambda\Sigma|\beta_j|$$

after scaling all features to have unit sample variance. This procedure is called lasso (Tibshirani, 1996). The tuning parameter λ controls the strength of the penalty: $\lambda = 0$ yields the standard regression estimates, $\lambda \to \infty$ yields all zero estimates, and the values of λ in between these two extremes yield compromises between fitting the traditional logistic model and shrinking all of its coefficients towards zero. A suitable value for λ is usually found by assessing the fit. Once this is done and the model fitted, many coefficients will inevitably shrink all the way to zero, essentially performing feature selection. A highly effective and efficient algorithm for lasso and a related procedure called elastic net was developed by Friedman *et al.* (2010).

The OSCC dataset was analysed using lasso and resulted in a model with 25 non-zero coefficients. As an initial check to see how good the fit was, the grouping for the 48 samples was predicted using the fitted model (samples with $\hat{\pi}_i > 0.5$ were assigned to group 1); the groups of all 48 were identified correctly; there were zero errors. This may seem to indicate a very good fit, but since the fitting and the prediction were done using exactly the same data, the zero error rate could be highly over-optimistic. Therefore, it would be unwise to rely on this as an assessment of goodness-of -fit.

It is best to measure the predictive ability of a model by validating it on a set of data that was not used to fit the model. This could be done by dividing the dataset into two parts, then using one part (called the 'training set') for fitting the model and using the other part (called the 'test set') for testing it. The group of each sample in the test set can be predicted using the fitted model and the proportion of errors can be calculated to give an assessment of the predictive accuracy of the model. This will not be over-optimistic because the test data were not used for model fitting.

However, there is often not enough data to allow a part of it to be left out for testing. A simulated version of the same idea is leave-one-out cross-validation (LOOCV) (Stone, 1974). This is carried out in *n* steps as follows. At the *i*th step, all the samples except sample *i* form the training set. The model is fitted using this training set. Then the group of the *i*th sample, which is now temporarily the test set, is predicted using the fitted model and it is noted whether or not the prediction is correct. This is repeated for each *i* (i.e., for each sample), in turn and the percentage of total errors is calculated and reported as the LOOCV error rate.

The LOOCV error rate for the OSCC data was 4.2 % (i.e., 2/48), indicating a good fit. We also kept track of which genes (features) appear in at least half the 48 fitted models; there were 22 of them, out of which many had appeared in the primary logistic model fit based on all 48 samples and also in the individual feature analysis. These are possibly the genes most influential for separating the two groups, although because of correlations among genes, certain influential genes may not show up. We put together these genes and the genes in the primary model and constructed a heatmap, which is shown as Figure 4. The separation between the two groups can be clearly seen, with again the UK samples on the left and the SL samples on the right.

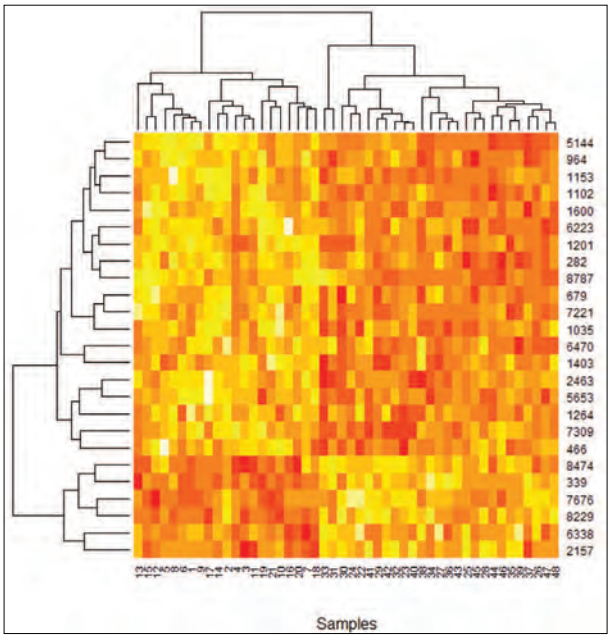

Figure 4: Heatmap of the top 25 genes of the OSCC data as determined by the combination feature analysis

Variations of cross-validation include leave-*k*-out cross-validation (in which *k* samples are left out at each step) and *k*-fold cross-validation (in which the original set of samples is randomly partitioned into *k* subsets, one of which is left out at each step). Another type of variation is the bootstrap (in which a random set of *n* samples is left in at each step, with the random sampling being done with replacement) (Efron, 1981; 1983; Breiman, 1996). There are variants of the bootstrap too, such as the 0.632 + bootstrap (Efron, 1983; Efron & Tibshirani, 1997).

When there are a large number of features, ensemble techniques, which iterate through samples of both rows and columns of X with the findings aggregated and collated at the end, have been found to work well. The initial popular ensemble technique was Random Forest (Breiman, 2001), which was mostly developed for mining early Big Data where n was very large. A variation called Enriched Random Forest (Amaratunga *et al*., 2008b) works well when $n < p$. An additional variation, in which lasso is used in a Random Forest like procedure (Amaratunga *et al*., 2012) has also been found to work well.

DISCUSSION

Big Data is increasingly becoming an important part of knowledge discovery in many fields. From a data analysis point of view, there are certain aspects of Big Data that are specific and unique. The most prominent is the very high dimensionality of the data and hence the need for dimension reduction *via* techniques such as principal components analysis and singular value decomposition. Another aspect, which occurs for example with data from internet search engines or data streams is sparseness. Such data might consist of only a few words. The data is typically stored as zeroes and ones where each variable is a word of the dictionary or a pair of words. This would create about 30 million features and in any observation there may be only 20 or 30 ones representing the words of a search string plus pairs of words; the rest are all zeroes. This data are so sparse that it is useful to compress it so that it occupies a very small space, and there are algorithms to extract the cases when needed and to do simple computations such as one step sparse logistic regression when an analysis is needed (Genkin *et al*., 2007).

In this paper, we have given an overview of techniques that are useful for analysing two group high-dimensional genomics data. Analogous techniques can be applied if there are r (where $r > 2$) groups or if the objective is to connect the p features to a continuous response variable. When there is no group information, the goal of the analysis might actually be to try and infer groups among the samples. For this, unsupervised classification (also called cluster analysis) techniques can be applied. Here too special methods have been developed for high-dimensional data, an example being an ensemble technique called ABC (Amaratunga *et al.*, 2008a).

Finally, as an important note; it is imperative that any findings be independently validated. Due to the high dimensionality, overfitting always remains a possibility, particularly in the selection of important features. Independent verification maybe sought through a repeat or similar study or through contextual subject-matter means. The importance of such independent qualification cannot be stressed enough.

REFERENCES

1. Amaratunga D. & Cabrera J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. John Wiley & Sons, New York, USA.
2. Amaratunga D. & Cabrera J. (2009). A conditional *t* suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication. *Statistics in Biopharmaceutical Research* **1**(1): 26 – 38. DOI: http://dx.doi.org/10.1198/sbr.2009.0003
3. Amaratunga D., Cabrera J., Cherkas Y. & Lee Y.S. (2012). Ensemble classifiers. *IMS Collection Volume 8, Contemporary Developments in Bayesian Analysis and Statistical Decision Theory : A Festschrift for William E. Strawderman* (edited by D. Fourdrinier, É. Marchand & A.L. Rukhin). Institute of Mathematical Statistics, Beachwood, Ohio, USA. DOI: http://dx.doi.org/10.1214/11-imscoll816
4. Amaratunga D., Cabrera J. & Kovtun V. (2008a). Microarray learning with ABC. *Biostatistics* **9**: 128 – 136. DOI: http://dx.doi.org/10.1093/biostatistics/kxm017
5. Amaratunga D., Cabrera J. & Lee Y.S. (2008b). Enriched random forests. *Bioinformatics* **24**: 2010 – 2014. DOI: http://dx.doi.org/10.1093/bioinformatics/btn356
6. Amaratunga D., Cabrera J. & Shkedy Z. (2014). *Exploration and Analysis of DNA Microarray and Other High Dimensional Data*. John Wiley & Sons, New York, USA. DOI: http://dx.doi.org/10.1002/9781118364505
7. Benjamini Y. & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**: 289 – 300.
8. Benjamini Y. & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**: 1165 – 1188.
9. Breiman L. (1996). Bagging predictors. *Machine Learning* **24**: 123 – 140. DOI: http://dx.doi.org/10.1007/BF00058655

10. Breiman L. (2001). Random forests. *Machine Learning* **45**: 5 – 32.
DOI: http://dx.doi.org/10.1023/A:1010933404324
11. Efron B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**: 589 – 599.
DOI: http://dx.doi.org/10.1093/biomet/68.3.589
12. Efron B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**: 316 – 331.
DOI: http://dx.doi.org/10.1080/01621459.1983.10477973
13. Efron B. & Tibshirani R. (1997). Improvement on cross-validation: the .632+bootstrap method. *Journal of the American Statistical Association* **92**: 548 – 560.
14. Friedman J., Hastie T. & Tibshirani R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *Journal of Statistical Software* **33**: 1 – 22.
DOI: http://dx.doi.org/10.18637/jss.v033.i01
15. Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**: 453 – 467.
DOI: http://dx.doi.org/10.1093/biomet/58.3.453
16. Genkin A., Lewis D.D. & Madigan D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**: 291 – 304.
DOI: http://dx.doi.org/10.1198/004017007000000245
17. Raghavan N., Amaratunga D., Cabrera J., Nie A., Jie Q. & McMillian M. (2006). On methods for gene function scoring as a means of facilitating the interpretation of microarray results. *Journal of Computational Biology* **13**: 798 – 809.
DOI: http://dx.doi.org/10.1089/cmb.2006.13.798
18. Saeed A.A., Sims A.H., Prime S.S., Paterson I., Murray P.G. & Lopes V.R. (2015). Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas. *Oral Oncology* **51**(3): 237 – 246. (Data available on *Gene Expression Omnibus* GSE51010)
DOI: http://dx.doi.org/10.1016/j.oraloncology.2014.12.004
19. Smyth G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**: Article 3.
20. Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* **36**: 111 – 147.
21. Storey J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**: 479 – 498.
DOI: http://dx.doi.org/10.1111/1467-9868.00346
22. Tibshirani R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B* **58**: 267 – 288.
23. Wouters L., Goehlmann H., Bijnens L., Kass S.U., Molenberghs G. & Lewi P.J. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* **59**: 1131 – 1140.
DOI: http://dx.doi.org/10.1111/j.0006-341X.2003.00130.x