Abstract

1 Pre-processing

1.1 Cohort-bias removal

For the cohort-bias removal we apply a genome-wise Location and scale (L/S) adjustment per cohort. Using a normalisation per cohort guarantees that the features have the same bounds over the cohorts and that the means are similar. The caveat of this approach is that we asume that the genome expression measurements are independent and we have no outliers. The standard normalisation transforms the genome expression values $\mathbf{x}$ per genome as follows

$$\mathbf{x}^* = \frac{\mathbf{x} - \overline{\mathbf{x}}}{\sigma}, \tag{1}$$

where $\mathbf{x}$ is the genome expression vector for some genome over all samples. This centers the mean and normalises the expression values with the standard deviation. To limit the influence of outliers we can center the median and use the interquantile range (IQR) for the scaling, i.e.

$$\mathbf{x}^* = \frac{\mathbf{x} - median\left(\mathbf{x}\right)}{IQR}, \tag{2}$$

To demonstrate the effect of these transformations with regard to cohort bias we take two genomes, one with high and one with low variance over the classifications.

There are various more elaborate methods to remove bias such as the SVD-based method from Alter et al.[1], the PCA-based bias removal methods EIGENSTRAT by Price et al.[9], MANCIE by Zang et al.[10], the distance weighted discrimination (DWD) approach from Benito et al.[2] or the ComBat method by Johnson et al.[7] who apply an empirical Bayes approach. A comparison of

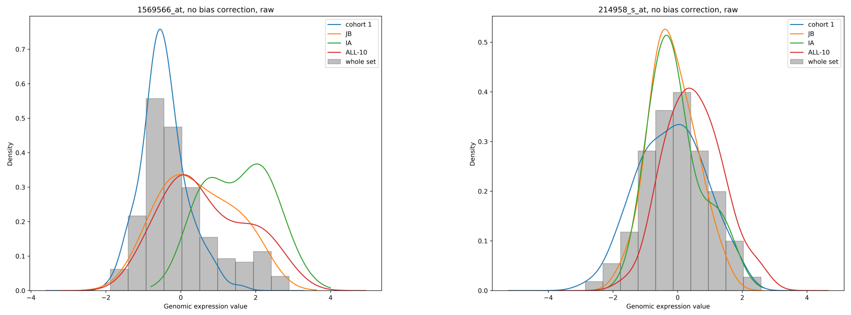

Figure 1: Two sets of distributions prior the bias correct, for, (left) a strong predictor and (right) a weak predictor

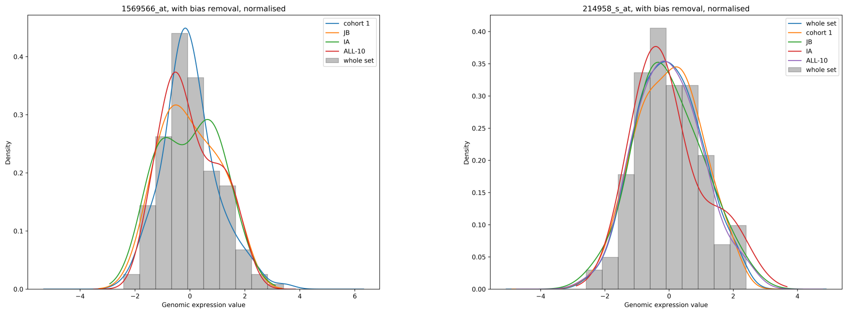

Figure 2: Two sets of distributions with L/S cohort correct, for, (left) a strong predictor and (right) a weak predictor

bias removal methods is out of scope for this work, for more details we refer the reader to Johnson et al[7]. The basic underlying assumption for all methods is that the samples are stratified over the cohorts, i.e. that in terms of patients each cohort represents a random selection from the total set of patients. Also, it is assumed that the distribution has only one mode.

1.2 Dimension reduction

1.2.1 Covariance based transformation

Principle Component Analysis (PCA): transformation of the feature space based on the eigenvectors of the covariance matrix. Can be applied to the entire dataset. The downside is that we obfuscate the biological meaning of the features: any value in the feature set of the transformed matrix is now a linear combination of N genome expression values, where N is the number of dimensions. Linear Discrimination Analysis: requires availability of classification for fitting, hence the transformation is biased to the training set, also the features are obfuscated similar to PCA.
Because the Covariance based transformation obfuscates the biological meaning of the feature vectors we choose variance-based feature reduction as the most suitable method to reduce the number of dimensions. As for LDA, variance-based feature reduction has a bias towards the training set because it dismisses features solely on the basis of variance across the different classifications which are obviously not available for the test set.

1.2.2 Variance-based feature reduction

We apply the False Discovery Rate method, with the Benjamin-Hochberg approach and the ANOVA model to determine the F-values, with the maximum p-value set at 0.05.

2 Classification

We will shortly describe the methods used for the predictions and the determination of genome importances. We will not go in detail on the selection of the method parameters, we refer the reader to the appendix for parameter selection.

2.1 Tree based

Single decision trees are known to be sensitive to changes in the input data. These ensemble methods help to decrease the variance without increasing the bias, i.e. increasing the ability to be generalised. We employ several tree-ensemble methods: Random Forest (RF) by Breiman[3], ExtraTrees (ET) by Geurts et al.[6] XGBoost (XGB) by Chen and Guestrin[5] and (Light)GBM (LGBM) by Ke et al.[8]. The RF and ET methods are ensemble methods that combine an arbitrary number of decision trees, using bootstrapped samples, random feature selection and a majority vote classification. The XGB and LGBM methods are ensemble methods that apply a technique called gradient boosting by Breiman[4].

2.2 Neural networks

We use 2 types of neural networks, a Deep Neural Network (DNN) and a Convolutional Neural Network (CNN).

2.3 Linear methods

Logistic Regression (LR), linear Support Vector Machines (lSVM), linear discriminant analysis (LDA)

Simplicity, transparancy.

2.4 Probabilistic methods

Naive Bayes (NB), Gaussian Processes (GPC), Relevance Vector Machines (RVM)

Table 1: Mean accuracies over 10 runs with 1% added random noise per run

	RF	DNN	CNN	LSVM	XGB	LDA
FDR $\alpha = 0.05$	0.38	0.29	0.38	0.43	0.25	0.42
FDR $\alpha = 0.1$	1.59	1.70	1.68	1.65	1.79	1.66
PCA $N = 200$	1.86	2.10	1.88	1.79	1.88	1.76
LDA $N = 200$	1.54	1.73	1.65	1.56	1.48	1.55
PCA $N = 500$	1.86	2.10	1.88	1.79	1.88	1.76
LDA $N = 500$	1.54	1.73	1.65	1.56	1.48	1.55

The tree-methods are not sensitive to the bias removal, or to the normalisation.

3 Post-processing

Description of weight/importance retrieval

4 Discussion

- if we choose PCA, LDA, check for inflection point in eigenvalue magnitude to 'smartly' select the number of components
- successively apply standard scaling and maxabs scaling to center cohort data?
- improve bias removal method L/S by ignoring outliers during normalisation
- we can combine the different models in one meta-model. This bagging of models increases the accuracy, removes method-specific biases and at the same time its helps reduce overfitting. The downside of bagging is that it obfuscates the results.

References

[1] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[2] Monica Benito, Joel Parker, Quan Du, Junyuan Wu, Dong Xiang, Charles M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Leo Breiman. Arcing the edge. Technical Report Technical Report 486, Statistics Department University of California, 08 1997.

[5] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[6] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[7] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[8] G. Ke, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T-Y Liu. Lightgbm: a highly efficient gradient boosting decision tree. In *31st conference on Neural Information Processing Systems*. NIPS, 2017.

[9] A.L. Price, Patterson N.J, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[10] C. Zang, T. Wang, K. Deng, B. Li, T. Xiao, S. Zhang, C.A. Meyer, H.H. He, M. Brown, J.S. Liu, Y. Xie, and X.S. Liu. High-dimensional genomic data bias correction and data integration using mancie. *Nature Communications*, 7, 2016.