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.

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

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-

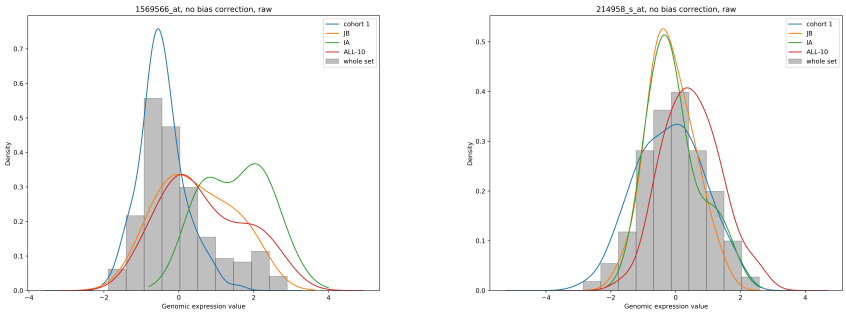

Figure 1: Two sets of distributions prior the bias correct, for, (left) a strong predictor and (right) a weak predictor

based method from Alter et al.[1], the PCA-based bias removal methods EIGENSTRAT by Price et al.[13], MANCIE by Zang et al.[16], the distance weighted discrimination (DWD) approach from Benito et al.[2] or the ComBat method by Johnson et al.[8] who apply an empirical Bayes approach. A comparison of

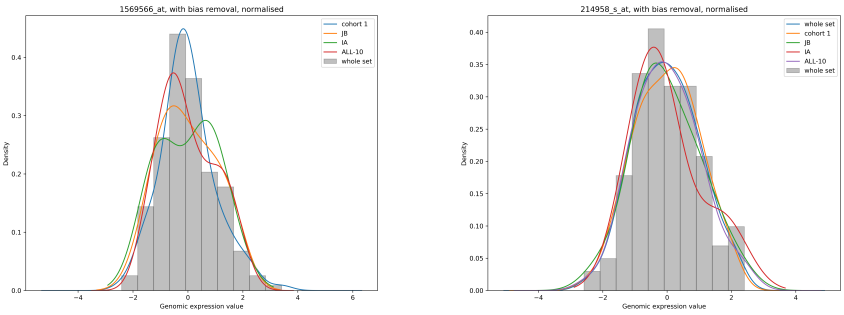

Figure 2: Two sets of distributions with L/S cohort correct, for, (left) a strong predictor and (right) a weak predictor

bias removal methods is out of scope for this work, for more details we refer the reader to Johnson et al[8]. 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. In figures 1 and 2 we show examples of distributions for two different genomes without and with bias removal respectively.

1.2 Dimension reduction

We considered several dimension reduction techniques such as Principal Component Analysis (PCA, see e.g. Shlens[15]), Linear Discriminant Analysis (LDA) and the False discovery rate (FDR, see Yoav and Hochberg[3]).
PCA is basically a transformation of the feature space based on the eigenvectors of the covariance matrix and can be applied to the entire dataset, including the test set. Using the eigenvectors of the covariance matrix as the basis for the features ensures maximal variance perpendicular to the coordinate axes. This is a coarse of saying that we maximize the information content per dimension. 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.
In LDA we try to find a linear transformation that maximizes the seperation **between** classes with respect to the seperation **within** classes, requiring two covariance matrices. LDA requires availability of the classification label for fitting, hence the transformation is biased to the training set, also the features are obfuscated similar to PCA. For both LDA and PCA we need to select the number of dimensions a priori.
The FDR method is basically a feature selection based on a minimum statistical seperarability of the distributions over the different classes. This minimum seperability is in this case the p-value for the rejection of the null hypothesis that the samples are drawn from the same distribution.
Because the Covariance or linear-discrimination based transformation obfuscates the biological meaning of the feature vectors we choose the FDR method as the most suitable method to reduce the number of dimensions. Also, the

FDR method is commonly applied in genomic research.
We apply the FDR method with the Benjamin-Hochberg approach and the ANOVA model to compare the distributions with a maximum p-value set at 0.05.
Arguably, a shortcoming of the FDR method is that, as for LDA, it 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. Another shortcoming is that it ignores feature interdependency, i.e. we may accidentally dismiss feature combinations as predictors because we have removed their constitutive parts.
For this reason, the use of a generic dimension reduction technique such as PCA is advised to improve the robustness of the model in terms of classification accuracy. As we are currently primarily interested in the importance of individual genomes and not per se the accuracy of the predictor we consider this out of scope.

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 the 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[4], ExtraTrees (ET) by Geurts et al.[7] XGBoost (XGB) by Chen and Guestrin[6] and (Light)GBM (LGBM) by Ke et al.[9]. 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[5].

2.2 Neural networks

We use 2 types of neural networks, a Deep Neural Network (DNN) [11] and a Convolutional Neural Network (CNN) [12]. The main advantage of neural networks is that they can learn nonlinear relationships between features. Despite the small sample size, it is interesting to apply neural networks in this context due to the high dimensionality of the data. Neural networks with multiple layers are proficient in discerning more subtle patterns in the data compared to other approaches. Shallow neural networks have been succesfully applied to similar sets of genetic expression data in the past, such as in [10]. A DNN, as shown in 3, uses several fully connected layers of nodes as a network architecture. A high level explanation of the difference between DNNs and CNNs is that a DNN looks at the entire dataset in each node (layers are fully connected). While a

CNN contains operations that allow it to focus on smaller subsets of the data (convolutional layers) and operations that allow it to filter out irrelevant data (pooling layers). An example of a CNN architecture for image classification is shown in 4.

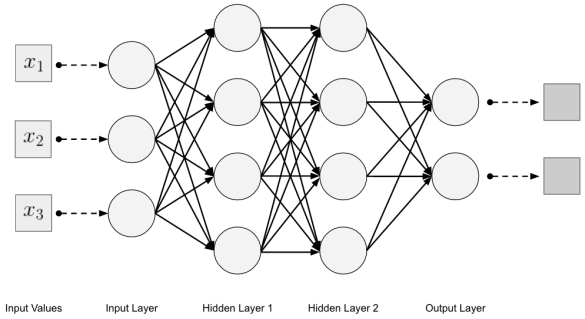

Figure 3: A generic architecture for a deep feedforward neural network.

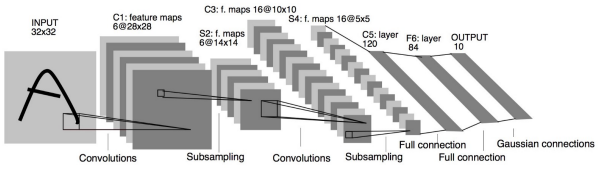

Figure 4: A typical convolutional neural network architecture for image classification from [12].

It is not possible out-of-the-box to discern which features are important for the final prediction. Using Local Interpretable Model-Agnostic Explanations (LIME) developed by Ribeiro et al.[14] we can get an idea of the so-called local decision boundary for individual samples which may assist in understanding the model decision in that particular case but I cannot

2.3 Linear methods

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

Simplicity, transparancy, robustness. However, what we gain in expressiveness we may lose in accuracy.

2.4 Probabilistic methods

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

3 Results

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

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

4 Post-processing

Description of weight/importance retrieval

5 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] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673, 2001.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[12] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[13] 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.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[15] Jonathon Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014.

[16] 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.