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 it is sensitive to 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 ensure similar bounds we can then scale the values by largest absolute value per genome vector, i.e.

$$\mathbf{x} = \frac{\mathbf{x}}{\max(\mathbf{x})}, \tag{2}$$

and as an alternative to the standard normalisation we can also apply a minmax normalisation

$$\mathbf{x} = \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}. \tag{3}$$

There are also robust scalers which are appropriate if outlying expression values are still present in the data. In our case we assume that these values are filtered out a priori.

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.[4], MANCIE by Zang et al.[5], the distance weighted discrimination (DWD) approach from Benito et al.[2] or the ComBat method by Johnson et al.[3] who apply an empirical Bayes approach. A comparison of bias removal methods is out of scope for this work, for more details we refer the reader to Johnson et al[3]. 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.

1.2 Dimension reduction

1.2.1 Covariance based transformation

Partial least squares: overfitting

Latent Dirichlet Allocation: requires availability of classification for fitting, hence the transformation is biased to the training set.

Principle Component Analysis: 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 . Any value in the feature set of the transformed matrix is a linear combination of N genome expression values, where N is the number of expression

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.

1.2.2 Variance-based feature reduction

Mann-Whitney U FDR, Benjamini-Hochberg procedure - ANOVA F-value FDR, Benjamini-Hochberg procedure - Wilcoxon

2 Classification

2.1 Tree based

2.2 Neural networks

Accuracy, but not transparant

2.3 Linear methods

Simplicity, transparancy.

2.4 Bagging

increases the accuracy, removes method-specific biases and at the same time helps reduce overfitting

3 Post-processing

importance/weights, what are they, how are they obtained?

4 Discussion

- check for inflection point in ordering of eigenvalue to 'smartly' select the number of components, if we choose PCA, LDA
- improve bias removal by mimicking featurewise distribution of expression values
- improve bias removal method L/S by ignoring outliers during normalisation

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] 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.

[4] 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.

[5] 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.