Introduction to Random Projection

Random projection maps original data onto low-dimensional subspace using a linear transformation matrix whose values are assigned randomly. 

\[ X_{\text{new}} \times W_{\text{new}} = X_{\text{new}} \]

The linear transformation matrix \( W \) is sampled as follows:

\[ w_{ij} = \sqrt{c} \begin{cases} 1 \text{ with prob } \frac{1}{2c} \\ 0 \text{ with prob } \frac{1}{2} \\ -1 \text{ with prob } \frac{1}{2c} \end{cases} \]

where \( c \) is set to \( \sqrt{d} \).

Characteristics of Random Projection

Advantages
- RP is data independent
- RP is fast, and meet the need of real-time processing
- RP reduce dimensionality of data without introducing significant distortion

Disadvantages
- RP generates transformation matrix without considering the structure of data

Our Contributions
- Improve classification accuracy of RP by adopting Feature Selection (FS), Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and etc.
- Give a faster and more accurate solution for large-p small-n problem whose dataset have few data points and many features.

Combination Methods

In this paper, we compared performance of following combination methods:
- Random Projection + Principle Component Analysis (RP+PCA)
- Random Projection + Linear Discriminant Analysis (RP+LDA)
- Feature Selection + Random Projection (FS+RP)
- Feature Selection + Random Projection + Linear Discriminant Analysis (FS+RP+LDA)
- Feature Selection + Random Projection + Post Feature Selection (FS+RP+PFS)

As an example, the procedure of FS + RP + PFS for classification is described as:

**Algorithm 1** - FS + RP + PFS

**Input:** Training data matrix \( X \in \mathbb{R}^{n \times d} \), label vector \( Y \in \mathbb{R}^{n \times 1} \), dimension after FS \( k_1 \), and dimension after RP \( k_2 \)

**Output:** The label of a testing sample \( TS \in \mathbb{R}^{1 \times d} \)

1. Map \( X \in \mathbb{R}^{n \times d} \) to \( X_1 \in \mathbb{R}^{n \times k_1} \) using transformation matrix \( W_1 \) generated by FS
2. Map \( X_1 \) to \( X_2 \in \mathbb{R}^{n \times k_2} \) using transformation matrix \( W_2 \) generated by RP
3. Map \( X_2 \) to \( X_3 \in \mathbb{R}^{n \times k_3} \) using transformation matrix \( W_3 \) generated by PFS
4. Train a SVM using \( X_3 \) and corresponding label \( Y \)
5. Reduce dimension for \( TS \) using \( W_3, W_2 \) and \( W_1 \) one by one, and predict the label for \( TS \) by SVM trained in step 4.

Datasets

We used three breast cancer gene expression datasets: BC-TCGA [1], GSE2034 [2], and GSE25066 [3].

Experimental Results

**Comparison among dimensionality reduction techniques based on Random Projection for cancer classification**

**Conclusions**
- Improvement of 14.77% of FS+RP compared to RP on BC-TCGA dataset
- Improvement of 13.65% of RP+LDA compared to RP on BC-TCGA dataset
- RP+RPCA yield lower performance compared to RP on genomic datasets

References

Acknowledgements

This work is partially supported by the National Key Research and Development Program of China (Grant No.2016YFC0901905), National Natural Science Foundation of China (Grant No.61471147), National Science Foundation of Heilongjiang Province (Grant No. F2016016) and the Fundamental Research Funds for the Central Universities (Grant No.HIT.NSRIF.2017037).