OXFORD

Gene compression using Machine Learning for Cancer type Classification

Sidarth Srinivasan [1], Natarajan Balaji Shankar [1] and Shruti Mohanty [1]

[1] University of California, Los Angeles, 90024, USA

Abstract

Motivation: In this project, we focus on using compression algorithms to capture latent features in gene data. These features reveal valuable information about the genomic space and can be used to generate hypotheses and develop robust models for cancer type classification. We also focus on using RNAseq data for this task. Choosing the right latent dimensionality and compression algorithm is an important step, and we rely on heuristics and the targeted task to make these decisions. Overall, our approach allows us to better understand the genomic space and improve cancer type classification.
Results: In this study, we trained three compression algorithms - PCA, ICA, NMF - on the TCGA dataset, and evaluated them across different metrics. We found that certain dimensionalities were optimal for each algorithm, and we used these to identify the best biological representations. Our approach achieved 92.22% accuracy for predicting cancer types. Overall, our study showed that the use of compression algorithms can help to improve the accuracy of cancer type prediction.
Availability: Code to perform all analysis and generate the results provided in this report is present in the GitHub repo.
Contact: sidarthsrini@ucla.edu ,balaji1312@ucla.edu , shrutimohanty@ucla.edu

1 Introduction

In recent years, advances in high-throughput sequencing technologies have made genome sequencing more affordable and allowed for the processing of genomes from a wide range of subjects on a large scale. However, the resulting collection of massive amounts of data requires the development of sophisticated techniques to effectively compress and extract the rich information from these sequences. Gene sequence compression algorithms are being developed for this specific purpose, and they have started to produce promising results.

There are simple algorithms, such as Principal Component Analysis, Independent Component Analysis, and Non-negative Matrix Factorization, that use linear expressions to reveal the relationships and expressions of the compressed feature set. There are also nonlinear methods that use advanced concepts such as deep learning networks and reconstruction networks to obtain even better representations of the feature set. In this project, we only experimented with linear techniques.

Compression techniques can find hidden biological and technical representations in gene expression data. Important details about the samples are revealed by these biological representations, which can also be used to develop hypotheses that are challenging to test in the original genomic space. The linear compression methods have been applied to large transcriptomic compendium to reveal the influence of copy number alterations on gene expression measurements, and to estimate cell-type proportion in bulk tissue samples Fehrmann *et al*., 2015. The nonlinear compression methods have revealed latent signals characterizing oxygen exposure, transcription factor targets, cancer subtypes, and drug response Tan J *et al*., 2017. Other latent variable approaches have been used to detect and remove technical artifacts, including batch effects Johnson *et al*., 2007.

Previous research has extensively explored these algorithms in terms of latent space dimensions, but the effectiveness of compression in downstream tasks such as cancer type classification has not been thoroughly studied. In this study, we focus on multi-class cancer type classification using data obtained from various gene compression techniques and propose our own approach, which outperforms all existing methods.

2 Approach

We implemented a pipeline shown in Figure 1 on TCGA dataset (https://gdc.cancer.gov/about-data/publications/ pancanatlas) using three different algorithms - PCA, ICA, NMF. These data consisted of 11,060 samples with 16,148 measured genes quantified with RSEM and normalized with log transformation. We then calculated performance metrics such as reconstruction loss, Pearson correlation, and Spearman correlation for the compressed features in various latent dimensionalities. Since there is no generalized latent space that captures all biological features from RNAseq data, we focused on the task of cancer-type classification. We initially implemented a logistic regression model to

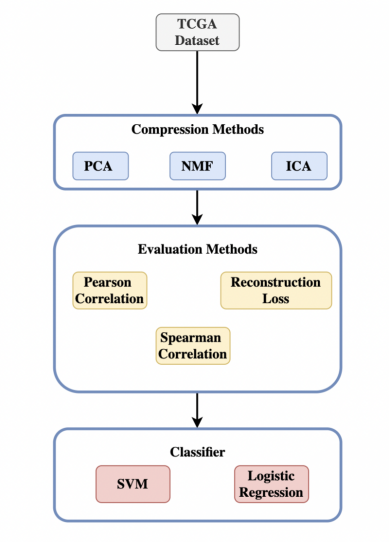

Fig. 1. Proposed model pipeline

perform one-vs-all classification to get a baseline performance for this task. We then performed multi-class classification using both logistic regression and SVM on the reduced feature space, achieving an accuracy of 92.22%. We also compared the performance of the model by comparing the classification accuracy on the entire dataset without any feature reduction.

3 Methods

The model pipeline is described in detail in this section. It provides information about the different steps and processes involved in the pipeline, and explains how they are used to achieve the desired results.

3.1 Gene Compression

We applied one-step linear compression algorithms such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Non Negative Matrix Factorisation (NMF) to reduce the dimensionality of the data. These algorithms are well-established and have been widely used in dimensionality reduction.

PCA identifies a unique and deterministic solution that represents the compressed features with a decreasing amount of variance. The directions identified by PCA are orthogonal to each other and linearly uncorrelated. As explained in Jolliffe *et al.*, 2016, selecting the topmost features in PCA often represents the entire dataset. However, there may be circumstances where the last few features are of interest for outlier detection. Therefore, selecting features across multiple latent dimensions is a more generalized representation model.

ICA, on the other hand, does not have an inherent ordering for its feature sets. As discussed in Kairov *et al.*, 2017, ICA defines a new coordinate system in the multidimensional space such that the distributions of the data points in the new axes are as mutually independent as possible. ICA is particularly effective at separating mixed signals when the subcomponents are derived from a non-Gaussian distribution. It has been widely used for the analysis of transcriptomic data to blind-separate biological and technical factors affecting gene expression data.

As explained in Frigyesi *et al.*, 2008, Non-negative matrix factorization (NMF) is a relatively new approach to analyze gene expression, which models data by additive combinations of non-negative basis vectors. Gene expression data is split into a product of two non negative matrices W and H. The k columns of W are the basis vectors using which factorization is performed in the genetic space. Its use has allowed for the linear application of large transcriptomic compendiums in gene expression data to reveal the influence of copy number alterations, identify coordinated transcriptional programs, and estimate cell-type proportion in bulk tissue samples.

The dataset was split into 90% training and 10% test sets balanced by the cancer type. Feature reduction was performed across 8 latent dimensionalities (k) in the range k=2 to k=200 using Scikit-learn for PCA, ICA, and NMF. These features were then fed into the classifier models for binary and multiclass classification. Their performance results are discussed in the next section.

3.2 Cancer Type Classification

We developed a unified pipeline that is able to accurately predict all 33 cancer types from the TCGA data set. After observing a definite trend in the latent space dimensions, we leveraged this for cancer classification. To being with, we developed a binary one-vs-all classifier that predicts in a 'yes' or 'no' format. We observed good model performance on that and since that is not an efficient method, we further extended it to multi-class to predict all cancer types.

4 Results

This section presents the evaluation metrics and the results obtained from our pipeline for gene compression and cancer-type classification.

4.1 Evaluation Metrics

The following metrics were used to evaluate the different compression techniques used in the project:

Pearson Correlation: Pearson's correlation coefficient is a measure of the linear correlation between two variables. It ranges from -1 to 1, where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation. In our case, the correlation between input and reconstructed output is denoted by:

$$\rho = \frac{cov(X,Y)}{\sqrt{std(X)} * \sqrt{std(Y)}}$$

Reconstruction Cost: It is denoted as the MSE between input features, and reconstructed output from latent dimensions. Mean Square Error (MSE) is denoted by:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

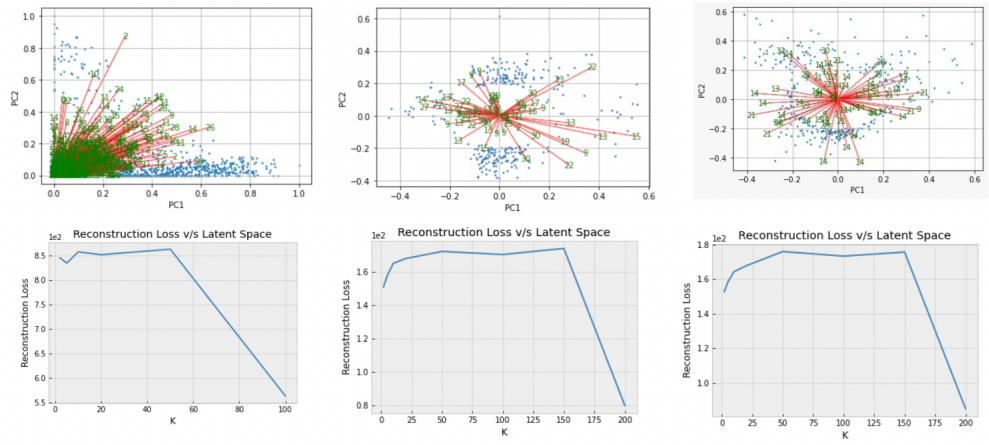

Fig. 2. Score and loadings biplot, and Reconstruction loss plots for the algorithms NMF, ICA, and PCA respectively.

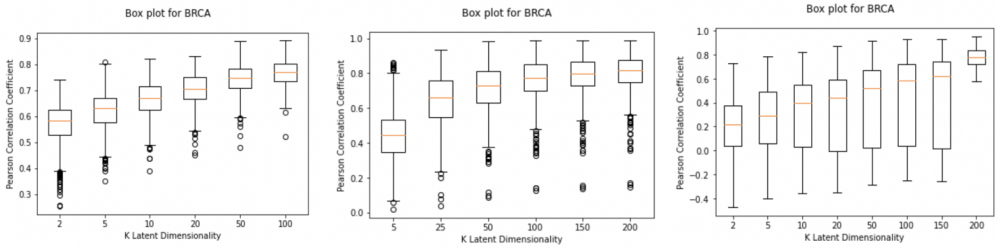

Fig. 3. Pearson Correlation box plots for the compression algorithms NMF, ICA, and PCA respectively.

Spearman Correlation: Pearson Correlation between ranked variables. The Spearman correlation coefficient is calculated using the ranks of the values rather than the actual values of the variables, which makes it less sensitive to extreme values and allows it to capture non-linear relationships:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

4.2 Qualtitative and Quantitative Analysis

We analyze the compression algorithms using the evaluation metrics mentioned above, along with the score, and loadings biplot. This is seen in Figure 2.The points on the biplot represent the individual observations in the dataset, and the position of each point is determined by the values of the two sets of variables for that observation, after compression. The biplot also includes vectors, called "loadings," which represent the direction and strength of the relationship between the two sets of variables. These loadings can be used to interpret the relationship between the variables and to identify any clusters or trends in the data. The observations are positive as expected for NMF, and for PCA, and ICA we can see the possible direction of maximum variance for different cancer-type clusters. From the reconstruction loss plots, we can eyeball an optimal latent space dimension for good reconstruction loss values when compared to larger dimensions at k=100, for PCA, and ICA, and k=50 for NMF. The plots for PCA, and ICA are similar as they are rotations of one another. Figure 3 displays the Pearson correlation coefficient box plots. It is the ratio of covariance between two standard variables and their standard deviation. This metric indicates the ability of models to capture specific information about sample composition across latent dimensionalities. An overall increasing trend was seen in the median correlation value and decreasing variance as the latent dimensionalities increased. Similar trends were observed for Spearman Correlation coefficients.

Beyond the compression metrics analysis, we analyzed our models based on their classification accuracies. The training set was used to train each compression algorithm, that was evaluated on the testing set. A binary logistic regression classifier was trained using the compression features from each individual algorithm for a one v/s all comparison, and a multiclass classifier was trained using compression features for a one v/s one comparison. The dataset was label encoded for this purpose. In almost all models, an increase in accuracy was reported with an increase in k, and then the accuracies saturated. The accuracies reported were better than the raw features, as the classifier failed to converge when the number of features was more than the number of samples. The highest accuracy of 92.22% was observed.

5 Discussion and Conclusion

Our observation from this entire analysis is that there is no correct dimensionality and several biological representations can be better revealed across a range of dimensions. Biological representations are enhanced for cancer-type classifications when gene expression data is compressed using linear compression algorithms into different latent space dimensionalities. We explored various compression techniques to derive the optimal latent space for the gene expression dataset and evaluated these techniques using correlation and reconstruction loss methods. In addition, we explored various classification techniques such as one vs all (binary classification) and multi-class classification (33 different cancer types)

Num. Components	Classifier Type	Data Compression (Accuracy)		
		NMF	PCA	ICA
2	Logistic Regression	15.01%	13.65%	11.03%
	SVM	12.93%	14.65%	25.86%
5	Logistic Regression	31.56%	49.91%	11.30%
	SVM	32.73%	56.69%	55.42%
10	Logistic Regression	58.59%	77.67%	11.75%
	SVM	62.75%	80.56%	80.74%
20	Logistic Regression	75.59%	87.70%	25.50%
	SVM	76.85%	88.43%	88.25%
50	Logistic Regression	87.61%	91.95%	35.08%
	SVM	88.25%	91.41%	90.05%
100	Logistic Regression	89.69%	**92.22%**	38.79%
	SVM	88.61%	91.77%	**92.22%**

Table 1. Accuracy results for Multiclass Classifier-Compression pairs

using LR and SVM classifiers. PCA reduced to 100 dimensions followed by logistic regression, and ICA reduced to 100 with SVM gives us the best accuracy of 92.22%. Overall, we identified that a reduced number of components tends to outperform the full dataset, due to the failure of the convergence of the classifier.

Along with this analysis, we also performed other experiments that include cross-validation. Because of the size of the dataset, we limited ourselves to kfold CV for latent dimension 5. We also tried unsupervised clustering using K-means and GMM, however, the accuracy for these methods was $<20\%$.

As machine learning continues to be applied to derive insight from biomedical data sets, researchers should shift focus away from optimizing a single model based on certain mathematical heuristics, and instead towards learning good and reproducible biological representations that generalize to alternative data sets regardless of compression algorithm and latent dimensionality. Subtle patterns in input signals can be identified in this approach which aids the task of cancer classification.

Code to perform all analyses and generate the results provided in this report is present in the GitHub repo. There are 3 different scripts in the code folder for each compression algorithm, and the respective binary and multi-class classification results. We have also provided scripts to perform a comparative analysis between the 3 algorithms, and cross-validation for a single latent dimension. In terms of the split of the work, each of us worked on one compression algorithm followed with analysis for cancer type classification. Balaji worked on refactoring the code for submission, and presenting results, Sidarth and Shruti worked on the report and documenting the findings.

References

Jolliffe IT, Cadima J. (2016) Principal component analysis: a review and recent developments, *Philos Transact A Math Phys Eng Sci.*, **374**, 20150202.

Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, et al (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis, *BMC Genomics.*, **18**, 712.

Frigyesi, A. & Höglund, M (2008) Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes, *Cancer Inf.*, **6**, 275–292.

Fehrmann, R. S. N., Karjalainen, J. M., Krajewska, M. (2015).Gene expression analysis identifies global gene dosage sensitivity in cancer, *Nature Genetics*, **47**, 115-125.

Tan J, Doing G, Lewis KA, Price CE (2017) Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks, *Cell Syst*, **5**, 63-71.

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics*, **8**, 118-27.

6 Appendix

Folders in the GitHub repo with additional results -

- Results for Cross Validation - Under Cross_Validation
- Results for Comparison Analysis across compression algorithms - Under Comparison
- Results for one v/s all classifier for all compression algorithms - Under ICA, NMF, PCA.
- Results for Pearson, Spearman correlation box plots - Under ICA, NMF, PCA