Dimensionality Reduction Techniques in Cancer-Screening Classification Models∗

Predicting p53 Tumor Suppressing Protein Mutations

Anthony Clark†  
 Computer Science  
 Wentworth Institution of Technology (WIT)  
 Boston MA U.SA  
 [clarka14@wit.edu](mailto:clarka14@wit.edu)

ABSTRACT

Since biophysical genetic assays have become more efficient and cost-effective, the corresponding genetic datasets have become extremely voluminous where each feature represents a structural attribute of the specific gene1. Classification models built on genetic datasets have become some of the leading methods in cancer screening techniques2. Biophysical structural attributes have the potential to indicate the probability of gene mutation and the downstream impact on associated cancers1. As the dimensionality of genetic datasets increases, the training of cancer screening classification models (CSCMs) requires exponentially more observations to generate accurate predictions, making data analysis computationally demanding. Through dimensionality reduction, CSCMs can achieve high prediction accuracy rates with exponentially fewer observations2. I use a high-dimensional genetic dataset pertaining to the tumor suppression protein p53 to train an array of CSCMs. I use both supervised and unsupervised dimensionality reduction techniques in an attempt to improve the prediction accuracy rate of CSCMs without increasing the total number of observations in the dataset. Contrary to what some might believe, standard dimensionality reduction via simple covariance and correlation matrixes offers a feasible CSCM accuracy optimization alternative to more advanced machine learning dimensionality reduction techniques such as linear discriminant analysis and principal component analysis. This is especially relevant when analyzing the occurrence of rare events such as gene mutations.

∗Abbreviated CSCMs

†Sole author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*WOODSTOCK’18, June, 2018, El Paso, Texas USA*

© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

https://doi.org/10.1145/1234567890

KEYWORDS

Principal Component Analysis, Linear Discriminant Analysis, K-Nearest Neighbors, p53 Tumor Suppressing Protein

1 Introduction

Cancers are associated with genetic mutations. Biophysical models provide accurate diagnosis of mutated cancer-associated genes1. These biophysical models, however, can be extremely large with a total number of features ranging from hundreds to tens of thousands. CSCMs built on high-dimensional biophysical genetic datasets require an enormous number of observations to train2. This is because cancerous gene mutations are rare events and as the dimensionality of a given biophysical model increases, the likelihood of a rare event being expressed in each dimension/ becomes exponentially low4,5. Additionally, biophysical attributes in large genetic datasets tend to possess a great deal of collinearity. The values and trends of one biophysical attribute are correlated with those of all the other biophysical attributes. This correlation can be exploited. The values and activity of certain biophysical attributes can be used to predict and classify mutations to a high degree of accuracy without having to measure all the associated biophysical attributes or the mutated gene itself2. Although collinearity can be exploited, it can also be a hindrance with respect to classification models. High degrees of collinearity inherent in each dataset implies a great deal redundancy5. High levels of redundancy can impact the accuracy of classification models. Collinearity does not just provide unnecessary redundant information but can harm the accuracy rates of classification models. To exploit collinearity enough to make accurate predictions without sacrificing model efficiency due to feature redundancy, certain features in each dataset must be selected and others neglected. There is a considerable amount of ambiguity around which features are to be selected and which features are to be neglected as there are countless approaches, each with different recommendations, depending on the domain and application4. The process of reducing the number of features in a dataset to favor model efficiency and accuracy is known as dimensionality reduction. Dimensionality reduction, in the context of machine learning algorithms, is a mathematical process of projecting high-dimensional data onto a lower dimensional representation5. The use of dimensionality reduction allows for CSCMs to be built on high-dimensional data without the need for excessive numbers of observations. In this paper, three methods of dimensionality reduction; linear discriminant analysis (LDA), principal component analysis (PCA) and simple covariance matrix dimensionality reduction are explored. Several CSCMs have been constructed before and after dimensionality reduction and a comparison has been made. In contrast to the two machine-learning dimensionality reduction techniques (LDA & PCA) a simpler covariance and correlation dimensionality reduction has also been applied to the dataset.

2 Data

The dataset used in this paper is a collection of biophysical models pertaining to the tumor suppressing protein “p53”.

2.1 Source of dataset

The dataset used in this paper was taken from the UC Irvine Machine Learning Repository. All the data was obtained by in vivo assays. The data was collected used by Danziger et al. to train cancer-screening active learning models1

2.2 Characters of the datasets

The entire dataset contains 5408 variables with “p53” representing the response variable and the remaining 5407 variables representing physical and biochemical attributes associated with the p53 gene. The first 4826 predictor variables represent 2D electrostatic and surface-based properties while the remaining 582 predictor variables represent 3D distance-based features. Each observation is a measurement of a different p53 gene, and the mutation of each gene is denoted by the response variable, which is categorical. The response variable has two possible outcomes. The first response variable outcome is “inactive”, which corresponds to non-mutated, non-cancerous p53 genes. The second response variable outcome is “active”, which corresponds to mutated and cancerous p53 genes. The dataset heavily favors “inactive”, non-cancerous p53 genes. The dataset has 16,722 in total, but due to processing constraints only the first 200 observations were used. The data was cleaned of all missing and null values. The biophysical attribute values (the features) were expressed as characters, so they had to be converted to numeric values for downstream classification models to be constructed.

3 Methodology

The cancer-screening classification models (CSCMs) were trained using K-nearest neighbors (KNN), linear discriminant analysis (LDA) and multiple logistic regression. KNN is known to decrease in prediction accuracy as dimensionality increases, but it has the potential to be highly accurate given certain feature parameters and adequate data. Principal component analysis (PCA) was used to analyze the biophysical structural dataset, sans response variable, to identify the primary principal component and the constituent primary loading features. CSCMs were made before and after PCA and their respective prediction accuracy rates were compared. CSCMs were also trained after being informed by standard correlation matrices. In this paper, classification models are trained to predict the mutation of the tumor suppression protein p53. This protein is associated with almost every known cancer and is paramount in the biochemical maintenance of cell proliferation3. Classification models trained to predict the mutation status of p53 offer vital insights in the detection of countless cancer types. For these classification models to be properly optimized, proper feature selection must be performed or else accuracy rates will suffer, and cancers have the potential to go undetected.

3.1 K-Nearest Neighbors

The K-nearest neighbors is a machine learning algorithm that utilizes proximity or nearness of observations to classify and predict data. The standard KNN algorithm uses Euclidean distances to compute nearness. While other distances (for instance: Manhattan distances) can be used, the standard method was used in this analysis6. The KNN algorithm does not include dimensionality reduction and so its prediction accuracy rate is subject to decline given an increase in the number of features4. K is a hyperparameter that informs the algorithm how many neighbors are required to determine the class of a prediction. Hyperparameter tuning was used to optimize K in KNN models throughout the analysis.

*Equation 1: KNN uses Euclidean distances to compute proximity of values and classify predictions.*

3.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised machine learning classification algorithm that utilizes linear combinations of features that maximize separability of classes. It takes high-dimensional data and projects it onto lower-dimensional space where class distinctions are most pronounced4. LDA utilizes a combination of Bayes theorem, multivariate gaussian probability distributions and linear algebra to project high-dimensional data onto lower-dimensional space and maximize class separability.

*Equation 2: LDA uses Bayes Theorem in dimensionality reduction*

3.3 Principal Component Analysis

Principal component analysis is an unsupervised machine learning model that reduces the complexity of high-dimensional datasets by projecting them into a lower-dimensional representation. Like LDA, PCA uses linear combinations to converge multiple dimensions or features into shared axes, thus decreasing the amount of representing dimensions in the dataset4. Unlike LDA, PCA is unsupervised meaning it is a re-expression, non-predictive algorithm with no target/response variable.

*Equation 3: The eigenvector equation is used in PCA to compute linear combinations of features*

3.4 Python Libraries

The following python libraries were used in this analysis

|  |  |
| --- | --- |
| **Library Name** | **Function** |
| Numpy | Numpy arrays for pandas functionality |
| Pandas | Pandas dataframes used to hold large datasets efficiently |
| Matplotlib | Used for quick visualization and data exploration |
| Sklearn | KNN, LDA, Log Reg and PCA models used from sklearn |

*Table 1. Details python libraries used in analysis.*

3.5 Comparative Analysis

KNN and LDA models trained before and after PCA-mitigated dimensionality reduction were compared to test the improvement of classification model performance. The loading coefficients of both LDA and PCA were used to construct lists of the most influential biophysical attributes, respectively. The biophysical attribute lists were then used to construct KNN and logistic regression models. The performance of the models was compared. Lastly, a KNN model, an LDA model and a logistic regression model using covariance matrix guided dimensionality reduction was constructed and compared with the PCA and LDA guided models. The p53 mutant accuracy rates are a measurement of how many mutated p53 genes were properly classified by a given model, the p53 normal accuracy rates are a measurement of how many normal p53 genes were properly classified by a given model and the total accuracy is a combination of the two

4 Results

The dataset was explored, and a considerable amount of collinearity was observed. Of the 5408 features, there were 29,246,464 potential linear combinations. After a covariance matrix was generated, there were found to be 304,839 combinations of features that had a correlation coefficient above 0.90. First the entire dataset was split into training and test data using an 80/20 train: test ratio. Both a KNN and LDA model were trained and evaluated prior to any dimensionality reduction. Since LDA has dimensionality inherent in its algorithm, it could process a vast number of features and obtain a relatively high accuracy rate. Its “active” gene prediction rate was only 50%, meaning it missed half of the cancerous p53 mutants. Unfortunately, the dataset strongly favored

4.1 Collinearity in the Dataset

The amount of collinearity inherent in the feature dataset rendered standard classification models impossible without excessive amounts of computing power. Also due to the immense number of features in the dataset, the full extent of feature collinearity could not be visualized. Collinearity had to be expressed through covariance and correlation matrices as well as LDA and PCA component loading coefficients.

A diagram of a number of blue dots

Description automatically generated

*Figure 1. Shows the correlation between the two biophysical attributes X0.018.26 and X0.003.34. Strong correlations, above 0.90 were observed between over 300,000 pairs of biophysical attributes indicating enormous amounts of collinearity in the dataset*.

4.2 Classification Models Before and After PCA

KNN and LDA models were trained before PCA-mitigated dimensionality reduction and compared with KNN and LDA models trained after PCA-mitigated dimensionality reduction. Results are shown in table 2. The total accuracy is misleading because the data is heavily skewed towards normal p53 genes. Therefore, the total accuracy can be misleading. A model can obtain a relatively high total accuracy (above 90%) while still maintaining a 0% P53 mutant accuracy rate. The p53 mutant accuracy rate is the most informative statistic in this context. It describes how many cancerous, mutated versions of p53 were properly identified in the training data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | PCA | P53 Mutant Accuracy | P53 Normal Accuracy | Total Accuracy |
| KNN | No | 0% | 92% | 92% |
| LDA | No | 67% | 94% | 92% |
| KNN | Yes | 0% | 93% | 89% |
| LDA | Yes | 0% | 97% | 89% |
| Log Reg | Yes | 0% | 92% | 92% |

*Table 2. Details the accuracy rates of 5 models. A KNN and LDA model built before PCA-mitigated dimensionality reduction, a KNN and LDA after PCA-mitigated dimensionality reduction and a logistic regression model after PCA-mitigated dimensionality reduction.*

The use of PCA-mitigated dimensionality reduction actually decreased overall classification model accuracy. There are two major reasons for this. One being that PCA is unsupervised, meaning it has no response variable. It is not predicting the class of an unknown value; it is simply re-expressing the predictor variables into lower dimensional space. The top ten biophysical feature list provided by the interpretation of the first principal component are not necessarily correlated with the response variable. They may explain most of the variance in the overall dataset, but this does not translate to the overall covariance observed with respect to the p53 mutant. There were also only 200 observations in the dataset and the mutation of the p53 gene is a rare event. Coupled with the small number of observations, PCA could not properly re-express the fluctuations in biophysical feature behavior that would account for rare mutation events. The only model in table 2 that had a p53 mutant accuracy rate above 0% was the LDA model trained prior to PCA-mitigated dimensionality reduction. Taking into consideration the unsupervised nature of PCA, this observation is reasonable. LDA has dimensionality reduction inherent in its algorithm and it is a supervised classification algorithm trained with respect to the p53 response variable. After PCA-mitigated dimensionality reduction, the LDA algorithm has lost pivotal biophysical feature information pertaining to the p53 mutant, which is why its accuracy rate decreased after PCA.

4.3 Comparison of PCA and LDA

The LDA loading coefficients provided a list of the top 10 biophysical features similar to PCA, but the LDA loading coefficients are trained with respect to the p53 response variable. A KNN and logistic regression model were trained using the biophysical features informed by the LDA loading coefficients and both models yielded 0% mutant p53 accuracy rates. The exact same was observed from the KNN and logistic regression models trained by the list of biophysical attributes informed by PCA. This was peculiar because the LDA model prior to PCA performed relatively well, however, the biophysical attributes derived from the LDA loading coefficients did not improve either the KNN or logistic regression model. Since dimensionality reduction in the LDA algorithm is concerned with projecting so many features onto lower dimensional space, in this case 5407 features, the linear combinations of features may not be as representative of the overall p53 mutant covariance as individual features themselves. The advanced mathematics involved in projecting so many features into lower-dimensional expressions may convolute the simpler correlations the mutant p53 response variable has with select features. A standard correlation matrix was compared with the two LDA and PCA dimensionality reduction techniques.

4.4 Comparison of Machine Learning Models with Simple Collinearity Matrix Guided Dimensionality Reduction

A list of top ten biophysical attributes was constructed using simple correlation/covariance matrices with response p53. The list of attributes was then used to create a data subset and build classification models using only these ten attributes.

|  |  |  |
| --- | --- | --- |
| Biophysical Attribute Name | Correlation with p53 | Type |
| X.0.012.6 | 0.58 | 2D electrostatic |
| X.0.012675 | 0.56 | 2D electrostatic |
| X0.024 | 0.56 | 2D electrostatic |
| X.0.094210543463157895 | 0.55 | 2D electrostatic |
| X0.6105068493150686 | 0.55 | 3D distance |
| X11.725 | 0.55 | 2D electrostatic |
| X.0.018.3 | 0.55 | 2D electrostatic |
| X0.012.1 | 0.55 | 3D distance |
| X0.026.3 | 0.55 | 2D electrostatic |
| X0.102 | 0.54 | 2D electrostatic |

*Table 3. Shows the correlation coefficients for the top 10 biophysical attributes out of the total 5407 features. 8 are 2D electrostatic and surface biophysical attribute while two are 3D distance-based*

Both a KNN and a logistic regression model were training using the list of biophysical attributes seen in table 3. This sub dataset yielded the most accurate results without increasing the number of observations. The simple approach of finding the biophysical attributes most correlated with the p53 mutant response variable and then selecting those attributes to train classification models yielded considerably higher p53 mutant accuracy rates than both the supervised (LDA) and unsupervised (PCA) dimensionality reduction techniques. The KNN model trained on simple correlation had a 100% mutant p53 accuracy rate.

|  |  |  |
| --- | --- | --- |
| Model Type | Dimensionality Reduction Technique | P53 Mutant Accuracy Rate |
| KNN | None (Negative Control) | 0% |
| KNN | Linear Discriminant Analysis | 0% |
| KNN | Principal Component Analysis | 0% |
| KNN | Simple Correlation | 100% |
| Log Reg | None (Negative Control) | 0% |
| Log Reg | Linear Discriminant Analysis | 0% |
| Log Reg | Principal Component Analysis | 0% |
| Log Reg | Simple Correlation | 50% |

*Table 4.* *Shows the p53 mutant accuracy rates for 8 classification models trained in this analysis. Details the dimensionality reduction technique used in attempts to improve p53 mutant accuracy rate.*

4.5 Addressing Overfitting in Classification Models

The KNN accuracy rate of 100% suggested a likelihood of overfitting. The biophysical attributes informed by simple correlation dimensionality reduction were used to re-query the full dataset, querying only the top 10 biophysical attributes informed by simple correlation with 16,000 observations. Both a KNN and logistic regression model were trained using this new subset and compared with the models shown in table 4. The p53 mutant accuracy rates in both models declined heavily with KNN decreasing from 100% to 67%. This suggests that the KNN model was highly subject to overfitting.

5 Discussion

The lack of improvement observed after PCA and LDA mitigated dimensionality reduction was surprising. Both PCA and LDA are leading dimensionality reduction techniques in the field of CSCMs4. The dataset was highly skewed towards non-cancerous p53 gene biophysical models. This paired with a relatively small number of observation-to-feature ratio rendered both LDA and PCA unreliable. That being said, LDA on its own, with no only it’s built-in dimensionality reduction obtained a relatively high p53 mutant accuracy rate of 67%. The list of biophysical attributes chosen in accordance with the pre-PCA LDA loading coefficients did not help improve downstream KNN of logistic regression models. The accuracy observed in the pre-PCA LDA model did not translate to other classification models using LDA-guided biophysical attribute preferencing. The biophysical list was somewhat arbitrarily cutoff at ten biophysical attributes. This could greatly impact the lack of accuracy improvement seen when applied to KNN and logistic regression models. The KNN and logistic regression models trained with the biophysical attributes informed by simple correlation dimensionality reduction, had the highest p53 mutant accuracy rates with the KNN model having a accuracy rate of 100%. When the data was re-queried with a larger number of observations and these models were applied to larger testing sets, the accuracy rate dropped enormously (in the case of KNN, down to 67%), suggesting overfitting in both models. This is to be expected in datasets with such a small observation-to-feature ratio. The analysis should be repeated on an instrument with more computing power so that correlation coefficients of the pairs of all 5407 biophysical attributes can be compared with respect to the p53 gene. The laptop used in this analysis was not capable of running such an analysis without crashing, which is why the data had to be parsed. Hyperparameter tuning was performed for all models, altering k for KNN, test size for others. This also inevitably led to some variation between models trained on the same data subset.

6 Conclusion

Despite the overfitting observed in the KNN and logistic regression models trained on data informed by simple correlation dimensionality reduction, the simple correlation methods proved to yield considerably higher p53 mutant accuracy rates than more advanced PCA and LDA dimensionality reduction techniques. When training cancer-screening classification models, multiple approaches of dimensionality reduction need to be explored and compared. Genetic datasets are very large and very complex, because of this PCA and LDA are leading dimensionality reduction techniques in the field due to their potential to project very high-dimensional data onto low dimensional representations. Despite their impressive sophistication, both PCA and LDA have the potential to overlook pivotal simple linear relationships between predictor features and response variable. These simple relationships can sometimes be observed using standard correlation/covariance matrices. The accuracy of cancer-screening classification models often can be lifesaving, which is why models need to be trained on extremely large datasets and multiple approaches of dimensionality reduction need to be explored. The cancer-screening models also must be tested on multiple datasets to ensure that overfitting has not occurred. The models developed in this analysis should not be used in clinical practice because they have only been trained on one dataset. It is often observed when examining classification models broadly, that simpler dimensionality reduction techniques can yield more accurate classification rates than more complicated techniques. It is because of this phenomenon that when developing cancer-screening classification models, simpler dimensionality reduction techniques should not be neglected purely in favor of more advanced techniques.

ACKNOWLEDGMENTS

I would like to acknowledge Professor Pang for her enthusiasm and knowledge. Her guidance in python programming and the foundations of data science have greatly helped me improve my skills in data analysis.

REFERENCES

[1] Danziger, Samuel A., et al. “Predicting Positive P53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning.” *PLoS Computational Biology*, edited by James M. Briggs, vol. 5, no. 9, Sept. 2009, p. e1000498. *Semantic Scholar*, <https://doi.org/10.1371/journal.pcbi.1000498>.

[2] Keshavarz-Rahaghi, Faeze, et al. “A P53 Transcriptional Signature in Primary and Metastatic Cancers Derived Using Machine Learning.” *Frontiers in Genetics*, vol. 13, Aug. 2022, p. 987238. *Semantic Schola*DOI: <https://doi.org/10.1145/567752.567774>

[3] Soussi, T. “The P53 Tumor Suppressor Gene: From Molecular Biology to Clinical Investigation.” *Annals of the New York Academy of Sciences*, vol. 910, June 2000, pp. 121–37; discussion 137-139. *PubMed*, <https://doi.org/10.1111/j.1749-6632.2000.tb06705.x>.

[4] Borade, Sushma Niket, and Ramesh P. Adgaonkar. “Comparative Analysis of PCA and LDA.” *2011 International Conference on Business, Engineering and Industrial Applications*, 2011, pp. 203–06. *IEEE Xplore*, <https://doi.org/10.1109/ICBEIA.2011.5994243>.

[5] Motsinger, Alison A., and Marylyn D. Ritchie. “Multifactor Dimensionality Reduction: An Analysis Strategy for Modelling and Detecting Gene - Gene Interactions in Human Genetics and Pharmacogenomics Studies.” *Human Genomics*, vol. 2, no. 5, Mar. 2006, p. 318. *Springer Link*, <https://doi.org/10.1186/1479-7364-2-5-318>.

[6] Al-Hadidi, Moh’d Rasoul, et al. “Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm.” *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*, 2016, pp. 35–39. *IEEE Xplore*, https://doi.org/10.1109/DeSE.2016.8.Conference Name:ACM Woodstock conferenceConference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00