Dimensionality Reduction Techniques in Cancer Screening Classification Models∗

Predicting p53 Tumor Suppressing Protein Mutations

Anthony Clark†  
 Computer Science  
 Wentworth Institution of Technology  
 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 gene. Classification models built on genetic datasets have become some of the leading methods in cancer screening techniques. Biophysical structural attributes indicate the probability of gene mutation and the downstream impact on associated cancers. 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 arduous. Through dimensionality reduction, CSCM can achieve high prediction accuracy rates with exponentially fewer observations. 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 to improve the prediction accuracy rate of CSCMs without increasing the total number of observations in the dataset. Contrary to what some might believe, dimensionality reduction offers a feasible CSCM accuracy optimization alternative to sampling methods such as boosting and bagging.

∗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 genes. 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 train. 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 low. Dimensionality reduction is a mathematical process of projecting high-dimensional data onto a lower dimensional representation. 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, I explore two method of dimensionality reduction. The first

Example format: xxxx.

2 Data

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

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

What’s the format and size of the datasets? What parameters/columns/rows/character and their units are included in this dataset. Use a table to explain this is recommended. Did you clean the data or convert any unit in the dataset? If so, what’s the formula/rule did you apply? Did you combine any datasets? If so, how do you combine them? Did you create any new category for analysis in the datasets? If so, what and how do you create?

3 Methodology

In this part, you should give an introduction of the methods/model. First, what’s the method/model. What’s the assumption of this method/model. What’s the advantage/disadvantage of this method/model. Why did you choose it. What Python module or function do you apply to apply this method/model. Any optional input/extra work did you adjust to make the results better. If you have multiple methods, feel free to use subsection 3.1, 3.2, 3.3, … to separate them.

3.1 Heading Level 2

3.2 Heading Level 2

…

Example format: The updated template, user manuals, samples, and required fonts, all are available at the URL <https://www.acm.org/publications/proceedings-template>. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

 (1)

**Continuation part of Paragraph Text** The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**



The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

4 Results

In this part, you need to select a reasonable way to deliver the result of your topic. For example, equation or numerical results, or visualization of your result. You also need to provide a clear explanation of all results and how to understand the results. If there exist any unexpected results, please explain why or possible cause of this special result. You can use subsection 4.1, 4.2, … to separate your results.

4.1 Heading Level 2

Example format: In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

5 Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

6 Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what’s the effect of this result on the real-world?

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here.

REFERENCES

Use the following ACM Reference format for your citation

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

[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] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

Conference Name:ACM Woodstock conference

Conference 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