Effective Data Dimensionality Reduction Workflow for High-Dimensional Gene Expression Datasets

Utsha Das*, Azmain Yakin Srizon†, Md. Al Mehedi Hasan‡, Julia Rahman§ and Md Khaled Ben Islam¶
Department of Computer Science & Engineering
*†‡§*Rajshahi University of Engineering & Technology*, Rajshahi, Bangladesh
¶*Pabna University of Science & Technology*, Pabna, Bangladesh
Email: *utshadas5@gmail.com, †azmainsrizon@gmail.com, ‡mehedi_ru@yahoo.com,
§juliacse06@gmail.com, ¶mdkhaledben@gmail.com

***Abstract*—While moving towards the era of 'Big Data', the scourge of dimensionality is growing an example of the most concerned obstacles in bioinformatics and biomedical research. Typically, an omics classification involves irrelevant and unnecessary features that can take a long time to compute and reduce classification performance. Previously, various researches showed that combining univariate and multivariate feature selection methods may enhance the enforcement of classification. In this research, we have proposed a workflow that can provide better classification performance by using fewer variables for gene expression data. To establish our statement, we started by taking four gene expression datasets: GSE5325, GSE6919/GPL8300, GSE6919/GPL92, and GSE6919/GPL93. We applied Student's t-test to discard redundant features. After that, Principal Component Analysis (PCA) was exercised to reduce the dimension of data. Wrapper Recursive Feature Elimination (RFE) method was performed over the reduced data to obtain the best combination of PCAs for better performance. Finally, the Support Vector Machine (SVM) was utilized to measure performance, and outcomes were compared with the previous researches. The results showed that our proposed approach produced a better performance with much fewer variables for gene expression data. All our research resources, documents, programs and snippets are located at https://github.com/Srizon143005/DataReductionWorkflow.**

***Index Terms*—T-Test, Principal Component Analysis, Recursive Feature Elimination, Random Forest, Support Vector Machine**

I. Introduction

'Big Data' refers to the tremendous amount of data created by modern technologies such as portable tools, sensors, and tracking gadgets [1], [2]. In bioinformatics and biomedical investigation, the increase of high-throughput methodologies such as gene expression datasets has produced an exponential increase in the dimensionality [3]. The term 'curse of dimensionality', originally asserted by Bellman in 1957 [4], is now a reality. High dimensional data can often lessen the classification correctness because of possessing redundant features that are present in the primary dataset [3].

Hence, the apparent solution will be to discard the redundant features using a univariate filtering method before treating the high dimensional data [5]. Although univariate feature selection approaches have been used widely for cancer classification handling microarray data [6], [7], it's still not sufficient because some features may seem insignificant by appearance but can act promising in incorporation. To defeat the obstacle, a collection of methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) has been proposed for blending the original features into a different and smaller subset of variables [3]. Finally, wrapper techniques like forward selection and backward elimination can be applied to data that combines a feature selection approach while learning or classification step [8].

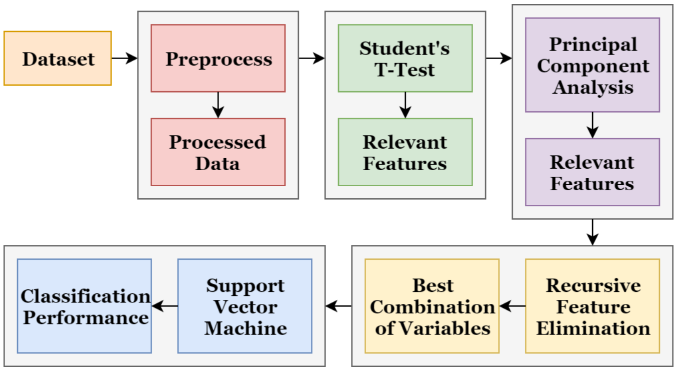
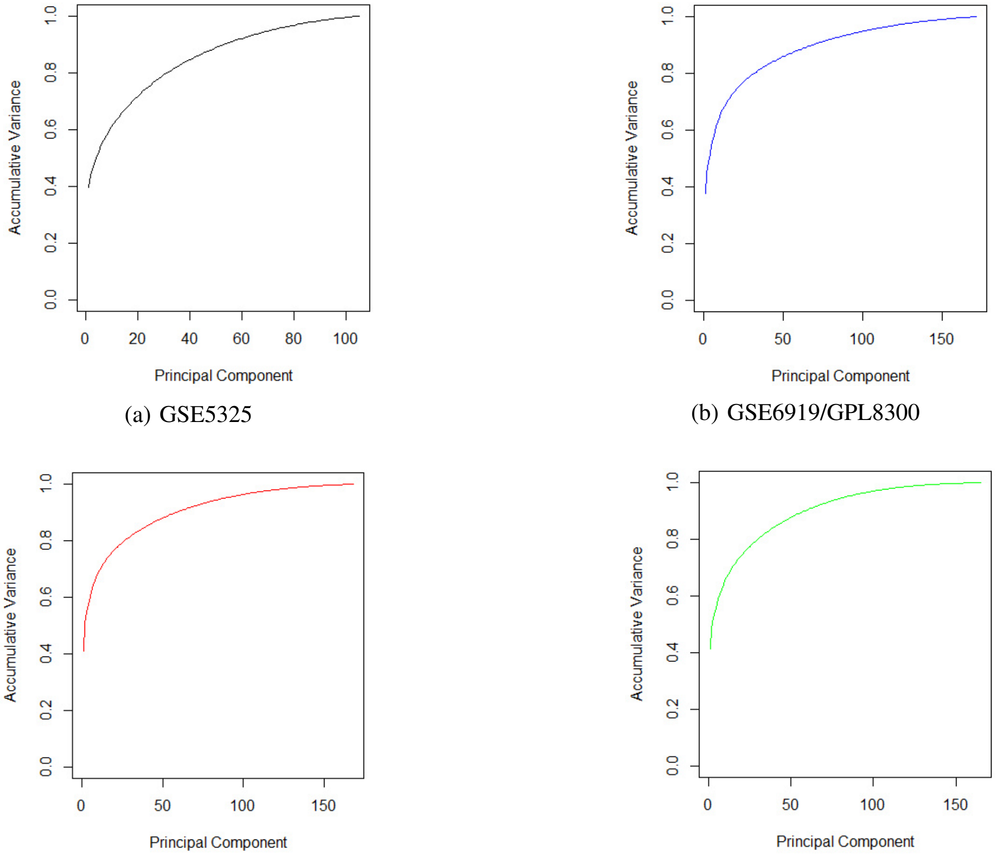

Dataset
Preprocess
Processed Data
Student's T-Test
Relevant Features
Principal Component Analysis
Relevant Features
Recursive Feature Elimination
Best Combination of Variables
Support Vector Machine
Classification Performance

Fig. 1: Proposed workflow diagram of our research with the assistance of Student's T-Test, Principal Component Analysis, Wrapper Recursive Feature Elimination merging with Random Forest and Support Vector Machine.

Previously, various researches showed that by applying one or multiple univariate and multivariate data reduction methods in sequential order, while taking a small number of variables into consideration, high classification accuracy can be obtained [3], [9]. In this paper, we have introduced a workflow to achieve better correctness by using even fewer variables (see Figure 1). First, we took four gene expression datasets and preprocessed them and a univariate parametric feature selection approach (Student's t-test) was applied. Principal Component Analysis (PCA) was implemented upon the reduced data afterward. The best combination of PCAs was selected by employing Wrapper Recursive Feature Elimination (RFE) method merged with the Random Forest (RF) approach. Finally, Support Vector Machine (SVM) [10] classifier was applied to obtain classification correctness and the outcomes were compared with the previous researches.

978-1-7281-7366-5/20/$31.00 ©2020 IEEE



(a) GSE5325

(b) GSE6919/GPL8300

(c) GSE6919/GPL92

(d) GSE6919/GPL93

Fig. 2: The aggregate variance of principal components for the investigation of corresponding datasets

II. Materials and Methods

In this study, Student's t-test was applied to discard irrelevant features and performed principal component analysis to reduce dimensionality. Afterwards, recursive feature elimination based on random forest has been used to obtain best combination of PCAs. Finally, support vector machine has been used to measure the performances.

A. Dataset Description and Processing

We selected four datasets for this research: GSE5325, GSE6919/GPL8300, GSE6919/GPL92 and GSE 6919/GPL93. All of them were downloaded from Gene Expression Omnibus [11]. GSE5325 contains 105 breast tumor samples where 24,648 genes were studied [12]. The dataset also contains estrogen receptor alpha status (1 indicates positive and 0 indicates negative), a transcription determinant perceived being critical to mimicking the completion of a considerable dimension of breast carcinomas and employed to investigate coformulation. On the contrary, we chose GSE6919 as it has been analyzed in details for feature selection studies [13]. GSE6919 holds three separate datasets in three platforms. GPL8300 platform has 81 normal and 90 tumor samples, GPL92 platform has 77 normal and 91 tumor samples and GPL93 platform has 75 normal and 90 tumor samples. Missing values were present in the picked datasets. Firstly, the genes with no values for any samples were discarded. Secondly, the missing values were replaced by the median of the corresponding feature values. For the datasets of all three platforms from GSE6919, log transformation was employed primarily.

B. Student's T-Test

Student's T-Test [14] was employed to recognize if there endured an equivalent or separate organization amongst two groups. Think of two blind individuals $a_{11}, a_{12}, ..., a_{1n_1}$ and $a_{21}, a_{22}, ..., a_{2n_2}$ where both of them indicates a regular arrangement holding means μ_1 and μ_2 having variances σ_1^2 and σ_2^2 respectively. Applying t-test, hypothesis $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ can be examined. While $\sigma_1^2 = \sigma_2^2$, test statistic is,

$$t = \frac{\overline{a_1} - \overline{a_2}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where

$$\overline{a_1} = \sum_{i}^{n_1} \frac{a_{1i}}{n_1}; \; \overline{a_2} = \sum_{i}^{n_2} \frac{a_{2i}}{n_2};$$

$$s_1^2 = \frac{\Sigma(a_{1i} - \overline{a_1})^2}{n_1 - 1}; \; s_2^2 = \frac{\Sigma(a_{2i} - \overline{a_2})^2}{n_2 - 1};$$

TABLE I: Accuracy, sensitivity, specificity in percentage, and their corresponding standard deviation (SD) after applying Student's T-Test, Principal Component Analysis (PCA), Recursive Feature Elimination (RFE) merging with Random Forest (RF) and Support Vector Machine (SVM) on all four datasets.

Dataset	Accuracy (%)	SD	Sensitivity (%)	SD	Specificity (%)	SD
GSE5325	90.24	5.23	86.24	5.98	94.21	6.43
GSE6919/GPL8300	82.75	4.22	89.00	6.61	77.81	3.69
GSE6919/GPL92	80.80	2.85	84.28	4.74	77.60	3.39
GSE6919/GPL93	82.00	2.00	84.19	2.31	79.90	4.12

and

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

While $\sigma_1^2 \neq \sigma_2^2$, the test statistic is,

$$t = \frac{\overline{a_1} - \overline{a_2}}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

Lastly, the P value was discovered utilizing the identical t values, having $n_1 + n_2 - 2$ degrees of freedom.

C. Principal Component Analysis

Principal Component Analysis is an example of the multivariate method that reduces the quantity of characteristics. Although PCA decreases the high dimensionality, it preserves maximum disparity among the predictor factors [15]. For gene expression analysis, PCA can express an individual by employing a comparatively small number of variables rather than thousands of them. At first, PCA takes mean normalized data as input. The following action is to compute the covariance matrix, which is meaningful for PCA.

$$cov(X, Y) = \frac{\sum_{i=0}^{M}(X_i - \overline{X})(Y_i - \overline{Y})}{M}$$

where X and Y signify features and $\overline{X}$ and $\overline{Y}$ indicate mean values for the corresponding characteristics and M denotes the amount of samples.

In assistance with the covariance matrix, eigenvalue and eigenvector can be calculated. Finally, PCA projects the earlier data to new data in terms of newly found basis or eigenvectors without reducing information. By taking only several variables with most of the variance can represent a sample rather than all the original features, therefore reduces dimensionality. In gene expression analysis, the amount of features is mostly higher than the amount of samples. PCA can lessen dimensionality at most the amount of samples by not dropping information [15], [16].

D. Wrapper Recursive Feature Elimination

Recursive Feature Elimination is a backward elimination technique developed on the idea of discarding the least significant variables in various steps [17]. In the Caret library, resampling techniques are considered in because of the change induced [18]. We applied Random Forests along with 10-fold cross-validation since assistant receptions for our research domain. Random forest was applied as it preserves the high accuracy of a large proportion of data. The purpose of cross-validation is to determine a dataset to examine the design in the preparation period to restrict obstacles such as overfitting [19].

E. Support Vector Machine

Support Vector Machine is a superintended training approach which can classify data with the guidance of discriminant function. The generalized form of discriminant function for Support Vector Machine classifier is [20], [21],

$$f(z) = \sum_{i=1}^{n} \alpha_i y_i k(z_i, z_j) + b$$

where $K(z_i, z_j)$ denotes the kernel of the SVM classifier. In this research, we have applied the radial basis function (RBF) kernel [22]. Hence, $K(z_i, z_j) = exp(\gamma|z_i - z_j|^2)$.

III. Experimental Analysis

At first, the Student's t-test is applied to all four datasets with Bonferroni adjustment. We chose only those features that have an adjusted P-value of less than 0.05, therefore, identified 705, 957, 431, and 278 features in GSE5325, GSE6919/GPL8300, GSE6919/GPL92, GSE 6919/GPL93 datasets respectively. After that, we applied Principal Component Analysis (PCA) on the reduced data. As the amount of features is extremely higher than the amount of samples for all four datasets, PCA reduces the amount of characteristics to at most the amount of samples for each dataset which are 105, 171, 168, and 165 respectively. We observed that 71,101, 88 and 84 number of PCAs can preserve 95% information with the guidance from the cumulative variance curve of principal elements (see Figure 2).

As PCA only concerns about the reduction of data, it's unclear if taking top PCAs can classify the normal and tumor samples correctly or not. Hence, we selected all the PCAs and implemented the Wrapper Recursive Feature Elimination (RFE) process in combination with the Random Forest (RF) method to identify the best combination of variables to achieve higher accuracy. We implemented 10-fold cross-validation to all four datasets, and this procedure was repeated 10 times. We observed that several features were repeated in each iteration. Hence, we chose those variables that produced higher accuracy and were present in almost every iteration. We identified 4, 4, 3 and 5 variables respectively. Hence, not more than 5 variables were selected for any datasets (Table II).

TABLE II: Principal components selected after applying Recursive Feature Elimination (RFE) merged with Random Forest (RF) on PCA-reduced data for each dataset.

Dataset	PCAs Selected
GSE5325	PCA1, PCA2, PCA102, PCA105
GSE6919/GPL8300	PCA1, PCA118, PCA162, PCA171
GSE6919/GPL92	PCA1, PCA8, PCA168
GSE6919/GPL93	PCA1, PCA2, PCA10, PCA82, PCA130

TABLE III: Comparison of accuracy and required variables between proposed and previous research [3].

Dataset	Method	Accuracy (%)	Variables
GSE5325	Proposed	90.24	4
	[3]	88.00	8
GSE6919/GPL8300	Proposed	82.75	4
	[3]	77.00	35
GSE6919/GPL92	Proposed	80.80	3
	[3]	80.00	5
GSE6919/GPL93	Proposed	82.00	5
	[3]	81.00	6

The selected variables were then used by the Support Vector Machine (SVM) considering the 10-fold cross-validation. SVM was applied as considered datasets were small and previous researches found better results with SVM. Tuning the SVM classifier with the 'radial' kernel provided the values of cost and gamma to calculate the accuracy, sensitivity, and specificity. To minimize the problem of class imbalance, we applied weighting factors for normal and tumor classes during learning. This process was repeated 20 times for GSE5325 and 10 times for others as the random state was set to random value and average of accuracy, sensitivity and specificity along with their corresponding standard deviation (SD) were calculated (Table I). Finally, we compared the outcome considering accuracy and the amount of variables with previous research for each dataset and observed that the methodology employed required fewer variables to provide better accuracy (Table III).

IV. Conclusion

In this paper, four gene expression datasets were observed. The cause of the observation was to propose a general workflow to reduce the dimensionality and increase the classification accuracy at the same time for gene expression datasets. In this research, we applied a univariate feature selection method (Student's t-test), a multivariate feature extraction method (Principal Component Analysis), Wrapper Recursive Feature Elimination (RFE) method merging with Random Forest (RF) and Support Vector Machine (SVM) classifier in sequential order. The experimental results revealed that by adopting Student's t-test as univariate feature selection, PCA produces better predictor variables. The outcomes also showed that applying RFE several times provide a better combination of PCAs and SVM requires at most five variables to offer greater accuracy. Hence, we concluded that the proposed workflow offers fewer variables and higher accuracy for gene expression datasets compared to the previous researches.

References

[1] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, p. 28, 2008.

[2] Y. Perez-Riverol, M. Bai, F. da Veiga Leprevost, S. Squizzato, Y. M. Park, K. Haug, A. J. Carroll, D. Spalding, J. Paschall, M. Wang *et al.*, "Discovering and linking public omics data sets using the omics discovery index," *Nature biotechnology*, vol. 35, no. 5, p. 406, 2017.

[3] Y. Perez-Riverol, M. Kuhn, J. A. Vizcaino, M.-P. Hitz, and E. Audain, "Accurate and fast feature selection workflow for high-dimensional omics data," *PloS one*, vol. 12, no. 12, p. e0189875, 2017.

[4] R. Bellman, "Dynamic programming and lagrange multipliers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 10, p. 767, 1956.

[5] K. Michalak and H. Kwaśnicka, "Correlation-based feature selection strategy in classification problems," *International Journal of Applied Mathematics and Computer Science*, vol. 16, pp. 503–511, 2006.

[6] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes, "Gene selection from microarray data for cancer classification—a machine learning approach," *Computational biology and chemistry*, vol. 29, no. 1, pp. 37–46, 2005.

[7] Y. Wang, F. Makedon, and J. Pearlman, "Tumor classification based on dna copy number aberrations determined using snp arrays," *Oncology reports*, vol. 15, no. 4, pp. 1057–1059, 2006.

[8] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[9] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[10] Z. R. Yang, "Biological applications of support vector machines," *Briefings in bioinformatics*, vol. 5, no. 4, pp. 328–338, 2004.

[11] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[12] L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmström, L. Memeo *et al.*, "Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7564–7569, 2007.

[13] S. Li and S. Oh, "Improving feature selection performance using pairwise pre-evaluation," *BMC bioinformatics*, vol. 17, no. 1, p. 312, 2016.

[14] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[15] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[16] S. Chambers, P. HOSKTNS, N. Haddad, F. Johnstone, W. McDicken, and B. Muir, "A comparison of fetal abdominal circumference measurements and doppler ultrasound in the prediction of small-for-dates babies and fetal compromise," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 96, no. 7, pp. 803–808, 1989.

[17] T. M. Phuong, Z. Lin, and R. B. Altman, "Choosing snps using feature selection," *Journal of bioinformatics and computational biology*, vol. 4, no. 02, pp. 241–257, 2006.

[18] M. Kuhn, "Variable selection using the caret package," *URL http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection. pdf*, 2012.

[19] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, p. 91, 2006.

[20] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[21] M. A. M. Hasan, S. Ahmad, and M. K. I. Molla, "Protein subcellular localization prediction using multiple kernel learning based support vector machine," *Molecular BioSystems*, vol. 13, no. 4, pp. 785–795, 2017.

[22] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, and M. D. Meyer, "Package 'e1071'," *The R Journal*, 2019.