Dimensionality Reduction of RNA-Seq Data

Isra Al-Turaiki
ialturaiki@ksu.edu.sa
College of Computer and Information Sciences,
Information Technology Department
King Saud University,
Riyadh, Saudi Arabia

Summary
RNA sequencing (RNA-Seq) is a technology that facilitates transcriptome analysis using next-generation sequencing (NSG) tools. Information on the quantity and sequences of RNA is vital to relate our genomes to functional protein expression. RNA-Seq data are characterized as being high-dimensional in that the number of variables (i.e., transcripts) far exceeds the number of observations (e.g., experiments). Given the wide range of dimensionality reduction techniques, it is not clear which is best for RNA-Seq data analysis. In this paper, we study the effect of three dimensionality reduction techniques to improve the classification of the RNA-Seq dataset. In particular, we use PCA, SVD, and SOM to obtain a reduced feature space. We built nine classification models for a cancer dataset and compared their performance. Our experimental results indicate that better classification performance is obtained with PCA and SOM. Overall, the combinations PCA+KNN, SOM+RF, and SOM+KNN produce preferred results.

Key words:
Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Self-Organizing Maps (SOM), RNA-Seq, Dimensionality Reduction

1. Introduction

There have been several major advances in the field of Bioinformatics over the past decades. Today, the scientific community is witnessing the accelerated production of biological data with the advent of high-performance technologies. As of February 2021, GenBank contains more than 776 billion bases and more than 226 million sequences [1]. Analyzing this huge data manually is impossible. Thus, computational power and tools are an important aspect of biological data collection and research.

RNA-Seq is a technology that facilitates the study of the entire transcriptome via next-generation sequencing (NSG) tools. It is a high-performance supplement to traditional RNA/cDNA cloning and sequencing techniques. Using RNA-Seq , it is possible to examine the expression levels of transcripts and alternate splice variants [2]. The rich information provided by RNA-Seq data can be utilized to advance many applications, such as disease classification and diagnosis, as well as the identification of potential biomarkers [3]. Like other data generated using NSG tools, RNA-Seq data are characterized as being a high-dimensional dataset because the number of variables (i.e., transcripts) far exceeds the number of observations (e.g., experiments). Using a large number of features to train machine learning algorithms leads to overfitting, which yields poor performance on real data. Thus, the *curse of dimensionality* limits the direct application of machine learning algorithms to RNA-Seq data. One important analytical aspect in the processing of RNA-Seq data is *dimensionality reduction (DR)*. Effective DR techniques transform the data to a lower dimension; in this way, the essence of the input data is retained, while noise and redundant features are eliminated. Dimensionality reduction is widely used to analyze high-dimensional data, such as bioinformatics datasets, in which hundreds of measurements are collected from a single sample simultaneously [4] [5].

Reducing the dimensions of the datasets reduces the time and storage space required. In addition, the removal of redundant and corelated features increases the ability of machine learning algorithms to learn from the dataset. When data are limited to very low dimensions, such as 2D or 3D, they become easier to visualize. In general, DR techniques can be classified as *linear* or *nonlinear* [6]. In linear DR techniques, a simple linear function is used to transform high-dimensional datasets into lower-dimensional datasets. Examples of *linear* DR techniques include *principal component analysis* (PCA), *singular value decomposition* (SVD), latent semantic analysis, locality preserving projections, independent component analysis, linear discriminant analysis, and projection pursuit. In *nonlinear* DR techniques, reduced dimensions are obtained through nonlinear transformations of the original dimensions [6]. Examples include kernel principal component analysis, multidimensional scaling, Isomap, locally linear embedding, self-organizing map, learning vector quantization, and T-Stochastic neighbor embedding.

Manuscript received March 5, 2021
Manuscript revised March 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.3.4

The choice of a DR technique depends on the nature of the dataset [7].

Given the wide range of dimensionality reduction techniques, it is not clear which is best for RNA-Seq data analysis. Little work has been done to evaluate and compare the effectiveness of various dimensionality reduction methods for RNA-Seq analysis. In this study, we attempt to fill this gap by studying the efficacy of three dimensionality reduction techniques to improve the classification of the RNA-Seq cancer dataset. In particular, we investigate the performance of PCA, SVD, and SOM. A total of nine classification algorithms are built using combinations of the three dimensionality reduction techniques and three classification algorithms. The goal is to study how dimensionality reduction techniques affect the performance of the classification models in terms of classification accuracy.

The paper is organized as follows: Section 2 describes the methodology used in the present work, including the dataset, dimensionality reduction techniques, and classification algorithms. Section 3 describes the experimental setup, performance measures, and experimental results and provides a relevant discussion. Section 4 presents the conclusions drawn from the study.

2. Methodology

In this section, a brief description of the RNA-Seq dataset used in this study is presented. We also describe the DR techniques under investigation and the classification algorithms used to assess their performance. Figure 1 shows a schematic view of our study design.

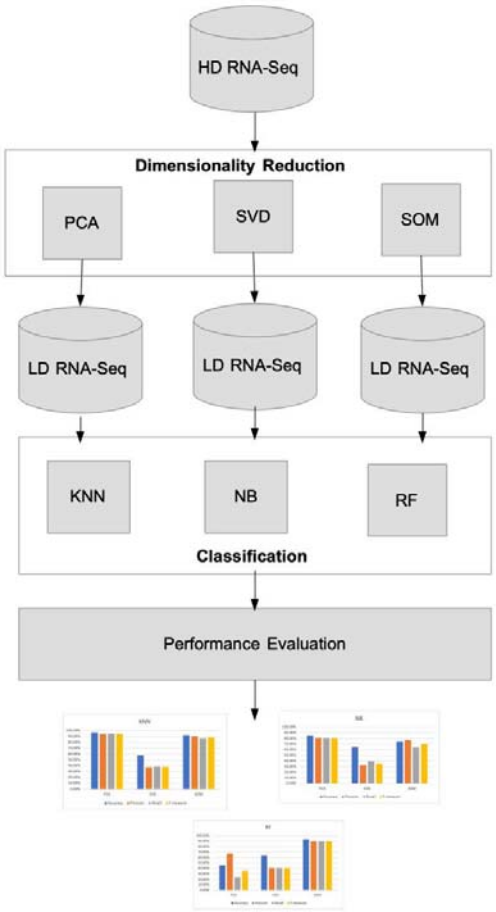

Figure 1 Schematic diagram of the experimental study design. (HD: high-dimensional, LD: low-dimensional)

2.1 Dataset Description

We conduct our study using tumor gene expression data collated by *The Cancer Genome Atlas Research Network* [8]. The RNA-Seq dataset was prepared by Ferles et al. [9]. The dataset consisted of 2,086 samples (records) with five classes of cancer labels: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and uterine corpus endometrial carcinoma (UCEC). There are a total of 972 features, each representing the *reads per kilobase of transcript per million mapped reads* (RPKM) RNA-Seq values of a specific gene. Table 1 summarizes the dataset's features.

Table 1: The distribution of cancer types in the RNA-Seq dataset

Cancer Type	Number of Samples
BRCA	878
KIRC	537
LUAD	162
LUSC	240
UCEC	269

2.2 Dimensionality Reduction

2.2.1 Self-Organizing Maps:

A self-organizing map (SOM) [10] is a nonlinear dimensionality reduction technique that is based on artificial neural networks. It utilizes an unsupervised learning algorithm to reduce the input dimensions. A map is iteratively formed in which similar records are placed closely together. The lower-dimension representation of the dataset is obtained using nonlinear mapping. An SOM is trained through *competitive learning*, in which a competition for activation takes place between output neurons. In each iteration, only one neuron at a time is activated. The selected neuron is thus the closest to the input record. The selection is based on the following calculation [11]:

$$\|z - c_c\| = \min_i \{\|z - c_i\|\}$$

where z is input record, c_c is the selected center, and c_i is the current center of the evaluation. Then, weight vectors that lie inside the neighborhood radius are updated as follows, where h_{ci} is the neighborhood radius:

$$c_i(t+1) = c_i(t) + h_{ci}(t)[z(t) - c_i(t)]$$

Self-organizing maps are easy to understand and offer an interactive and intelligible description of the results. They are effective in managing many forms of classification problems. Reducing dimensionality and grid clustering makes it possible to observe correlations in the results [12].

2.2.2 Principal Component Analysis:

Principal component analysis (PCA) [13] is a widely used linear dimensionality reduction technique. It eliminates noise and redundant variables while maintaining much of the variance of the data. With the lower-dimension representation of the dataset, PCA allows better visualization for assessing similarities and differences between data points, as well as clusters.

This is done by computing the covariance matrix in order to identify correlations between input variables. Eigenvectors and eigenvalues are then computed for the covariance matrix to find the *principal components (PCs).*

The eigenvalue λ can be determined by solving the following equation:

$$(\lambda I - A) = 0$$

where A is an n x n matrix and I is the identity matrix. The corresponding eigenvector v is calculated as follows:

$$(\lambda I - A)v = 0$$

Principal components are new variables that are created as a *linear combination* of the original input variables. Feature vectors are created using the eigenvectors and then used to project the data from the original axes to the PCs. This step requires multiplying the transpose of the original dataset by the transpose of the feature vector. Compared to other DR techniques, PCA is simple and has a lower computational cost [14].

2.2.3 Singular value decomposition:

Singular value decomposition (SVD) [15] is a dimensionality reduction technique that is based on matrix decomposition. Given a matrix A, containing real values and of size m x n, SVD decomposes A into three other matrices, U, S, and V*. Given the decomposition of any matrix, it is possible to reconstruct the original matrix. The dimensions of the three matrices are as follows: U is an *m x p* matrix, S is a *p x p* matrix, and V is an n x p matrix. The values contained in cells on the diagonal of S are called *singular values* of A. The columns of U are called the *left-singular vectors* of A, and the columns of V are called the *right-singular vectors* of A. The singular values play an important role in defining the variance of singular vectors. Accordingly, we can utilize these data to restrict the number of vectors to the preferred amount of variance, thus reducing noise in the raw dataset.
Although computationally expensive, SVD has been used in many applications, such as digital image processing, biological sequences classification, and pattern recognition, among many other [6].

2.4 Classification Algorithms

We assess the chosen dimensionality reduction techniques based on their ability to improve classification accuracy.
In this study, we use naïve Bayes (NB) [16], random forest (RF) [17], and K-nearest neighbor (KNN) [18] to perform the classification task. Below, we briefly describe each algorithm.
Naive Bayes is a simple probabilistic classifier based on the Bayes theorem. It is naive because it assumes independence between class attributes. However, it has comparable performance to other classification algorithms.
Random forest is a decision-tree ensemble method that creates multiple trees via a re-sampling process called

bagging (bootstrap aggregation). Many decision trees are constructed by re-sampling using bootstrapping with placement. Each node of the tree is divided using a subset of the attributes selected randomly for each tree. The class membership for a new example is predicted as the most commonly predicted class from the (aggregated) decision trees by a simple unweighted majority vote. This method is becoming widely used and has been established as highly effective for highly complex multi-criteria decision-making problems in a variety of fields

K-nearest neighbor is a similarity-based classification algorithm. In KNN, a new data point is classified based on the K neighboring data points. A majority vote between the neighbors takes place each time an unlabeled data point arrives. KNN is easy to understand. However, in the case of imbalanced datasets, classification results may be biased.

3. Experimental Results

3.1 Experimental Setup

In this study, we used the DR techniques and classification algorithms implemented in RapidMiner Studio Version 9.7.002 [19]. All algorithms were run using the default parameters. The experiments were run on MacBook Pro, with the macOS Catalina operating system, Version 10.15.7, and a 2.3 GHz 8-Core Intel Core i9 with 16 GB of RAM.

We reported the performance measures of the classification model in terms of accuracy, precision, recall, and F-measure using ten-fold cross-validation. Each measure is defined as follows:

- Accuracy is the percentage of correctly classified cancer samples, and it is calculated using Equation 2.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

- Precision is the percentage of samples of a given cancer type that are correctly classified out of all the samples predicted to belong to that cancer type. Precision is calculated as shown in Equation 3.

$$PR = \frac{TP}{TP + FP} \quad (3)$$

- Recall is the percentage of samples of a given cancer type that are correctly classified out of the total number of samples belonging to that cancer type. Recall is calculated as shown in Equation 4.

$$RE = \frac{TP}{TP + FN} \quad (4)$$

where TP represents true positives, FP represents false positives, FN represents false negatives, and TN represents true negatives.

The F-measure is the harmonic mean of precision and recall, and it is calculated via Equation 5 below.

$$F_1 = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

3.2 Experimental Results

A total of nine classification models were trained using reduced feature space with PCA, SVD, and SOM. Here, we discuss the performance of the nine models in order to highlight the effects of the three dimensionality techniques. The effect under investigation is related to the ability of the obtained models to improve classification accuracy.

First, we consider the effect of dimensionality reduction techniques with respect to each classification algorithm. Figures 1, 2, and 3 show the evaluation measures for KNN, NB, and RF, respectively. For KNN, the best accuracy, precision, recall, and f-measure values are obtained when PCA is used to reduce the feature space. Comparable results are obtained using SOM. However, performance degrades by a large margin with SVD. We observe similar behavior with NB because classification improves with PCA and SOM, as shown in Figure 2. For RF, the effect of dimensionality reduction is rather different. In fact, SOM performs considerably better than PCA and SVD, as shown in Figure 3.

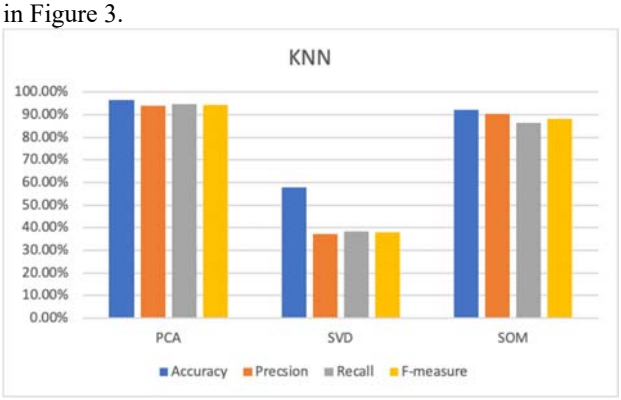

Figure 2 The evaluation measures for KNN models with dimensionality reduction techniques

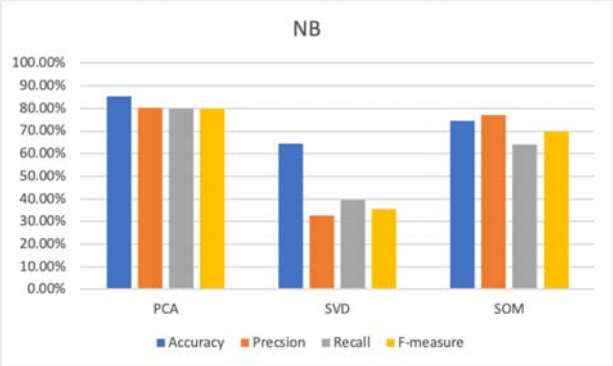

Figure 3 The evaluation measures for NB models with dimensionality reduction techniques

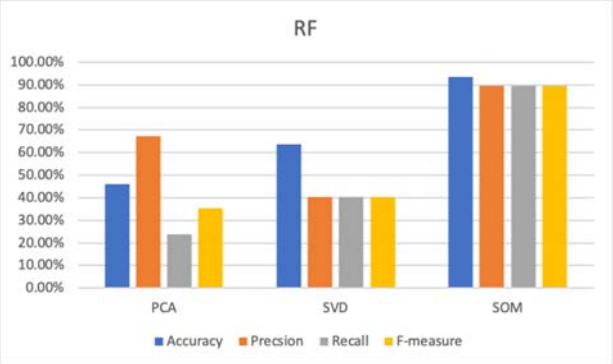

Figure 4 The evaluation measures for RF models with dimensionality reduction techniques.

Second, we consider each combination of DR techniques and classification algorithms. Overall, among the nine classification models trained in this study, the top three models are: PCA+ KNN, SOM+RF, and SOM+KNN. As shown in **Table 1**, the worst-performing models are those that utilize SVD in this dataset.

	Accuracy	Precsion	Recall	F-measure
PCA+KNN	**96.64%**	**93.94%**	**94.60%**	**94.27%**
SOM+RF	93.77%	89.56%	89.56%	89.56%
SOM+KNN	92.14%	90.26%	86.37%	88.28%
PCA+ NB	85.19%	80.32%	79.76%	80.04%
SOM+NB	74.69%	77.00%	64.10%	69.96%
SVD+ NB	64.38%	32.59%	39.39%	35.67%
SVD+RF	63.66%	40.55%	40.39%	40.47%
SVD+KNN	57.72%	37.21%	38.34%	37.77%
PCA+RF	45.97%	67.28%	24.00%	35.38%

The superiority of PCA is consistent with results in the literature on bioinformatics datasets [20]. Specifically, PCA is used for large-scale RNA-sequencing datasets because it avoids the *curse of dimensionality* while preserving the global structure of the dataset [21]. The transformation achieved by PCA can be used to render data more easily explored and visualized. Thus, almost all RNA-Seq research pipelines include this step. Self-organizing maps (SOM) have many benefits over other methods in terms of dimension reduction, multidimensional scaling, and visualization capabilities [22]. Early microarray studies reported the use of SOM, which attracted immediate interest in the field of bioinformatics due to the robustness of this method [23].

4. Conclusion

RNA sequencing high-throughput technologies are delivering unprecedented transcriptome resolution, and it has been particularly useful in exposing the complexity of the transcriptome at the sequence-level. The resulting datasets are complex and high-dimensional, which poses a major challenge for researchers. This study compared the performance of PCA, SVD, and SOM in reducing the dimensionality of the datasets for better machine learning performance. The results show that PCA and SOM outperform SVD. In the future, we will focus on the comparison of various linear versus nonlinear types of dimensionality reduction for use with RNA-Seq datasets.

References

[1] "GenBank and WGS Statistics." https://www.ncbi.nlm.nih.gov/genbank/statistics/ (accessed Jan. 17, 2021).

[2] D. Singh, P. K. Singh, S. Chaudhary, K. Mehla, and S. Kumar, "Chapter Three - Exome sequencing and advances in crop improvement," in *Advances in Genetics*, vol. 79, T. Friedmann, J. C. Dunlap, and S. F. Goodwin, Eds. Academic Press, 2012, pp. 87–121.

[3] A. Jabeen, N. Ahmad, and K. Raza, "Machine learning-based state-of-the-art methods for the classification of RNA-seq data," in *Classification in BioApps: Automation of Decision Making*, N. Dey, A. S. Ashour, and S. Borra, Eds. Cham: Springer International Publishing, 2018, pp. 133–172.

[4] K. Nirmalakumari, H. Rajaguru, and P. Rajkumar, "Performance analysis of classifiers for colon cancer detection from dimensionality reduced microarray gene data," *Int. J. Imaging Syst. Technol.*, vol. 30, no. 4, pp. 1012–1032, 2020, doi: https://doi.org/10.1002/ima.22431.

[5] M. O. Arowolo, M. O. Adebiyi, A. A. Adebiyi, and O. J. Okesola, "A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data," *IEEE Access*, vol. 8, pp. 182422–182430, 2020, doi: 10.1109/ACCESS.2020.3029234.

[6] "Overview and comparative study of dimensionality reduction techniques for high dimensional data - ScienceDirect." https://www-sciencedirect-com.sdl.idm.oclc.org/science/article/pii/S156625351930377X (accessed Jan. 17, 2021).

[7] L. H. Nguyen and S. Holmes, "Ten quick tips for effective dimensionality reduction," *PLOS Comput. Biol.*, vol. 15, no. 6, p. e1006907, Jun. 2019, doi: 10.1371/journal.pcbi.1006907.

[8] "The Cancer Genome Atlas Program - National Cancer Institute," Jun. 13, 2018. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed Jan. 17, 2021).

[9] C. Ferles, Y. Papanikolaou, and K. J. Naidoo, "Denoising Autoencoder Self-Organizing Map (DASOM)," *Neural Netw.*, vol. 105, pp. 112–131, Sep. 2018, doi: 10.1016/j.neunet.2018.04.016.

[10] T. Kohonen, *Self-Organizing Maps*. Springer Science & Business Media, 2012.

[11] T. Ahvenlampi, R. Rantanen, and M. Tervaskanto, "Fault tolerant control application for continuous kraft pulping process," in *Fault Detection, Supervision and Safety of Technical Processes 2006*, H.-Y. Zhang, Ed. Oxford: Elsevier Science Ltd, 2007, pp. 849–854.

[12] D. Miljković, "Brief review of self-organizing maps," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2017, pp. 1061–1066, doi: 10.23919/MIPRO.2017.7973581.

[13] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[14] S. A. Alsenan, I. M. Al-Turaiki, and A. M. Hafez, "Feature extraction methods in quantitative structure–activity relationship modeling: A comparative study," *IEEE Access*, vol. 8, pp. 78737–78752, 2020, doi: 10.1109/ACCESS.2020.2990375.

[15] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numer. Math.*, vol. 14, no. 5, pp. 403–420, Apr. 1970, doi: 10.1007/BF02163027.

[16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 1st edition. San Francisco: Morgan Kaufmann, 2000.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[18] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regressi0on," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.2307/2685209.

[19] "RapidMiner | Best Data Science & Machine Learning Platform," *RapidMiner*. https://rapidminer.com/ (accessed Mar. 15, 2020).

[20] S. Mahapatra, A. Kumar, A. Sharma, and S. S. Sahu, "Effect of dimensionality reduction on classification accuracy for protein–protein interaction prediction," in *Advanced Computing and Intelligent Engineering*, Singapore, 2020, pp. 3–12, doi: 10.1007/978-981-15-1081-6_1.

[21] K. Tsuyuzaki, H. Sato, K. Sato, and I. Nikaido, "Benchmarking principal component analysis for large-scale single-cell RNA-sequencing," *Genome Biol.*, vol. 21, no. 1, p. 9, Jan. 2020, doi: 10.1186/s13059-019-1900-3.

[22] H. Wirth, M. Löffler, M. von Bergen, and H. Binder, "Expression cartography of human tissues using self organizing maps," *Nat. Preced.*, pp. 1–1, Jun. 2011, doi: 10.1038/npre.2011.5825.2.

[23] L. D. Locati *et al.*, "Mining of self-organizing map gene-expression portraits reveals prognostic stratification of HPV-positive head and neck squamous cell carcinoma," *Cancers*, vol. 11, no. 8, Art. no. 8, Aug. 2019, doi: 10.3390/cancers11081057.

Isra AL-Turaiki is an associate professor of Computer Science at King Saud University. She received her PhD in 2014 from the College of Computer Sciences at King Saud University. Her research interests include data mining, machine learning, and bioinformatics.