Machine Learning Techniques and Hybrid Feature Selection Methods for Efficient Prediction of Cancer

Remyamol K M[1], Philip Samuel[2]

[1]Department of Information Technology, School of Engineering, CUSAT, India
E-mail: rems84@gmail.com
[2]Department of Computer Science, CUSAT, India.
E-mail: philipcusat@gmail.com

Abstract

The high-dimensional genomic data presents significant challenges, and traditional analytical methods often struggle to capture the complex, non-linear relationships within these datasets. This study elaborates into the application of machine learning methods for dimensionality reduction and predictive modeling of binary phenotypes using gene expression data. Various dimensionality reduction techniques are explored, including t-distributed stochastic neighbor embedding (t-SNE), Non-negative matrix factorization (NMF), Principal component analysis (PCA), and manifold learning methods. Additionally, various algorithms such as logistic regression, random forests, support vector machines (SVMs)cand naive Bayes models are evaluated for predicting phenotypes. The study employs rigorous cross-validation, permutation testing, and evaluation metrics like the Matthews Correlation Coefficient (MCC) to assess model performance. The study rigorously assesses current genomics strategies, pinpointing their drawbacks and suggesting areas for future investigation, while delving into the potential of machine learning to overcome these hurdles and offer valuable insights in genomics.

Keywords- Gene Expression , Dimensionality Reduction, Biomarkers, Cancer Prediction, Epigenetics, Machine Learning, Feature extraction.

1. INTRODUCTION

Genomics, the study of an organism's complete set of genetic instructions, has experienced an unprecedented expansion by the technological advances in genomic high-throughput sequencing [1] [2]. The rapid pace of innovation in genomic technologies has enabled researchers to generate and analyze genomic data at an unprecedented scale, fueling discoveries that were once thought impossible. However, analyzing and interpreting the complex and high-dimensional nature of these genomic datasets presents significant challenges that traditional analytical methods often struggle to overcome. Conventional approaches, which often rely on linear models and simplifying assumptions, fail to capture the non-linear relationships and intricate patterns inherent within these datasets, leading to suboptimal results and potentially missing critical insights [3] [4].

Machine learning techniques, with their ability to learn non-linear representations and uncover complex patterns from data, have emerged as powerful tools for addressing these challenges. By leveraging machine learning algorithms, researchers can effectively handle large-scale genomic datasets, identify relationships between genes, and unravel the intricate regulatory mechanisms governing biological processes [5] [6]. Through the utilization of machine learning methods, we can discover hidden patterns unlocking new avenues for understanding and treating complex diseases [7] [8]. The application of machine learning in genomics has already shown promising results, enabling researchers to gain deeper insights into the intricate workings of the genome and its role in health and disease.

The crucial limitations in the interpretation and proper analysis of genomic data is the curse of dimensionality, where the number of features (e.g., gene expressions, genetic variants) far exceeds the number of samples. This high-dimensional nature of genomic data can pose significant challenges for traditional statistical methods, as they often struggle to handle such vast numbers of variables effectively. Dimensionality reduction methods are essential for addressing this issue, as they reduce noise, eliminate irrelevant features, and expose the most informative variables for downstream tasks such as clustering, classification, and biomarker discovery [9] [10]. By reducing the dimensionality of the data, these techniques can improve the interpretability and computational efficiency of subsequent analyses, while preserving the informations that are most relevant.

Here, we investigate the applications of machine learning methods for dimensionality reduction and predictive modeling of binary phenotypes using gene expression data. Specifically, we evaluate the performance of logistic regression, support vector machines (SVMs), random forests, and naive Bayes models for predicting phenotypes and assess their performance using rigorous cross-validation, permutation testing, and evaluation metrics such as the Matthews Correlation Coefficient (MCC) [11]. These machine learning models offer diverse approaches to handling high-dimensional data and capturing non-linear relationships, making them well-suited for genomic applications. Through this comprehensive analysis, the limitations that machine learning techniques can address is

identified, providing insights into the potential of these techniques to revolutionize our understanding of genomic data and its applications in disease diagnosis, treatment, and precision medicine.

2. MACHINE LEARNING TECHNIQUES FOR FEATURE REDUCTION

Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) are two widely used linear dimensionality reduction techniques in genomics data analysis. PCA is a classical method that aims to project the lower-dimensional subspace from high-dimensional data in order to preserve the maximum variance in the data. It identifies the principal components, which are the directions of maximum variance, and represents the data in terms of these components. PCA has been extensively applied in genomics for tasks such as visualizing high-dimensional gene expression data, identifying patterns and sources of variation, and reducing dimensionality as a pre-processing step for downstream analyses [12].

NMF, on the other hand, is a matrix factorization technique that decomposes the high-dimensional data into a product of two non-negative matrices, representing the basis vectors and the corresponding coefficients. Unlike PCA, which finds orthogonal components, NMF imposes a non-negativity constraint on the factors, which can lead to more interpretable and biologically meaningful representations. NMF has been widely used in genomics for tasks such as identifying gene modules, discovering biological processes, and extracting metagenes or metagene expression patterns from gene expression data [13].

Both PCA and NMF have their strengths and limitations. PCA is a well-established and computationally efficient method, but it can be sensitive to scaling and may not always capture the underlying biological structure of the data. NMF, on the other hand, can provide more interpretable and biologically relevant representations, but it may suffer from issues such as non-unique solutions and potential instability.

The choice between PCA and NMF, or the use of other linear dimensionality reduction techniques, often depends on the specific goals and characteristics of the genomic data analysis task at hand. In many cases, a combination of different techniques or the integration of dimensionality reduction with other methods, such as feature selection or non-linear techniques, can lead to improved performance and more comprehensive insights into the underlying biological processes.

Nonlinear dimensionality reduction techniques like Principal Component Analysis (PCA), Negative Matrix Factorization (NMF) t-Distributed Stochastic Neighbor Embedding (t-SNE) and Non- and manifold learning methods have gained significant traction in genomics data analysis due to their ability to handle complex, nonlinear relationships present in high-dimensional datasets. t-SNE is a powerful method that maps lower-dimensional data from high-dimensional data into a while preserving the local structure of the data. It has been extensively used for visualizing high-dimensional genomic data, such as gene expression data, by revealing underlying patterns, clusters, and potential subgroups. t-SNE has proven invaluable in exploratory data analysis and hypothesis generation, enabling researchers to gain insights into the intricate relationships and structures within their data [13].

Manifold learning techniques are based on the assumption that high-dimensional data lies on a lower-dimensional space. These methods aim to identify and preserve the intrinsic geometric structures present in the data while reducing dimensionality. By leveraging the underlying manifold structure, these techniques can effectively capture nonlinear relationships and uncover meaningful patterns that may be obscured in the high-dimensional space. Manifold learning methods have been applied to various tasks in genomics, including gene expression data analysis, protein structure prediction, and integrating multi-omics data [15].

Complementing these nonlinear dimensionality reduction techniques, sparse coding and dictionary learning methods offer alternative approaches to representing and analyzing high-dimensional genomic data. Sparse coding algorithms represent the data as a linear combination of a few basis vectors, known as a sparse representation. This sparse representation can be used for feature extraction and dimensionality reduction, capturing the most relevant and informative aspects of the data while reducing noise and redundancy. Dictionary learning methods take this concept a step further by learning a set of basis vectors, or a dictionary, that can represent the data as a sparse linear combination. These techniques have been applied to capture sophisticated patterns in genomic data while reducing dimensionality, enabling more efficient and effective downstream analyses. While nonlinear dimensionality reduction techniques excel at preserving local structures and capturing complex relationships, sparse coding and dictionary learning methods provide a complementary approach by representing the data in a sparse and interpretable manner. The specific characteristics of the data determines the choice of technique, the desired properties of the reduced representation, and the downstream analysis tasks. In many cases, a combination of multiple dimensionality reduction approaches may be beneficial, leveraging the strengths of different techniques to gain a more comprehensive understanding of the underlying biological processes and patterns within the high-dimensional genomic data.

3. PREDICTIVE MODELING AND EVALUATION

For predictive modeling of binary phenotypes, several powerful machine learning algorithms were employed. Logistic regression modeled the probability of belonging to a class as a function of input features. Support vector machines (SVMs) found the optimal hyperplane separating classes with maximum margin, effectively leveraging the dataset geometry. Random forests, an ensemble approach combining multiple decision trees, provided robust and accurate predictions. The naive Bayes classifier estimated conditional probabilities of features given class labels, assuming feature independence. The Matthews correlation coefficient (MCC), suitable for binary tasks and robust to class imbalance, served as the primary evaluation metric. Furthermore, a permutation testing framework empirically determined the statistical significance of the observed predictive performance by randomly permuting phenotype labels, generating a null distribution of scores, and

assessing the true association between features and phenotypes against this null.

Logistic regression estimates coefficients that represent the linear combination of features, which is then passed through the sigmoid to obtain predicted probabilities between 0 and 1. The coefficients are learned from training data via maximum likelihood estimation to maximize the likelihood of observing the actual class labels [8]. Logistic regression offers several advantages like interpretable coefficients indicating the relative importance and influence of each feature, probabilistic outputs easily interpretable for binary tasks, ability to handle continuous and categorical features, and compatibility with regularization techniques like L1/L2 to prevent overfitting. However, a linear relationship between features of the outcome is assumed and it can be sensitive to outliers and multicollinearity. The Logistic regression is a, widely-used technique, despite of these limitations, for binary classification problems across various domains due to its effectiveness, interpretability, and flexibility in handling different types of input data [9].

Support Vector Machines (SVMs) are powerful supervised learning algorithms that excel at binary classification tasks. The core idea behind SVMs is to identify the two classes in the feature space and find a hyperplane that maximally separates it . This optimal hyperplane is the one that has the maximum margin, which is the distance between the nearest data points from each class to the hyperplane. In this higher-dimensional space, the data is more likely to be linearly separable, allowing for the construction of an optimal hyperplane that separates the classes with the maximum margin. Support Vector Machines (SVMs) are powerful classifiers that find the optimal hyperplane separating classes with maximum margin in the feature space, mapped using kernel functions, achieving good generalization and robustness. The support vectors define the decision boundary, and SVMs maximize the margin while minimizing misclassification error through optimization, incorporating regularization. SVMs handle non-linear decision boundaries via the kernel trick, result in sparse solutions for computational efficiency, and have been extensively used in genomics for tasks like gene expression analysis, protein structure prediction, and disease risk prediction due to their ability to handle high-dimensional data, robustness to noise, and capturing complex non-linear relationships. They are particularly well-suited for classification tasks, including binary classification problems prevalent in genomics[12] [13].

The core idea behind Random Forests is to construct a collection of decision trees, by training random subset of the input features and training data. This randomization process helps to introduce diversity among the individual trees, reducing the risk of overfitting and improving the overall generalization performance of the ensemble. During the training phase, each decision tree in the ensemble is grown using a subset of the training data. This involves randomly sampling the original training data with replacement to create multiple bootstrap samples. Each bootstrap sample is then used to grow a separate decision tree, where at each node, a random subset of the input features is considered for splitting. When making predictions on new instances, the predictions of all the individual decision trees in the ensemble are aggregated by the Random Forest algorithm. The class with the more vote among the trees is assigned as the final prediction for the classification [14]. Random Forests have been successfully applied to various classification problems in genomics, such as gene expression analysis, disease risk prediction, and genomic biomarker discovery. Their ability to handle high-dimensional data, robustness to noise and outliers, and capability to capture complex non-linear relationships make them a powerful tool for analyzing genomic data. Random Forests combines multiple decision trees, by training random subsets of data and features through bagging, introducing diversity and reducing overfitting and it is an ensemble learning method. Predictions are made by aggregating tree votes, capturing complex patterns robustly and accurately. Random Forests offer feature importance estimation, parallelization for efficiency, handling of mixed data types, and robustness to overfitting high-dimensional data like genomics. They have been successfully applied to genomics classification tasks like gene expression analysis, disease risk prediction, and biomarker discovery due to their ability to handle complex, non-linear relationships in high-dimensional noisy data [10].

Naive Bayes classifiers are a family of simple yet effective probabilistic classifiers based on the Bayes' theorem. They are particularly well-suited for binary classification tasks and have been widely used in various domains, including genomics. In genomics, Naive Bayes classifiers have been successfully applied to tasks such as gene expression analysis, protein function prediction, and disease risk assessment. Their simplicity, efficiency, and ability to handle high-dimensional data make them a valuable tool in the analysis of genomic data, particularly when the assumption of feature independence is reasonable or when the primary goal is to obtain robust and interpretable predictions[11].

4. LIMITATIONS AND FUTURE DIRECTIONS

Feature selection methods identify the most informative subset of genomic features (e.g., gene expressions, genetic variants) to improve model performance, interpretability, and efficiency. Ensemble techniques like bagging, boosting, and stacking combine multiple models to capture diverse aspects of high-dimensional, complex genomic data robustly. Integrating multi-omics data (genomics, transcriptomics, epigenomics, proteomics) through approaches like multi-view learning, multi-kernel methods, and graph-based integration provides a comprehensive understanding of biological processes [13] [14]. Interpretability methods like feature importance analysis, saliency maps, and interpretable models elucidate biologically relevant patterns learned by machine learning models. Transfer learning and domain adaptation leverage knowledge from data-rich source domains to improve model performance in data-scarce target domains, addressing the challenge of limited annotated genomic datasets.

4. CONCLUSION

The application of machine learning techniques in genomics has demonstrated significant potential for addressing the challenges posed by high-dimensional and complex datasets. This study explored various dimensionality reduction and predictive modeling approaches for analyzing gene expression data and predicting binary phenotypes. Feature selection methods identify the most informative subset of genomic features, improving model performance, interpretability, and efficiency. Ensemble techniques like bagging, boosting, and stacking combine multiple models to capture diverse aspects of high-dimensional, complex genomic data robustly. Integrating multi-omics data (genomics, transcriptomics, epigenomics, proteomics) through approaches like multi-view learning, multi-kernel methods, and graph-based integration provides a comprehensive understanding of biological processes. Interpretability methods, such as feature importance analysis, saliency maps, and interpretable models, elucidate biologically relevant patterns learned by machine learning models. Transfer learning and domain adaptation leverage knowledge from data-rich source domains to improve model performance in data-scarce target domains, addressing the challenge of limited annotated genomic datasets. While the study presents promising results and insights, further research is required to address the limitations and explore new avenues in the field of genomics. Continued advancements in machine learning techniques, coupled with the integration of multi-omics data and improved interpretability, hold the potential to unlock deeper insights into biological systems and pave the way for personalized medicine and targeted therapeutic interventions.

REFERENCES

[1] I. C. Kim et al., "Iterative recursive dimension reduction: a technique for dimension reduction in machine learning," in Proc. IEEE Int. Conf. Data Min., 2017, pp. 321-330.

[2] K. M. Ting, “Confusion Matrix BT - Encyclopedia of Machine Learning and Data Mining,” C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2017.

[3] Abeer A. Raweh et al., “A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation” Digital Object Identifier 10.1109/ ACCESS .2018.2812734, 2018.

[4] Y. Pan, T. Liu, A. H. Aladailati, E. K. Dey, and S. Zhong, "Dimensionality Reduction of Single-Cell RNA-Seq Data Using Deep Generative Models," in IEEE Transactions on Computational Biology and Bioinformatics, Vol. 19, pp. 873-885, 2022.

[5] H.M. Mohamad, M.A.M. Abushariah, A.Y. Amin, "Gene expression-based disease classification using ensemble machine learning methods," Computers in Biology and Medicine, vol. 137, 104772, 2021.

[6] M. D. M. Rahman, C. Haoyuan, C. C. Armstrong, and L. Xiaoming, "Machine learning techniques for gene expression profile analysis of multiple cancer types," in Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI), 2018, pp. 251-254.

[7] J. Li, L. Liu, J. Liu, and R. Green, “Building Diversified Multiple Trees for classification in high dimensional noisy biomedical data,” Heal. Inf. Sci. Syst., vol. 5, no. 1, p. 5, 2017.

[8] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proceedings of International Conference on Machine Learning, pp. 161-168, 2006.

[9] S. Hosseini and B. Zohidian, "A Novel Dimensionality Reduction Method for Clustering Gene Expression Data," in IEEE Access, vol. 8, pp.64436-64447, 2020.

[10] M.S. Abirami, P. Vijayalakshmi, R. Lakshmi, "Disease prediction using DNA sequence and gene expression data with ensemble learning methods," Journal of King Saud University - Computer and Information Sciences, 2022.

[11] I. C. Kim et al., "Iterative recursive dimension reduction: a technique for dimension reduction in machine learning," in Proc. IEEE Int. Conf. Data Min., 2017.

[12] Sajid Shah, et al., “DNA Methylation Prediction Using Reduced Features Obtained via Gappy Pair Kernel and Partial Least Square”, 2022.

[13] C. H. Park and S. B. Kim, “Sequential random k-nearest neighbor feature selection for high-dimensional data,” Expert Syst. Appl., vol. 42, no. 5, pp. 2336–2342, 2015.

[14] M. Matsui, S. Corey, B. Ou, and B. Boman, "Microarray technology and its application to toxicogenomics," in Molecular Toxicology Protocols, 2nd ed., P. Keohavong and S. G. Grant, Eds. Totowa, NJ, USA: Humana Press, 2008, pp. 333-349