Feature Selection For High Dimensional Data Using Supervised Machine Learning Techniques*

Lazaros K.
Dept. of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
lakonstant@uth.gr

Vrahatis A.
Dept. of Informatics
Ionian University
Corfu, Greece
aris.vrahatis@ionio.gr

Tasoulis S.
Dept. of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
stasoulis@uth.gr

Plagianakos V.
Dept. of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
vpp@uth.gr

***Abstract*—In recent years, feature selection has become an increasingly active field of data science and machine learning research. Most of the datasets that are being used nowadays for various machine learning tasks consist of thousands of features (columns), which make them extremely complex and difficult to work with. In this paper, we propose a feature selection methodological pipeline that can be used to reduce the complexity of high dimensional datasets through the elimination of redundant and/or non-informative features as well as to improve the performance of machine learning models which are trained on high dimensional datasets. The proposed method has been applied to high-dimensional biomedical data and compared against a classic filter-based feature selection algorithm. Specifically, the method was applied to gene expression profiles of a single-cell RNA-seq dataset from healthy and infected by covid-19 human samples.**

I. Introduction

scRNA-seq is a next-generation sequencing technique that is used for transcriptomic analysis of individual cells within a cell population. It contributes to a better understanding of which genes are being expressed in an individual cell, to what extent they are being expressed, and how their expression differs from other individual cells within a cell population. In other words, this technique reveals previously unappreciated levels of heterogeneity among cell populations that at first glance seem to be homogeneous. Heterogeneity is a term that is used to describe the differences between cells in terms of function/behavior and by extension differences in terms of gene expression.

Through scRNA-seq, scientists can compare the transcriptomes of individual cells. That way, both similarities and differences between cells can be highlighted. By analyzing gene expression at the level of a single/individual cell, it becomes easier to identify extremely rare cell populations that would have probably gone unnoticed by the more classic bulk RNA-seq techniques (i.e; identification of malignant tumor cells and/or identification of highly specialized cells).

Since thousands of genes are being analyzed for each cell, the datasets that occur from scRNA-seq experiments are (in most cases) high-dimensional. It is also worth noting that scRNA-seq datasets are quite sparse, owing to a large number of dropout events (failure of detection of RNA molecules).

The term "High Dimensional Data" is used to describe datasets, in which the number of features (columns), exceeds the number of available observations (rows). This can be summed up by the expression $f >> s$ where f represents the number of features and s represents the number of samples/observations within the data high-dimensional number of features within a dataset, the harder it is for machine learning models that have been trained on the said dataset to make accurate predictions. That is because each new feature that is added to the dataset leads to an exponential decrease in the prediction power of the model (Hughes phenomenon/Curse of Dimensionality).

Feature selection is necessary for identifying and extracting a subset of features from a dataset which can be used to build useful and accurate models of the phenomena being studied. It is very effective at strengthening the performance of machine learning models in knowledge fields that are characterized by datasets with an enormous amount of features (i.e: medicine, biology, image processing, etc). It is also worth noting that aside from assisting in the simplification of datasets, feature selection can also provide a better understanding of the data (and by extension the problem) being used, as it assists in highlighting which features are directly associated with the target variable. Feature selection techniques generally fall into two broad categories: filter and wrapper-based methods.

Filter-based feature selection methods are usually applied during data preprocessing. They are independent of any machine learning algorithm. Features are selected based on the score they receive in specific statistical tests which measure the correlation of each feature to the target variable. On the other hand, wrapper-based methods are dependent on a

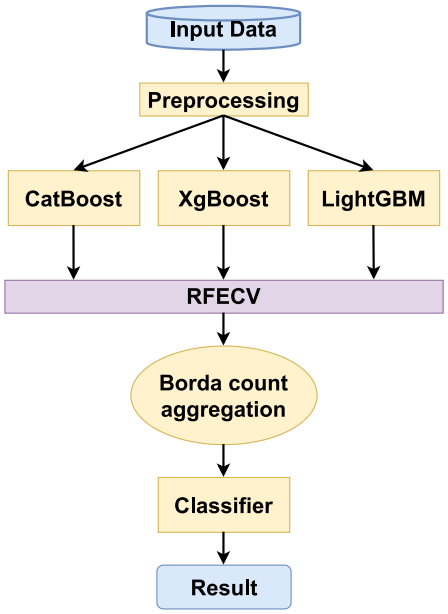

Fig. 1: Proposed methodology workflow

machine learning algorithm. A subset of features is used to train a machine learning model used depending on training performance, features are either added to or removed from the feature subset.

II. Related work

Over the years, many different feature selection methods have been proposed for dealing with scRNA-seq datasets. For instance, Feng et al. have proposed scTIM [20]. scTIM is a feature selection framework for scRNA-seq data. It utilizes a multi-objective optimization technique that maximizes gene specificity by taking into account relations between cells and genes as well as the potential of each gene to highlight relations between individual cells. The framework has also been designed to eliminate redundant genes. For all of the aforementioned tasks, scTIM creates objective functions which have to be optimized. The solution of each objective function is combined into one final solution. M3Drop [21] is an R package that includes well-established feature selection algorithms as well as two novel methods that utilize the null expression values of some genes (dropout events) to identify important genes. Chatzilygeroudis et al. [22] propose a feature selection and analysis framework which utilizes genetic optimization principles (genetic algorithm), to locate low-dimensional gene lists. A simple distance-based classifier is utilized for finding features that are well separated in Euclidean space. Lall et al propose sc-REnf [25], a feature selection framework that makes use of R'enyi and Tsallis entropies to identify informative genes within high-dimensional scRNA-seq datasets. sc-REnf also makes use of a very robust objective function which enables it to identify redundant features within noisy data. Last but not least, Li et al. [26], propose a feature selection method which is based on an intrinsic entropy formula in order to indentify and select important genes in high dimensional scRNA-seq datasets. They show that intrinsic entropy should be considered information instead of noise since genes with high I.E are highly informative for cell type and cell state analysis.

III. Methodology

Through this paper, we propose a feature selection methodological pipeline that can be used to deal with high-dimensionality scRNA-data. The steps of the proposed methodology are as follows:

- Data preprocessing; both the input data and the target variable are processed before any further use.
- 10-fold cross-validation is performed to evaluate the performance of 4 well-established machine learning models on the original dataset which consists of all the available features. These models are random forest, support vector machine, logistic regression, and k nearest neighbors.
- Feature selection is performed through the use of the rfecv wrapper-based algorithm. The algorithm is executed for three gradient-boosting tree-based models. Gradient boosting models were chosen as the basis for the wrapper-based feature selection algorithm due to their proven track record of dealing with high-dimensional data. The ones we chose for the proposed methodology are XgBoost, CatBoost, and LightGBM. Through this procedure three lists are obtained, each of which contains only the most important features of the dataset according to the gradient boosting model which was used.
- The three lists which were obtained from the feature selection procedure are combined through the use of a borda-based aggregation method into a "consensus" list which contains every feature that is considered important by the three chosen gradient boosting models. The

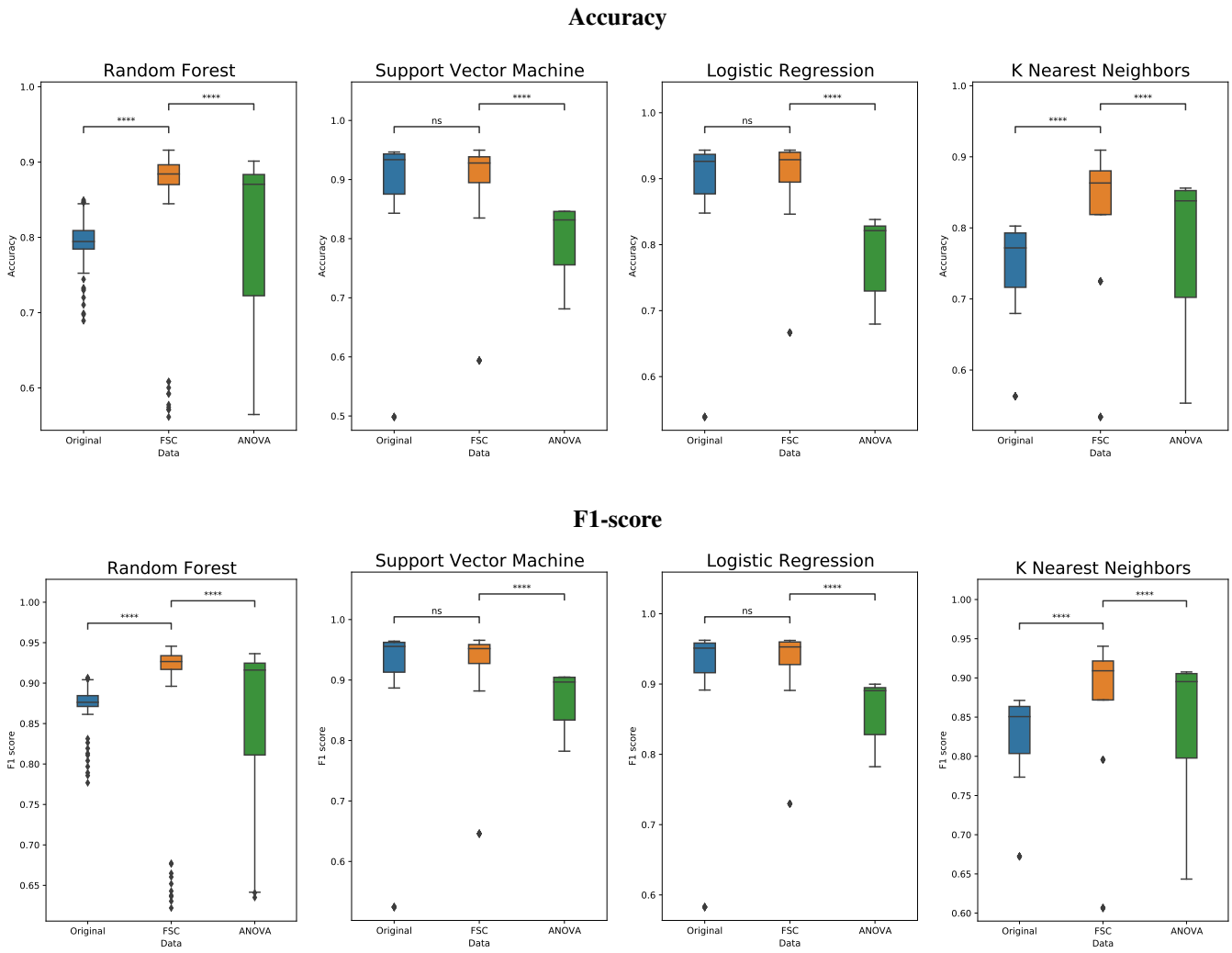

Fig. 2: Evaluation of selected features using different classifiers (10 runs of 10-fold cross-validation). The number of stars indicates that the p-value of the Mann-Whitney U test is less than 0.05, 0.01, 0.001 and 0.0001 respectively.

rationale behind the aggregation method is to enhance the robustness of the feature selection procedure and to minimize the bias that would occur because of the nature of wrapper-based feature selection.

- The performance of the 4 previously mentioned established models is assessed for the consensus/feature-reduced dataset through the use of 10-fold cross-validation.
- The results are compared.

IV. Experimental Analysis and Results

The dataset was obtained from the Gene Expression Atlas database and it consists of 6178 rows and 18958 columns. The target variable is categorical and binary, its two values being; ”covid-19” and ”normal”. 1651 samples belong to the category ”normal” and 4527 samples belong to the category ”covid-19”.

It should be noted that the dataset we used for evaluating the proposed method, is in .mtx format in its original configuration. mtx is a widely used file format for storing scRNA-seq data. In most cases, in a .mtx file rows represent genes and columns represent individual cells. We implemented our method using python and jupyter notebooks.

Due to the nature of .mtx files, the dataset had to be transposed before being converted to a pandas data frame. Min-max normalization was used on the input data to reduce the standard deviation of the samples and by extension the effect of outliers.

Because of limitations in computing power, we chose to shave off columns for which a percentage of rows have zero values. This was done so that the feature selection procedure can be completed in a reasonable amount of time. The percentage we chose was 9%. This optional step should be bypassed in cases where computing power is not an issue. Through this early dimensionality reduction step, the features of the dataset have been reduced from 18958 to 6042. This reduced version of the dataset (6042) was used for feature

selection. The dataset has been further reduced from 6042 features down to 762 by use of the feature selection and borda aggregation consensus procedures.

The values of the target variable have also been enumerated with 1 being the value for "covid-19" and 0 being the value for that". Both the processed input data and target variable have been stored in .csv files. Cross-validation has been performed using sklearn's cross_validate function which provides the ability to assess the performance of machine learning models through k-fold cross-validation using several evaluation metrics and not just one. The metrics we elected to use to evaluate the performance of the chosen machine learning models were accuracy and the F1 score.

To verify the effectiveness of the proposed feature selection method, we compare it to a well-established filter-based method; the ANOVA test. Sklearn's SelectKBest algorithm is run on the original dataset which includes every feature. Through SelectKBest an ANOVA statistical test is performed so that only the K most important features remain in the dataset. In our case, the 10 most important features have been selected by SelectKBest. Subsequently, cross-validation is performed for the 4 well-established machine learning models which were previously mentioned to compare their performance to that obtained from the proposed methodology. The performance of the 4 machine learning models we used for all 10 iterations of 10-fold cross-validation, on the full feature dataset (18958 features/genes), on the "consensus"/feature selected dataset (762 features/genes) and on the ANOVA reduced (10 features/genes) dataset is displayed in the box plots above;

V. Conclusions

The machine learning models we made use of to evaluate the effectiveness of the proposed method, performed better on the "consensus" dataset (762 features/genes) than on the complete dataset (18958 features/genes).

The results obtained from the proposed methodology are better than those obtained using the SelectKBest filter-based feature selection algorithm. This may be because we didn't initialize the algorithm with a specific value for K. Instead, we elected to use its default value of 10.

Our method, eliminated redundant/non-informative features from a high dimensional scRNA-seq dataset while also leading to performance improvements for the simpler of the chosen machine learning models. Additional testing will be carried out in the near future to rule out the possibility of satisfactory performance due to bias.

References

[1] Aurelien Geron. Hands-on Machine learning with Scikit-Lean, Keras TensorFlow. O'REILLY, 2nd edition, 2019.

[2] Andrew Glassner. Deep learning: A visual approach. No Starch Press, 1st edition, 2021

[3] Kjell Johnson Max Kuhn. Applied predictive modeling. Springer, 2nd edition, 2018.

[4] Weston J. Barnhill S. Vapnik V. Guyon, I. Gene selection for cancer classification using support vector machines. Machine Learning, 117:389–422, January 2002.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30:3146–3154, 2017

[7] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, NeurIPS, pages 6639–6649, 2018.

[8] Tin Kam Ho. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE, 1995.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[10] Mark Gerstein Zhong Wang and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. Nature Reviegradient-boosting):57–63, 2009.

[11] Papatheodorou I, Fonseca NA, Keays M, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018;46(D1):D246-D251. doi:10.1093/nar/gkx1158.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189–1232, 2000

[13] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-Validation, pages 532–538. Springer US, Boston, MA, 2009

[14] Xinchuan Zeng and Tony Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. Journal of Experimental Theoretical Artificial Intelligence, 12, 04 2001.

[15] Puneet Misra and Arun Singh. Improving the classification accuracy using recursive feature elimination with cross-validation. 11:659–665, 05 2020.

[16] Ashraful Haque, Jessica A Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. Genome Medicine, 9, 2017.

[17] Merijn Erp and Lambert Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. 01 2000

[18] Chen Qi, Zhaopeng Meng, Xinyi Liu, Qiangguo Jin, and Ran Su. Decision variants for the automatic determination of optimal feature subset in rf-rfe. Genes, 9:301, 06 2018

[19] Omer Sagi and Lior Rokach. Approximating xgboost with an interpretable decision tree. Information Sciences, 572:522–542, 2021

[20]] Zhanying Feng, Xianwen Ren, Yuan Fang, Yining Yin, Chutian Huang, Yimin Zhao, and Yong Wang. scTIM: seeking cell-type-indicative marker from single cell RNA-seq data by consensus optimization. Bioinformatics, 36(8):2474–2485, 12 2019.

[21] Tallulah S Andrews and Martin Hemberg. M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics, 35(16):2865–2867, 12 2018.

[22] Konstantinos I. Chatzilygeroudis, Aristidis G. Vrahatis, Sotiris K. Tasoulis, and Michael N. Vrahatis. Feature selection in single-cell rna-seq data via a genetic algorithm. In Dimitris E. Simos, Panos M. Pardalos, and Ilias S. Kotsireas, editors, Learning and Intelligent Optimization, pages 66–79, Cham, 2021. Springer International Publishing.

[23] Padraig Cunningham and Sarah Delany. k-nearest neighbour classifiers. Mult Classif Syst, 54, 04 2007. 64

[24] Silvin A, Chapuis N, Dunsmore G, et al. Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. Cell. 2020 Sep;182(6):1401-1418.e18. DOI: 10.1016/j.cell.2020.08.002. PMID: 32810439; PMCID: PMC7405878.

[25] Snehalika Lall, Abhik Ghosh, Sumanta Ray, Sanghamitra Bandyopadhyay, sc-REnF: An entropy guided robust feature selection for single-cell RNA-seq data, Briefings in Bioinformatics, Volume 23, Issue 2, March 2022, bbab517.

[26] Lin Li, Hui Tang, Rui Xia, Hao Dai, Rui Liu, Luonan Chen, Intrinsic entropy model for feature selection of scRNA-seq data, Journal of Molecular Cell Biology, Volume 14, Issue 2, February 2022, mjac008.