APPLYING THE BIG DATA ANALYTICS IN BIOINFORMATICS - A PRACTICAL EXAMPLE

Nenad Mitić[1], Aleksandar Veljković[1], Mirjana Maljković[1], Saša Malkov[1], Minjie Lyu[2], Xin Lin[2], Marek Michalewicz[3], Guanglan Zhang[4], Vladimir Brusić[2]

[1] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
e-mail: {nenad, aleksandar,mirjana,smalkov}@matf.bg.ac.rs
[2] University of Nottingham, Ningbo, China
e-mail: {minjie.lyu, xin.lin, vladimir.brusic}@nottingham.edu.cn
[3] University of Warsaw, ICM, Warsaw, Poland
e-mail: marek.michalewicz@icm.edu.pl
[4] Boston University, Metropolitan College, HiLab, Boston, United States
e-mail: guanglan@bu.edu

Abstract:

Transcriptomics is the study of all transcripts in a cell in response to biological changes. Single cell transcriptomics concurrently measures changes of gene expression in individual cells in biological material. Due to the large number of cells and high dimensionality of features (genes), obtaining the transcriptomic's results requires sofisticated tools and techniques. We performed a supervised classification, a Big Data analytics technique, to predict peripheral blood mononuclear cell (PBMC) types. The material used in prediction included more than 124,000 cells, where each had a partial measurement of 30698 gene expression values. Using supervised machine learning classification algorithms, we built prediction models to compare their ability to classify cell type within peripheral blood mononuclear cell samples. All methods showed classification accuracy between 95% and 99%. The results obtained confirm feasibility of using Big Data analytics classification techniques for characterization of PBMC cell types.

Keywords: Big Data Analytics, Classification, Supervised Machine Learning, PBMC

1. Introduction

The emergence of bioinformatics as a discipline was crucially influenced by the accelerated development of computing. This era brought forward the intensive development of software components, increased computational performance, and the decreasing prices of computer systems. Computer systems' capabilities keep surpassing the performance of supercomputers of previous decades, making ever-increasing computational power available to the wider scientific community. The development of computer systems has enabled the collection and storage of large amounts of data, their processing and the analysis of results. The use of Big Data Analytics / Machine Learning Techniques also contributed to the increase of the efficiency of data processing and analysis, which was performed manually in the previous period. These methods are mathematically based and give results with high reliability, enabling the processing and analysis of large amounts of data in areas where such research was not possible before. The results of bioinformatics projects involving Big Data Analytics techniques have significantly contributed to the efficiencies in various fields, especially in medicine and pharmacology. One such example of the application of data mining/machine learning techniques to recognize *peripheral blood mononuclear cells* (PBMC) is described in this article. PBMC are a set of specialized cells that are essential parts

of the immune system. PBMCs play complex roles in initiating and launching immune responses against various pathogens, cancer, and toxins [1]. There are five main PBMC types: B cells (BC), T cells (TC), Natural Killer cells (NK), Dendritic cells (DC) and Monocytes (MC) [2].

Gene expression values in PBMC were obtained using *Single Cell Transcriptomics* (SCT) technology. SCT can provide expression values of tens thousands of genes from thousands of individual cells in a very short time [3, 4]. Different cell types display patterns of gene expression that are characteristic for cell types and subtypes. Data obtained from SCT can be produced in a stream-like manner, efficiently and reproducibly. Accurate classifications can be done using Big Data Analytics / Machine Learning methods [3]. We present a method for prediction of PBMC cell types and compare the performance of several supervised classification techniques.

2. Material and methods

2.1 Data

We used 138 SCT datasets from four sources: Broad Institute Single Cell Portal (http://singlecell.broadinstitute.org/single_cell) - BroadS1 (BS1) and BroadS2 (BS2) datasets, 10x Genomics (http://support.10xgenomics.com/single-cell-gene-expression) - 10xG dataset, and a selection of data sets from NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo) - GEO dataset.

The initial set of data was preprocessed as different datasets contained the expression data of the standardized lists of genes. Data used in classification includes more than 124,000 cells (16,405 from BroadS1; 12,146 from BroadS2; 13,183 from GEO, and 82,428 from 10x). Each cell had expression values of 30,698 genes in the form of sparse matrices. The data sets represent incomplete gene expression measurements, where a gene may be expressed in the cell but may be missed in measurements.

2.2 Classification methods

The aim of classification was to construct prediction models of PBMC cell types that generalize well, that is, predict cell types from unknown biological material with high accuracy. Models were constructed using four separate sets of cells by source (BS1, BS2, 10x, and GEO). Each set of cells was divided in training, test and validation partitions in the relation 50:30:20. Models were built on training partitions while test and validation were used for accuracy testing. In addition, constructed models (trained on one data set) were tested for classification accuracy on other three independent sets of cells. For example, the model built using BS1 data for training was tested with BS2, 10x and GEO data.

To build a prediction model, the input cells were represented as 30,698-dimensional vectors, where each gene expression value represents one dimension of the data vector. The aim of the research was to discover/construct the best prediction model of PBMC cell types. We used two software packages with 22 classification algorithms or algorithm variants: IBM SPSS Modeler [5] - 16 algorithms, and Python (Scikit-learn library) - 6 algorithms. The list of algorithms used in this study is presented in Table [1]. Because not all algorithms can successfully process data with a large number of dimensions, we performed dimension reductions for the selection of shortened lists of genes most suitable for classification. After applying the dimensionality reduction, 11 sets with different number of genes (lists with 15, 31, 41, 45, 48, 64, 78, 10800 and 30698 genes, and two lists with 23 genes) were selected.

IBM SPSS Modeler algorithms	Python (Scikit-learn) algorithms
C5.0	Gradient Boosting
CART	k-Nearest Neighbors
CHAID	Multinomial Naive Bayes Classifier
QUEST	Neural Network
Random Trees	Random forest
SVM - Polynomial Kernel	Support Vector Machine (RBF kernel)
SVM - RBF Kernel	
SVM - Sigmoid Kernel	
SVM - Linear Kernel	
ANN - Multilayer Perceptron (MLP) Model	
ANN - Radial Basis Function (RBF) Model	
Random Forest	
Bayes Network - TAN (Tree Augmented Naive Bayes Model)	
XGBoost-AS (Spark implementation)	
XGBoost Tree (Python implementation)	
XGBoost Linear (Python implementation)	

Table 1. List of algorithms used in this study

For each list of genes we performed 5 random sampling procedures to obtain training, test and validation partitions for internal cross-validation analysis. For this part of the study, we constructed more than 400 models that were applied on more than 2000 sets of data partitions.

The construction of the model using personal computers and single-cpu servers did not always go smoothly. For sets with a larger number of genes (10800 and 30698) some models needed more than 128GB of real memory and more than two weeks to complete. A small number of models could not be constructed within one month. Although most classification algorithms do not have the possibility of parallelization, the use of high-performance computers would improve performance and increase the number of algorithms that can be considered.

3. Results

The average accuracy of applied methods was between 91.7% and 99.8%. Models built on one cell type (on training partition) applied on sets of identical cell types (test, validation partition) have an accuracy from 95% to 99.8% (see Table [2]). The average accuracy in the cross-application of models built on one set of cells to sets of the other three cell types is greater than 91%, with the exception of a few methods that shoved an accuracy of 60%. These results occur in cross-application of 10X and GEO models to BS1 and BS2 data in selected sampling. The reason is that 10X and GEO did not include DC cell types and the constructed models have lower accuracy when applied on samples of BS1/BS2 data with a high percent of DC cell types. The accuracy of prediction for some data mining/machine learning algorithms is shown in Table 2.

	Group	C5.0	CART	CHAID	QUEST	RF	SVM RBF	XGB-AS	XGBT
Training	10x	99.82	99.59	99.31	99.37	99.9	99.44	99.74	99.89
	BS1	96.68	94.96	95.25	92.18	99.04	95.95	96.69	99.04
	BS2	96.95	95.21	94.88	92.90	99.04	96.43	96.54	98.65
	GEO	99.67	98.22	97.39	95.48	99.72	95.58	98.80	99.79
Test	10x	99.86	99.56	99.12	99.32	99.62	99.46	99.78	99.61
	BS1	96.58	93.85	93.67	91.74	95.05	95.50	96.51	95.63
	BS2	96.79	93.72	93.34	92.69	95.19	95.90	96.52	95.05
	GEO	99.68	96.70	95.07	95.43	97.48	94.99	98.67	97.46
Validation	10x	99.81	99.50	99.16	99.40	99.61	99.43	99.73	99.61
	BS1	96.06	93.46	93.19	91.24	94.19	94.64	95.64	94.15
	BS2	96.66	93.24	93.24	92.36	94.66	95.33	96.58	95.03
	GEO	99.54	96.71	95.85	95.57	97.48	95.26	99.05	97.94

Table 2. Part of PBMC types prediction results. The first group indicates accuracy of prediction in Training phase for different algorithms. Results are presented for all four groups of cells. The second and third group (Test and Validation) represent the accuracy of applying the model from the Training phase to sets of cells on previously unknown cell types in the Test/Validation partition.

A comparison of the results for models based on a different number of genes shows that the difference in accuracy was around 2%-3%, depending on the algorithm and data set. These results are encouraging, and since models with fewer genes require less computing resources and time to calculate, further work on this problem and an increase in efficiency and accuracy of the predictions will enable a relatively fast construction of devices that can be used in disease diagnosis.

4. Conclusions

This paper presents a practical example of applying Big Data Analytics by supervised classification techniques to bioinformatics data. Obtained results indicate the feasibility of classification algorithms for the effective prediction of PBMC cell types using expressions of a relatively small number of genes. Improving the Big Data Analytics process and increasing the efficiency of algorithms will enable rapid implementation of the proposed solutions in medical diagnostics.

References

[1] D. D. Chaplin, “Overview of the immune response”, J. Allergy Clin. Immunol., vol. 125, 2, S3-23, 2010.

[2] K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, et al., The Impact of Food Bioactives on Health: in vitro and ex vivo models [Internet]. Cham (CH): Springer; 2015. Chapter 15. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500157/

[3] L. Yang, Y. Zhang, N. Mitic, D. B. Keskin, G. L. Zhang, L. Chitkushev, et al., “Single-cell mRNA Profiles in PBMC”, in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1318-1323, IEEE, 2020.

[4] Z. Wang, M. Gerstein and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics”. Nat. Rev. Genet., vol. 10, 1, pp. 57-63, 2009.

[5] IBM SPSS Modeler 18.2 Algorithms Guide, https://www.ibm.com/support/pages/spss-modeler-182-documentation

[6] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar: Introduction to Data Mining, 2nd ed, Pearson Education, 2019

[7] C. Aggarwal (ed.): Data Classification Algorithms and Applications, CRC Press, 2015 of relapse, Nat. Med., vol. 24, 4, pp 474-483, 2018.