May 2018

Big Data Analytics for Obesity Prediction

Ahsan BILAL [a,b], Alfredo VELLIDO [a,c] and Vicent RIBAS [b 1]
[a] *Universitat Politècnica de Catalunya (UPC BarcelonaTech)*
[b] *EURECAT: Centre Tecnològic de Catalunya*
[c] *Intelligent Data Engineering and Artificial Intelligence (IDEAI) Research Center*

Abstract. Feature selection (FS) is essential for the analysis of genomic datasets with millions of features. In such context, Big Data tools are paramount, but the use of standard machine learning models is limited for data with such low instances to features ratios [1]. Apache Spark is a distributed in-memory big data system with the potential to overcome this bottleneck. This study analyzes genomic data related to prediction of human obesity. Since Apache Spark is unable to cope with our dataset containing ≈ 0.74 million features, we propose here a pipeline to solve this problem using partitioning strategies, both vertical, by dividing the data based on gender, and horizontal, by splitting each chromosome into 5,000-instances subsets. For each subset, *Minimum Redundancy and Maximum Relevance* FS was used to find rankings of the most relevant features. The challenge, thus, is making accurate obesity predictions with parsimonious subsets of features selected from millions of them. We tackle it by defining a 2-phase pipeline: first learning with individual chromosomes and then learning with joined 22 chromosomes from selected features.

Keywords. Big Data, Apache Spark, Feature Selection, Data Partitioning, Obesity

1. Introduction

Obesity is the result of an excess of fat in the body, defined by genetic and environmental factors. It is a chronic diseases and causes increased levels of circulating fatty acids. Individuals suffering obesity are at higher risk of developing a number of serious concurrent medical conditions [5]. The objectives of this paper are three-fold and can be briefly summarized as follows: (1) Design and implementation of a data pipeline to process and analyze feature-oriented genetic datasets with millions of items; (2) finding the most relevant features (single nucleotide polymorphisms, or SNPs [6]) and their ranking for a specific disease under analysis, namely obesity, for each chromosome; and (3) Forecasting obesity among males and females separately.

2. Experimental Dataset

The analyzed dataset comprises genomic data from a series of patients. This dataset was originally available in a compressed binary format. The base dataset consists of 22 chromosomes, whereas chromosome 23 is related to sex and is not considered. A total of 4,988 patients and 736,990 SNPs with categorical representation, denoted by 0, 1, 2 were available.

3. Proposed Data Pipeline

Our proposed data pipeline is divided into two phases. In the first phase, the data are horizontally (by rows) and vertically (by columns) partitioned and a data preparation strategy is applied to each partition. In the second phase, we merge the results from all 22 chromosomes and obtain a final model based on top relevant features which influence the obesity in males and females.

The first phase is further divided into five stages: *Data Partitioning, Data Transposition, Feature Selection, Data Merging, Classifier*. A high level view of the *Phase 1* architecture can be seen in **Figure 1**.

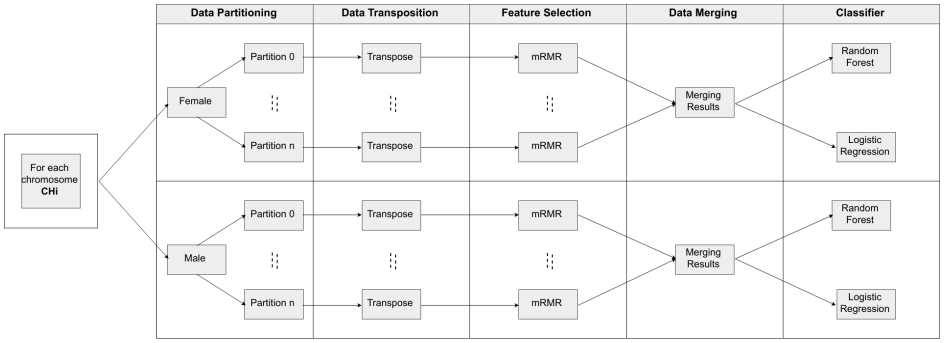

Figure 1. First phase of data pipeline.

The first stage of pipeline consists of dividing the data into horizontal and vertical partitions. This partitioning enabled us to run the job in the available Apache Spark cluster.

The partitioning solution initially implemented in Apache Spark [2] involved writing partitioned data in HDFS (a Java-based file system for data storage). The Apache Spark version 2.0 exceptions that were solved by turning to the use of PLINK[7][2] and Linux commands instead. This solution was fast and efficient.

In the second stage of the first phase, we transposed the data of each partition for each chromosome CH_i for males and females separately. This stage was needed due to the required format structure of the data (SNPs as *variables* and patients as *samples*) to apply FS, in contrast to the original structure of the provided data (patients in columns and SNPs in rows).

The third stage consists on the FS procedure as applied to each partition of the chromosome CH_i for males and females separately. In this stage, we applied the *Minimum Redundancy and Maximum Relevance (mRMR)* filter method [3,4] to select the top 20 features according to their ranking for each partition of the data. To summarize, from all 22 chromosomes of males and females, we selected only 3,040 SNPs variants; that is, a mere 0.41% of the total SNPs.

Finally, we merge the data in the fourth stage using the top 20 features from each partition of the chromosome CH_i (obtained in the 3^{rd} stage) selected to be the most relevant as obesity predictors and build the classifiers in the fifth stage for the data from

[2]PLINK is a widely used application for analyzing genotypic data. It can be considered the *de facto* standard in the field.

each chromosome (CH_i), for females and males separately. We split the data into training (70%) and test (30%).

The second phase of our pipeline is the backbone of the whole procedure. It combines the top relevant features from each partition of all 22 chromosomes to create a single dataset with a reduced number of features for males and females.

Approximately 200 features were selected from each chromosome and learning models were built individually, evaluating their accuracy. The final step of the proposed data pipeline involves finding common features that are available in both male and female datsets, ranking them according to the mRMR FS score. The conceptual view of phase 2 architecture is shown in **Figure 2**.

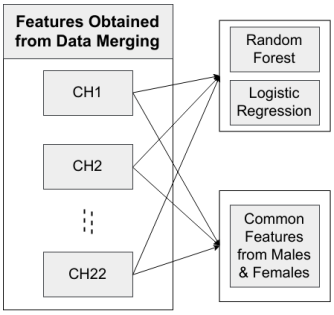

Figure 2. Second phase of data pipeline.

4. Experimental Results

The experiments were performed in two different types of machines: (1) YARN with 3 executers, 27GB RAM and 7 CPUs; (2) local with 1 executer, 8 GB RAM and 4 CPUs.

The performance of mRMR was extremely slow for 5,000 features using the maximum resources of YARN machine. It took several days to process all the partitions from all 22 chromosomes for both genders.

Next we report the classifier results. 5-Fold Cross Validation was used in order to find the best parameters of the model and increase efficiency. A Random Forest (RF) was first used, but accuracy did not increase beyond 0.5. After analyzing its implementation in Spark ML, we found that it could not handle imbalanced binary class distributions. To overcome this problem, a down-sampling technique was used to reduce the number of cases in the majority class (in our case "0"). Alternatively, Spark can manage the weights with imbalanced binary classification using a Logistic Regression (LR) model.

Finally, we merge the top selected 0.41% of SNPs from all chromosomes, combining them in a single dataset. During the evaluation of each chromosome, we found that LR performed well for the binary classification problem. We also observed that for LR method without sampling and with weights, does not change the AUC significantly, although with LR-Weights it slightly increased the AUC. The results for all 22 chromosomes combined are shown in **Table 1**.

Table 1. Combined performance, as measured by AUC, with all 22 chromosomes

Sampling (Classifier)	Gender	Test AUC	CV AUC
Weight (LR)	Male	0.971	0.962
Weight (LR)	Female	0.965	0.948
No Sampling (LR)	Male	0.963	0.941
No Sampling (LR)	Female	0.925	0.923
Down-sampling (RF)	Male	0.782	0.784
Down-sampling (RF)	Female	0.632	0.678
No Sampling (RF)	Male	0.500	0.501
No Sampling (RF)	Female	0.500	0.500

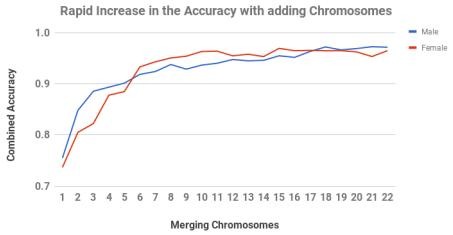

Figure 3. Graphical representation for combined performance of all 22 chromosomes

5. Discussion

From **Table 2**, we see that the group of a dozen chromosomes play a relevant role in achieving the best classification for males. Selected features from these chromosomes contribute individually for predicting obesity with accuracy higher than 70%. For females, chromosomes 1, 2, 5 and 6 achieved the highest results.

The accuracy for all 22 chromosomes also shows that the percentage of the accuracy is comparatively higher in the chromosomes with higher number of variants (SNPs), e.g., chromosome 22 shows the poorest performance for males and females.

The mutual relevant information ($\approx 67\%$ for males and $\approx 53\%$ for females) is contributed by chromosome 2 alone, whereas others provide less information. Thus, it appears that the top 220 features selected from chromosome 2 are more relevant in predicting the obesity.

The graphical illustration of the **Table 2** in **Figure 3** shows that the first six chromosomes already increase the performance of the final model over the 90% threshold for both females and males). From **Figure 4**, it can be concluded that performance increase by adding more chromosomes only marginally. In conclusion, we have found preliminary evidence that combining the features of just 6 chromosomes (a very parsimonious selection if compared with the complete original dataset), obesity can be predicted quite accurately.

References

[1] Li, J, and Liu, H. *Challenges of feature selection for big data analytics.* IEEE Intelligent Systems 32(2), 2017: 9-15.

[2] Salloum *et al. Big data analytics on Apache Spark*, International Journal of Data Science and Analytics, 2016: 1–20.
[3] Ramrez-Gallego, S, *et al. Fast-mRMR: Fast Minimum Redundancy Maximum Relevance algorithm for high-dimensional Big Data.* International Journal of Intelligent Systems, 32, 2017: 134-152.
[4] Ding, C, and Peng, H. *Minimum redundancy feature selection from microarray gene expression data.* Journal of Bioinformatics and Computational Biology 3(2), 2005: 185-205.
[5] Severinsen *et al. Genetic susceptibility, smoking, obesity and risk of venous thromboembolism*, British Journal of Haematology, 149(2), 2010: 273–279.
[6] *What are single nucleotide polymorphisms (SNPs)?*. U.S. National Library of Medicine https://ghr.nlm.nih.gov/primer/genomicresearch/snp
[7] Purcell *et al. PLINK: a toolset for whole-genome association and population-based linkage analysis.*, American Journal of Human Genetics, 81, 2007.

Table 2. Highest accuracy for all 22 chromosomes individually with Males and Females

Ch	mRMR (Male)	Accuracy (Male)	mRMR (Female)	Accuracy (Female)
1	0.6256	**0.755**	0.5113	**0.737**
2	**0.6663**	**0.762**	**0.5318**	**0.738**
3	0.6061	0.733	0.4681	0.697
4	0.5782	**0.765**	0.3942	0.692
5	0.4956	**0.727**	0.4159	**0.705**
6	0.5658	**0.752**	0.4457	**0.715**
7	0.4219	0.682	0.3537	0.682
8	0.4656	**0.742**	0.3704	0.666
9	0.3848	**0.710**	0.3177	0.685
10	0.2697	0.592	0.2672	0.698
11	0.4154	0.696	0.3514	0.685
12	0.4088	**0.732**	0.3185	0.669
13	0.3262	0.658	0.2543	0.691
14	0.2813	0.662	0.2318	0.650
15	0.2723	0.677	0.2273	0.646
16	0.2904	0.692	0.2258	0.636
17	0.2314	0.657	0.1828	0.627
18	0.2757	0.685	0.2292	0.651
19	0.1781	0.661	0.1357	0.628
20	0.2107	0.650	0.1778	0.635
21	0.1224	0.621	0.0939	0.627
22	0.1098	0.578	0.0981	0.586

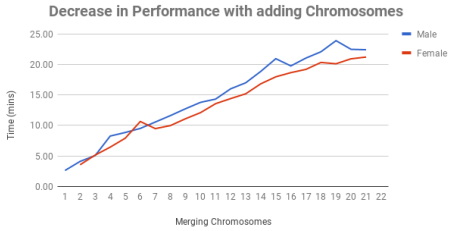

Figure 4. Decreasing performance with adding more chromosomes