Enabling interpretation of the outcome of a human obesity prediction machine learning analysis from genomic data

Ahsan Bilal[1,2], Alfredo Vellido[1,3], Vicent Ribas[2]

(1) Universitat Politècnica de Catalunya (UPC BarcelonaTech)
Barcelona 08034, Spain, avellido@cs.upc.edu

(2) EURECAT: Centre Tecnològic de Catalunya
Barcelona 08005, Spain, vicent.ribas@eurecat.org

(3) Intelligent Data Engineering and Artificial Intelligence (IDEAI) Research Center
Barcelona 08034, Spain

Keywords: Machine Learning, Feature Selection, Minimum Redundancy and Maximum Relevance, SNP, Big Data, Apache Spark, Obesity.

Abstract. In this brief paper, we address the medical problem of human obesity prediction from genomic data. Genomic datasets may contain a huge number of features and they often have to be analyzed within the realm of Big Data technologies. As a medical problem, obesity prediction would welcome interpretables outcomes. Therefore, the analyst would benefit from appraches in which the problem of very high data dimensionality could be eased as much as possible. Feature selection can be an essential part of such approaches. In this context, though, traditional machine learning methods may struggle. Here, we propose a pipeline to address this problem using partitioning strategies: both vertical, by dividing the data based on gender, and horizontal, by splitting each of the analyzed chromosomes into 5,000-instances subsets. For each, *Minimum Redundancy and Maximum Relevance* feature selection is used to find rankings of the single nucleotide polymorphisms most relevant for classification in the medical dataset.

1 Introduction

The pervasive use of networked computer systems in medical and clinical environments has made medical research an increasingly data-dependent discipline. This brings to the fore many challenges related to operational data management and knowledge extraction from data [**?**].

This paper addresses the medical problem of human obesity prediction from genomic data. Genomic datasets (in general and in the particular case of this study) may contain a huge number (even millions) of features. Not only that, but also, more often than not, showing very low ratios of instances to features. This has two immediate consequences: first, that the data require Big Data technologies for their management and analysis and, second, that traditional machine learning (ML) methods for data analysis and knowledge extraction may struggle in a low instances-to-features ratios scenario [**?**].

As a medical problem, obesity prediction would welcome interpretable outcomes that can be acted upon in an operational manner, even if for purely research-related puroposes. Therefore, the analyst would benefit from approaches in which the problem of large data dimensionality could be eased as much as possible. Feature selection (FS) for dimensionality reduction (DR) can be an essential part of such approaches and it is the strategy that we propose in our study.

Apache Spark is a distributed in-memory Big Data system with the potential to overcome these bottlenecks. Our analyses, though, show that Apache Spark is unable to cope with our dataset containing ≈ 0.74 million features. Here, as an alternative, we

propose a pipeline to address this problem using partitioning strategies: both vertical, by dividing the data based on gender, and horizontal, by splitting each of the analyzed chromosomes into 5,000-instances subsets. For each subset, *Minimum Redundancy and Maximum Relevance* (mRMR) FS is then used to find rankings of the most relevant single nucleotide polymorphisms (SNPs) in a medical dataset.

The remaining of the paper is structured as follows: first we describe the FS approach followed in the study. We then briefly describe the analyzed data and the pipeline used for their tractable analysis. This is followed by the reporting of the experimental results and some summary conclusions.

2 Methods: Feature Selection

FS can be described as a process of automatic tagging of subsets of features as relevant for model construction. FS is by itself useful, but it mostly acts as a filter, muting out features that are not useful for the purpose of analysis. As commented in [**?**], FS "has shown its effectiveness in many applications, but the unique characteristics of big data present challenges". Usually, real-world datasets come with a sizeable amount of irrelevant and redundant features. FS also helps data analysis by decreasing memory storage requirements and computational cost, while avoiding information loss in as much as possible [**?**].

FS methods can be used for identifying and removing from data those unwanted, irrelevant and redundant variants that do not contribute to the accuracy of a model, or may in fact decrease the accuracy of the model. Redundant features are adscribed to the category of irrelevant ones. Each feature is to some extent relevant and cannot be discarded manually, but redundancy implies the co-presence of another feature with similar peformance, and the model's learning performance will not be compromised by removing one of them [**?**].

According to Guyon and Elisseeff [**?**], the most important objectives of FS are:

- to reduce overfitting and improves the model performance in sense of prediction,
- to provide faster and cost effective models,
- to achieve an easy interpretation of the model by domain users using only a small subset of data.

Although FS techniques are very handy in large-scale datasets and are widely used, there are also a few aspects that require being careful about in the process. The advantages of FS techniques come at a certain price, because the search for a subset of relevant features introduces an additional layer of complexity in the modeling task. Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset, as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset [**?**, **?**, **?**, **?**].

From a ML point of view, the selection of biomarkers in our medical context can be stated as a FS problem for a classification task, where we have the objective of finding a reasonably small set of features (biomarkers) that is capable of best explaining the difference between the disease and the control samples [**?**].

From a biological point of view, Haury *et al.* explain that applying FS to biological case/control datasets allows to investigate the genes selected in the signature and evaluate the relationship to biological processes involved in the disease [**?**].

2.1 *Minimum-Redundancy-Maximum-Relevance (mRMR)*

mRMR was first developed by Peng *et al.* [**?**] and it is considered as one of the most powerful filter methods. It is based on mutual information and selects features

according to the maximum statistical dependency to the class label. Selecting a small but meaningful subset out of several thousands or millions of biomarkers is a most relevant task, not only for achieving the most accurate classification of biomedical data, but also for enabling biomedical interpretability. The mRMR algorithm was developed and intended to deal with the classification of DNA microarray data, which is a challenging task when faced with a huge number of features (SNPs in this case) paired with a limited number of observations. In their study, Peng *et al.* stated that selected genes via mRMR provide a more balanced coverage of the space and capture broader characteristics of phenotypes [**?**].

The mRMR algorithm ranks the importance of the features based on their relevance to the class. As the name suggests, the main goal is to achieve the maximum relevancy between the features *X* and the class *C*, using mutual information (MI).

$$I(A,B) = \sum_{b \in B} \sum_{a \in A} P(a,b) log \left(\frac{P(a,b)}{P(a)P(b)} \right) \tag{1}$$

In the above Eq.**??**, *I* represents the mutual information between the features *a* and *b*, which can be easily derived by calculating the marginal probabilities *P(a)* and *P(b)*, and the joint probability between both features *P(a,b)* [**?**].

The maximum relevance can be determined by Eq.**??** [**?**],

$$max D(X,C), D = \left(\frac{1}{|X|} \right) \sum_{X_i \in X} I(X_i; C) \tag{2}$$

Since the redundancy is a major issue in this feature selection task, specially when targeting the maximum relevancy criterion for large datasets, we can minimize the redundancy according to the following Eq.**??**, as suggested in [**?**].

$$min R(X), R = \left(\frac{1}{|X|^2} \right) \sum_{X_i, X_j \in X} I(X_i; X_j) \tag{3}$$

Finally, the combination of both Eq.**??** and Eq.**??** helps deriving the desired output i.e. mRMR in Eq.**??**, where *S* is the selected set of features [**?**].

$$max_{X_i \notin S} [I(X_i, C) - \left(\frac{1}{|S|} \right) \sum_{X_j \in S} I(X_j; X_i)] \tag{4}$$

3 Materials: Experimental Dataset

The analyzed dataset comprises genomic data from a series of patients. 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 were available.

4 Proposed Data Analysis Pipeline

Our proposed data analysis pipeline is based on a complex data pre-processing stage that includes *data partitioning, data transposition, feature selection, data merging* and building the classifier. First, the data are partitioned horizontally (by rows) and vertically (by columns) into subsets of 5,000 features and based on gender, respectively, and a data preparation strategy is applied to each partition. Second, we merge the results from all 22 chromosomes and obtain a final model based on top relevant features which influence

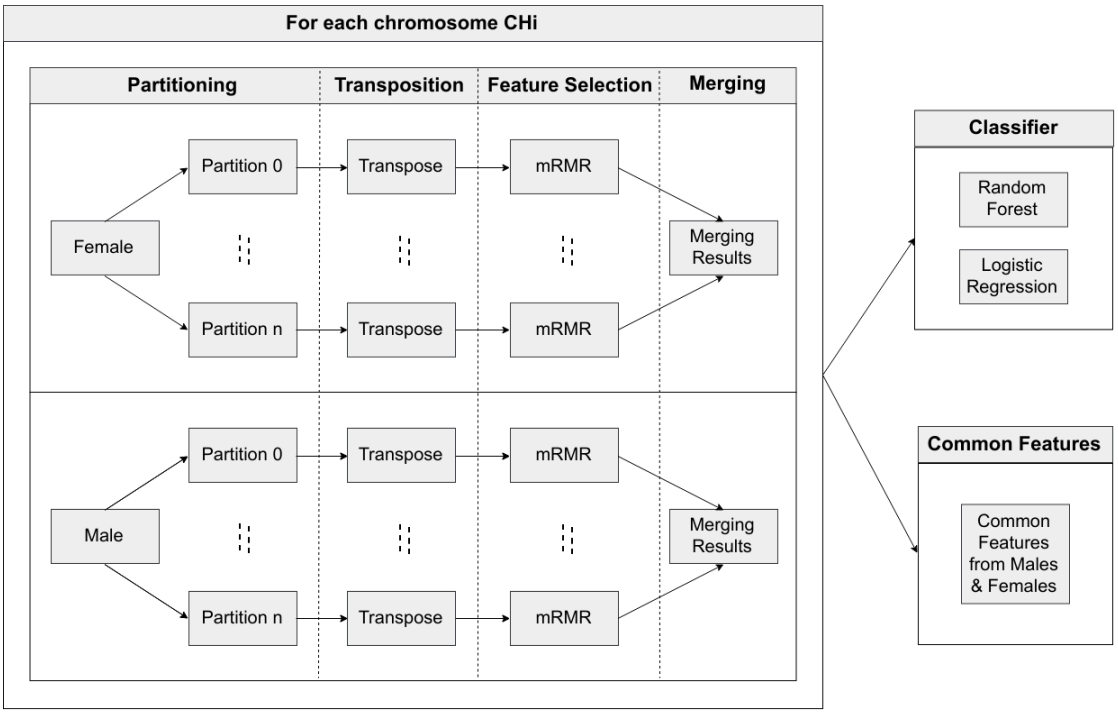

Figure 1: Data pipeline.

the obesity in males and females. A high level view of the proposed pipeline architecture can be seen in Fig. **??**.

The first stage of data preparation consists on dividing the data into horizontal and vertical partitions. This partitioning enabled us to run the job in the Apache Spark cluster available for data handling. The partitioning solution that was initially implemented in Apache Spark involved writing the partitioned data in HDFS (a Java-based file system for data storage). The Apache Spark version 2.0 generated exceptions that were handled by turning to the use of PLINK (a widely used application for analyzing genotypic data that can be considered the *de facto* standard in the field) and Linux commands instead, for partitioning the data first into gender-specific subsets and, second, into subsets containing 5,000 features. This solution was found to be fast and efficient.

Subsequently, data of each partition were transposed for each chromosome CH_i for males and females separately. This stage was necessary due to the required format structure of the data (SNPs as *variables* and patients as *samples*), so that the FS procedure could be applied. Note that in the original structure of the provided data, patients were described in columns and SNPs in rows.

Finally, FS was applied to each partition of the chromosome CH_i for males and females separately; the selected features found to be the most relevant as obesity predictors were merged; and the classifiers were built by splitting the data into training (70%) and test (30%) sets. Through the mRMR filter method [**?**, **?**], the top 20 features were selected according to their ranking, for each partition of the data. In summary, from all 22 chromosomes, both for males and females, only 3,040 SNPs variants were selected; that is, a mere 0.41% of the original total amount of SNPs available for analysis.

Approximately, 140 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 datasets, ranking them according to the mRMR FS score.

5 Results

The experiments were performed on a YARN machine with 3 executers, 27GB RAM and 7 CPUs. The performance of mRMR was extremely slow for 5,000 features using

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

the maximum resources. It took several days to process all the partitions from all 22 chromosomes for both genders.

A Random Forest (RF) classifier was first used, in a 5-Fold Cross Validation procedure used to find the best parameters of the model and increase efficiency. Unfortunately, the preliminary accuracy results were poor. After analyzing its implementation in Spark ML, we found that it could not properly 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. Alternatively, Spark can manage the weights with imbalanced binary classification using a Logistic Regression (LR) model, that was also used as classifier.

Finally, we merged 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 quite well in the binary classification problem. We also observed that not using sampling or weighting in the LR method did not have any significant negative impact in performance as measured by the the Area Under the ROC Curve AUC, although the weighting in the *LR-Weights* model slightly increased the AUC. The results for all 22 chromosomes combined are shown in Tab. **??** and common SNPs found in both males and females are listed in Tab. **??**.

6 Conclusions

Interpretability is paramount in medical applications of ML in general, but it is particularly difficult when the medical problem, obesity prediction in the case of this study, is defined according to genomic data. This setting requires the use of Big Data tools and technologies as we nned to extract knowledge from thousands of individuals described through millions of features (SNPs in this study). Systems still lack flexibility for bioinformatics data described through millions of *features* in a distributed manner and not just millions of records.

We have proposed a data analysis pipeline design using data partitioning for Big Data, which has solved feasibility issues in an Apache Spark 2.0 framework, allowing us to run jobs using the available resources. We reckon that running these tasks with maximized resources according to the proposed pipeline would definitely lead to a good computational performance.

Through feature engineering and FS-based DR, we have managed to reduce from the original bulk of 736,990 SNPs to an extremely lean 3,040 SNP selection, while providing a quite accurate obesity prediction (0.965 AUC for females and 0.971 AUC for males). This result, with specific SNP selections related to specific chromosomes, is the first and necessary step for guaranteeing the interpretability of any biomedical research oriented towards explaining human obesity from this type of genomic data.

Table 2: Common SNPs from Males and Females.

SR No.	SNP	Chromosome
1	2:4259627:C:T	2
2	2:224060700:G:A	2
3	2:233158545:C:T	2
4	3:125050868:T:C	3
5	4:130008848:T:C	4
6	6:30127079:T:C	6
7	6:32975283:G:T	6
8	6:30233192:T:C	6
9	6:28865417:T:C	6
10	7:1932780:G:C	7
11	8:30430742:T:C	8
12	8:133210054:T:C	8
13	8:143486205:G:A	8
14	9:119600196:T:C	9
15	11:91411734:A:G	11
16	13:67193281:C:T	13
17	13:101857816:G:A	13
18	14:57747325:C:T	14
19	14:92758540:G:C	14
20	15:96771641:T:G	15
21	16:7403274:T:G	16
22	19:11096293:G:A	19

Acknowledgements

This research was partially funded by the Spanish MINECO TIN2016-79576-R.

References

[1] J. Li and H. Liu. "Challenges of feature selection for big data analytics." *IEEE Intelligent Systems* vol.32, no.2, pp. 9-15, 2017.

[2] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective." *ACM Computing Surveys (CSUR)* vol.50, no.6, p. 94, 2017.

[3] H. Liu and H. Motoda, "Computational methods of feature selection.", CRC Press, 2007.

[4] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection." *Journal of Machine Learning Research* vol.3, pp. 1157-1182, 2003.

[5] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics." Bioinformatics vol.23, no.19, pp. 2507-2517, 2007.

[6] W. Daelemans, V. Hoste, F. De Meulder, and B. Naudts, "Combined optimization of feature selection and algorithm parameters in machine learning of language." *European Conference on Machine Learning*. Springer, Berlin, Heidelberg, 2003.

[7] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods." *Bioinformatics* vol.26, no.3, pp. 392-398, 2009.

[8] A.-C. Haury, P. Gestraud, J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures." PLoS ONE, vol.6, no.12, p. e28210, 2011.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol.27, no.8, pp. 1226-1238, 2005.

[10] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data." *Journal of Bioinformatics and Computational Biology* vol.3, no.2, pp. 185-205, 2005.

[11] S. Ramírez-Gallego, I. Lastra, D. MartínezRego, V. BolónCanedo, J.M. Benítez, F. Herrera, and A. AlonsoBetanzos, "Fast-mRMR: Fast Minimum Redundancy Maximum Relevance algorithm for high-dimensional Big Data." *International Journal of Intelligent Systems*, vol.32, pp. 134-152, 2017.