Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction

Beatriz López[a], Ferran Torrent-Fontbona[a,*], Ramón Viñas[a], José Manuel Fernández-Real[b,c]

[a] *University of Girona, Campus Montilivi, building EPS4, 17071 Girona, Spain*
[b] *Biomedical Research Institute of Girona, Avda. de França, s/n, 17007 Girona,Spain*
[c] *CIBERobn Pathophysiology of Obesity and Nutrition, Instituto de Salud Carlos III, Madrid, Spain*

Abstract

Objective: The use of artificial intelligence techniques to find out which Single Nucleotide Polymorphisms (SNPs) promote the development of a disease is one of the features of medical research, as such techniques may potentially aid early diagnosis and help in the prescription of preventive measures. In particular, the aim is to help physicians to identify the relevant SNPs related to Type 2 diabetes, and to build a decision-support tool for risk prediction.

Methods: We use the Random Forest (RF) technique in order to search for the most important attributes (SNPs) related to diabetes, giving a weight (degree of importance), ranging between 0 and 1, to each attribute. Support vector machines and logistic regression have also been used since they are two other machine learning techniques that are well-established in the health community. Their performance has been compared to that achieved by RF. Furthermore, the relevance of the attributes obtained through the use of RF has then been used to perform predictions with k nearest neighbour method weighting attributes in the similarity measure according to the relevance of the attributes with RF.

Results: Testing is performed on a set of 677 subjects. RF is able to handle the complexity of features' interactions, overfitting, and unknown attribute values,

*Corresponding author
Email addresses: beatriz.lopez@udg.edu (Beatriz López), ferran.torrent@udg.edu (Ferran Torrent-Fontbona), rvinast@gmail.com (Ramón Viñas), jmfreal@idibgi.org (José Manuel Fernández-Real)

providing the SNPs' relevance with an up to 0.89 area under the ROC curve in terms of risk prediction. RF outperforms all the other tested machine learning techniques in terms of prediction accuracy, and in terms of the stability of the estimated relevance of the attributes.

Conclusions: The random forest is a useful method for learning predictive models and the relevance of SNPs without any underlying assumption.

Keywords: Type 2 Diabetes, Random Forest, Feature Learning, Predictive model, Gini importance

1. Introduction

There is a great deal of interest in finding the SNPs that are related to a given illness in order to appropriately develop a corresponding personalised treatment. The first approaches with regard to studying the relationships between SNPs and diseases focused on single individual variable analysis, where a variable (SNP) is removed, and then some predictor indicator is analysed to measure the impact of the variable influence. However, interactions between variables meant that this approach did not perform well. Therefore, other approaches based on machine learning techniques, which enable the analysis of multiple combinations of variables, are preferred [1, 2]

The authors in [3] provide an overview of the different machine learning techniques applied to SNP data, from which two main approaches are distinguished: SNP association studies and predictive modelling. While SNP association studies consists of grouping SNPs according to their expression profiles (e.g. molecular function, biological process, cellular components), predictive modelling aims to identify which features are relevant to a specific function or class. For example, which features are particularly relevant with regard to Type 2 Diabetes (T2D).

Concerning the use of predictive models with SNP data, these methods suffer from the dimensionality problem: hundreds of subjects (samples), with thousands of SNPs per subject (features, attributes). As a consequence, [3] warns

about the risk of incurring an overfitting problem [4] when applying machine learning techniques to such kinds of datasets.

To overcome the overfitting problem, regularisation techniques [4] are applied, but they present some difficulties regarding tuning the appropriate regularisation parameter [5] and moving away from the current trends of personalised medicine [6]. On the other hand, the random forest (RF) technique [7] has been proven to outperform most of the current machine learning techniques when it comes to building classification models in general, and predictive models in particular, without any underlying assumptions [8]. Moreover, it is a computationally efficient technique and one that is almost free of parameters. The RF technique consists of building a given number of decision trees (a forest), which are combined in an ensemble mechanism (e.g. majority voting) in order to obtain a final classification outcome (e.g. ill or healthy), with a confidence degree associated with the result (a prediction indicator).

The relevance of features that conform to the RF model is obtained by aggregating the relative importance of the features over all of the trees [9, 10]. Therefore, no particular pre-processing techniques are required for feature selection [11, 12]. Moreover, as an ensemble technique, the obtained relevance feature set is stable [13, 14].

An additional property of the RF technique is its capacity for handling missing information [15], a common situation when dealing with SNP data [16]. This is due to the ensemble nature of the RF method, which combines several decision trees to provide a classification outcome (e.g. prediction) [7]. Each decision tree is learned by using a subset of features (SNPs) that are randomly selected, as well as a subset of samples that are also randomly chosen. However, the RF technique does not remove any information, maintaining the changes towards a personalised outcome.

This paper addresses the application of the RF technique to a dataset of SNPs, which has a significant percentage of missing information, classified in terms of people with T2D and people without it. In particular, our work concerns the identification of the relevant SNPs related to T2D. Furthermore,

RF performance is compared with other well-stablished machine learning techniques, such as Support Vector Machines (SVM) and Logistic Regresion (LR)

2. Material and methods

The Biomedical Research Institute of Girona has been gathering information about the SNPs of subjects with their corresponding diagnoses (T2D, glucose intolerance), as well as that of healthy subjects. Based on the available data, a T2D risk prediction model has been obtained with the use of the RF technique, from which the relevance of each feature is obtained by using the Gini importance [9].

2.1. The problem

The problem addressed in this paper is to find the relevance of a set of SNPs $g_1, g_2, \ldots, g_n$, given a set of samples P corresponding to people with and without T2D, in order to enable the prediction of T2D.

Each sample is noted as (x, y), where x is a list of attribute-value pairs $\langle g_i, v_i \rangle$ regarding the SNPs $g_1, g_2, \ldots, g_n$ and their values $v_1, v_2, \ldots, v_n$ for the given sample; and y is the class to which the person belongs. In this particular case, $y \in C = \{healthy, Type2diabetes\}$. Attributes, SNPs and features are used synonymously throughout this work [1].

Each SNP i has NVA_i values. In our particular case, $NVA_i = 4$ $(\forall i)$, with the following interpretation: 1: the SNP is not present; 2: the SNP is present; 3: the SNP has been expressed; 4: *unknown* value. Therefore, we are considering SNPs with missing information [2].

[1] Attributes is often the proper notation of supervised machine learning methods; SNPs of genetics, and features of feature learning methods

[2] In fact, this could be considered as a unique-value imputation method, as the unknown or missing value is treated as another attribute value [17].

2.2. Random Forest

RF is a supervised learning method, which means that each instance or sample is labelled with the outcome (class).

RF consists of an ensemble of k classifiers $h_1(x), h_2(x), \ldots, h_k(x)$, with $h(x)$ being the joint classifier [7, 18]. Each classifier $h_i(x)$ consists of a decision tree, in which nodes are attributes (see Figure 1). The selection of which attribute is collocated in a node n is performed as follows: 1) a subset of attributes is randomly selected, 2) an evaluation measure is applied to the selected attributes according to their capability for providing homogeneity partitions of the samples, and 3) the attribute with the highest score is chosen. In particular, we use the change of the Gini impurity[3] [18] to compute the score, as described in Equation (1)

$$\Delta G(g_i, n) = -\sum_{C_k \in C} p^2(C_k) + \sum_{j=1}^{NVA_i} p(v_{i,j}) \sum_{C_k \in C} p^2(C_k|v_{i,j}) \qquad (1)$$

where $v_{i,j}$ is the j value of the i SNP. Probabilities are estimated according to the instances that reach the n node.

Once a node is assigned to an attribute g_i, the data is split into as many sets as values the g_i attribute has (four). Then, the tree is grown with new nodes in each branch. These are obtained by repeating the attribute selection process. The stopping condition is defined according to the number of instances that remain in a node: if this number is lower than a given threshold τ, the algorithm stops. Samples used to build each tree are also selected randomly with replacements.

Once the RF is built, it can be used (tested) for predicting the T2D risk of a person. Given a query case q, with a list of SNP-value pairs $\langle g_i, v_i \rangle$, each decision tree provides an outcome, $h_i(q)$. The final prediction (class for q) is obtained by using an averaging mechanism that combines the probabilistic prediction of

[3]Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

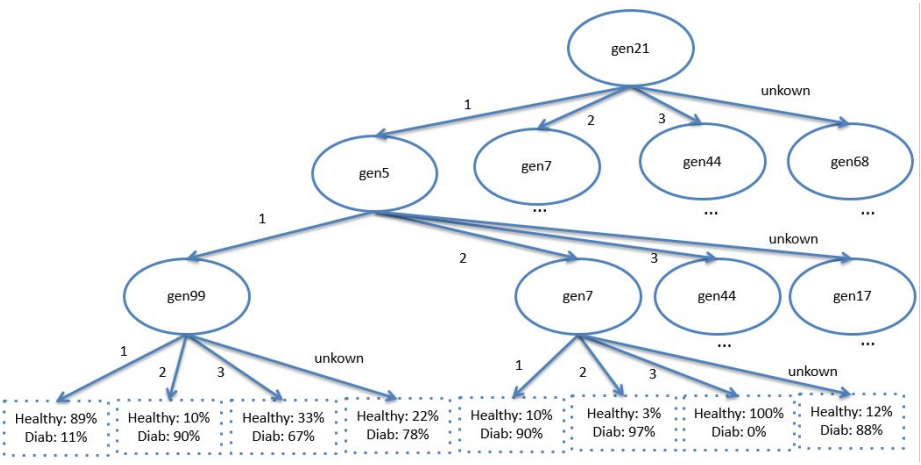

Figure 1: Example of a decision tree.

each tree regarding each class. The class with the highest prediction is assigned to q.

2.3. SNP relevance

The relevance of each SNP is obtained by averaging the information with regard to the SNP in each node n of each tree t, as shown below.

$$GI(g_i) = \frac{1}{T} \sum_{\forall t} \sum_{\forall n \in t} p(n) \Delta G(g_i, n) \delta(g_i, n) \tag{2}$$

where T is the total number of trees in the RF; $\delta(g_i, n)$ is a boolean function that returns 1 if g_i has been selected as the splitting feature in node n, 0 otherwise; and $p(n)$ the proportion of cases in node n, i.e. $p(n) = \frac{|P_n|}{|P|}$, where $|P_n|$ is the number of samples that reach n . This is known as the Gini importance or Mean Decrease Gini [19].

2.4. Dataset and quality control

The experiments have been carried out with a dataset of 1074 subjects, but 246 subjects had an unknown diagnostic, and therefore they were removed from the dataset, leaving a total of 828 participants for the experiments. Each sample contains 101 SNPs regarding T2D.

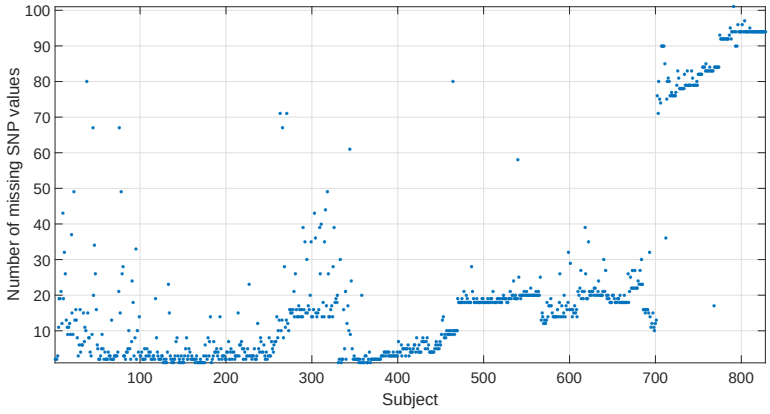

Figure 2: Rate of missing SNP values per sample. Horizontal-axis corresponds to the subject number of the 828 subjects with known diagnostic. Vertical-axis indicates the number of missing values (of the 101 SNPs) for each subject.

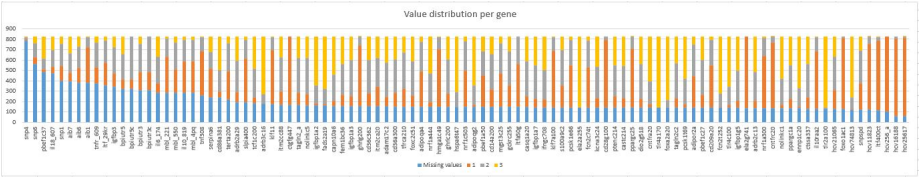

Figure 3: Distribution of SNP values of the 828 subjects with known diagnostic. SNPs are sorted according to the highest to lowest number of missing values. Blue: missing values; red: 1 value; grey: 2 value; yellow: 3 value. Vertical-axis represent the percentage of values (missing, 1, 2 and 3) for each SNP. Horizontal-axis correspond to SNPs.

Regarding missing information, Figure 2 shows the distribution of missing data among the different samples. It is worth observing that some of the samples accumulate a huge amount of missing information. On the other hand, Figure 3 shows the number of missing values per SNP[4] (blue colour). SNPs have been ordered in the x-axis according to the amount of missing values.

The elimination of samples with a large amount of unknown SNPs benefits the learning process, although RF can manage unknown data. Recent studies

[4]SNPs names are hidden for simplicity reasons and medical research confidentiality issues.

argue that RF can suffer from some instability regarding prediction outcomes if the amount of irrelevant information they handle is high [20]. The improving on stability of RF has been recently addressed by using a k-Nearest Neighbour (k-NN) algorithm to perform feature selection [21]. This alternative approach can be explored in a near future regarding prediction, but some challenges should be considered regarding the generation of the SNP's relevance, due to the nature of k-NN. Therefore, we removed samples with more than 25% missing information. This kind of pre-process has been proven useful in previous work [15].

Moreover, five SNPs (hcv256_a, ctgfg447, ela2g741, ela2a255, tlr4a170) were used for subjects recruitment criteria and, as a consequence, they were also removed. At the end, 677 samples were left, with 96 SNPs each one.

The remaining dataset has the following characteristics:

- The 10.94% of SNPs have missing value. The distribution of missing values respect to the class is 48.03% in the healthy class and 51.97% in the diabetes class.
- there are 429 samples of healthy subjects. All of them have missing values.
- there are 248 samples of diabetic subjects. All of them have missing values.

2.5. Experimental set-up

A 10-cross validation method has been used for experimentation purposes. The original dataset is imbalanced, and we generated balanced sets. Therefore, the percentage of both classes in each fold was 50%.

The number of decision tress has been set at k=1000. According to [7], as the number of trees increases, the RF tends to converge on the real predictor.

In order to analyse the implications of RF on learning SNP's relevance, the following experimental scenarios have been defined:

- **Raw data**: The dataset is used as provided.
- **Clinical data**: SNP data has been combined with clinical information to understand the prediction capacity of SNPs.

- **SNP-value relevance**: Each SNP value has been substituted for a boolean variable, so each SNP-value pair is converted to a boolean variable that indicates whether the SNP has a particular value or not. For example, SNP1 has three possible values (1, 2 or 3), then instead of a variable to indicate the value of SNP1, we have three boolean variables (SNP1-1, SNP1-2 and SNP1-3). If SNP1 value is unknown, then the three SNP-value boolean variables also have unknown value. The aim is to obtain fine grain information about SNP-value interaction or interactions about SNP variations.

Moreover, a comparison analysis with other state of the art machine learning methods, such as SVM and LR, has been performed.

Results are analysed in terms of the average AUC over all of the folds, where AUC is the area under the ROC[5] curve. An AUC of 1 represents a perfect test, i.e. all subjects of class diabetic have been classified as diabetic without any healthy subject being classified as diabetic.

3. Results

Regarding the SNP relevance, results are measured according to the Gini importance measure as defined in Section 2.3. The Gini importance of each SNP is averaged over all of the k-folds and the mean and standard deviation are provided.

3.1. Initial results

An AUC of 0.853 ± 0.050 (average $\pm$ standard deviation) was obtained for the raw dataset. The mean relevance values obtained for each SNP is plotted in Figure 4. SNPs are sorted in the x-axis according to their original sort in the dataset. The SNP with the highest relevance, tnf308, has been proved to

[5]The Receiver Operating Characteristic curve (ROC curve) illustrates the ability of a binary classifier by plotting the true positive rate against the false positive rate. Therefore, it illustrates the cost, in false positives, of achieving a particular true positive rate.

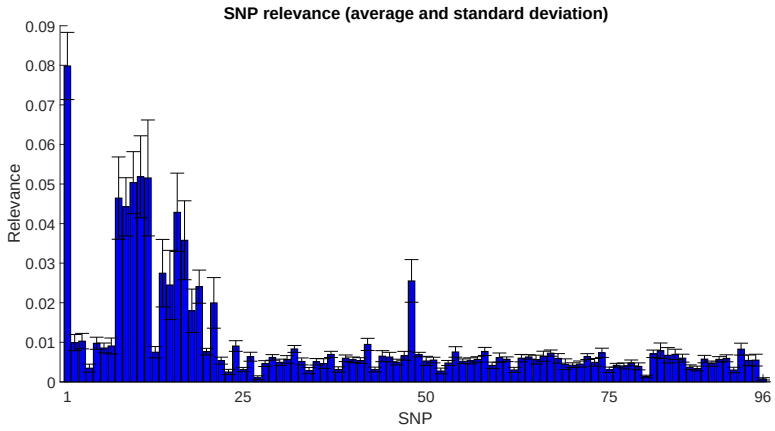

Figure 4: SNP relevance obtained with RF

be directly related to diabetes [22, 23]. Other high scored SNPs, such as snp4, are proven to be related to T2D [24, 25]. Moreover, it is worth noting in Figure 4 that the standard deviation of the weights is low, demonstrating the stability of the RF technique.

3.2. Adding clinical data

Information about sex, Body Mass Index (BMI) and age has been added to all of the samples in order to analyse the impact of such clinical variables on the prediction and, with them the AUC increased to 0.890 ± 0.041. Therefore, information about sex, BMI and age improves the prediction of T2D. However, when only clinical data is used, the AUC decreases to 0.624 ± 0.049, meaning that these variables are insufficient for T2D prediction.

The relevance values learnt for the clinical variables are important, except for sex, respect most of SNPs relevance values. This result coincides with the fact that an AUC of 0.890 is obtained without using information about sex (only SNPs, BMI and age), and the AUC is similar 0.854 when sex is added to the one achieved with only SNPs information.

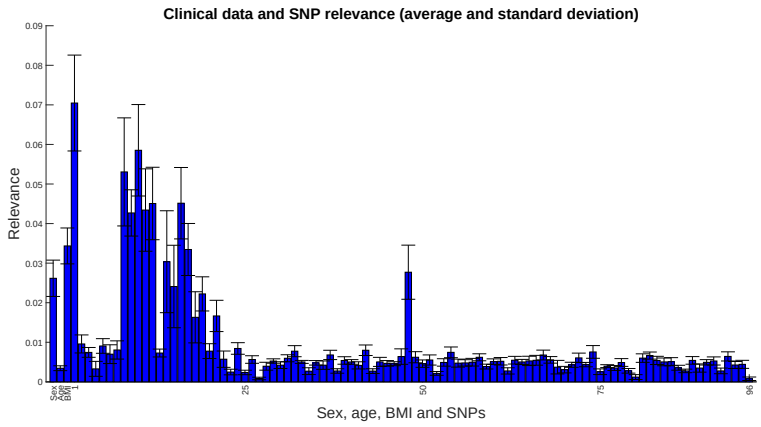

Figure 5: Feature relevance for SNPs and clinical variables (first three variables).

3.3. SNP-value relevance

The results in this scenario are worse since accuracy drops to 0.819 ± 0.046. The reason for the performance decrease can be found in the increase of the dimensionality problem, since the original set of 96 SNPs has been converted to a set of 96×3 variables, without increasing the number of samples.

The relevance values learnt in this scenario are shown in Figure 6, where SNPs with greater relevance in Figure 4 also have important relevance value for at least one SNP-value pair. However, there are interesting differences between SNP-value pairs for the same SNP. Therefore, despite the reduction in accuracy, using SNP-value attributes enables the comparison between the values of the SNPs. In this regard, in 30 of the 96 SNPs, the biggest relevance is when the value of the SNP is 1 (i.e. the SNP is not present). Similarly, in 43 of the 96 SNPs, the relevance is bigger for value 2 (i.e. the SNP is present); and in 23 SNPs, the relevance is bigger for a value of 3 (i.e. the SNP is expressed). Focusing on the 14 most relevant SNPs (see Figure 4), 4 times value 1 obtains the biggest relevance, 9 times value 2 and value 3 just once. Therefore, according to RF, the presence or not of these SNPs is more important than if they are expressed for the prediction of T2D.

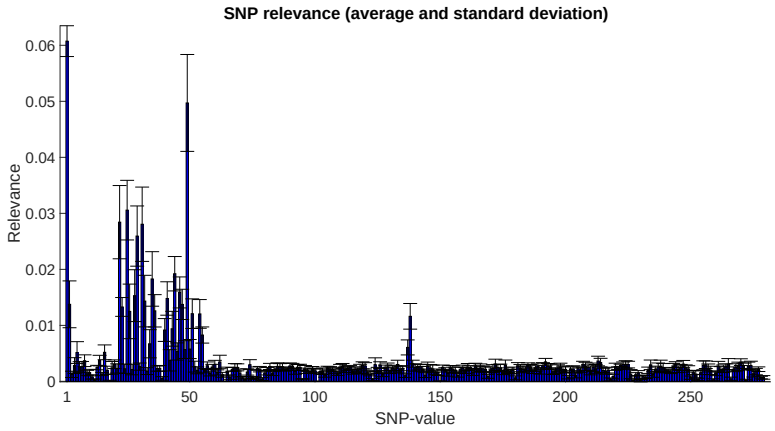

Figure 6: SNP-value relevance.

Scenario	RF	LR	SVM
Raw data	0.853 ± 0.050	0.835 ± 0.054	0.825 ± 0.044
With clinical data	0.890 ± 0.041	0.844 ± 0.050	0.825 ± 0.038
SNP-value	0.819 ± 0.046	0.791 ± 0.062	0.773 ± 0.059

Table 1: Summary of the results for all of the methods

3.4. Comparative analysis with other machine learning algorithms

Two other machine learning techniques that are well established in the health community have also been analysed: SVM and LR. SVM outperforms most of the other classification techniques, as does RF [8]. We have used a linear kernel since other kernels (polynomial, radial basis function, etc.) obtained worse results. On the other hand, we use an LR approach, since this is a traditional practice in medicine.

Table 1 shows the results for all of the methods in each of the experimental scenarios. It is possible to observe that RF outperforms all the other methods. The reasons for the bad performance of SVM and LR can be found in the overfitting problem caused by the dimensionality problem [3], but also on the issue that both methods constrain the predictive model to a linear split of the sample space.

On the other hand, the SNP relevance values for SVM and LR regarding the raw data set are shown in Figure 7 (top and bottom respectively). It is possible to observe that the values for both methods are similar, but they differ from the RF (see Figure 4). For example, the sets of weights found by SVM and LR in the different cross-validation subsets have an average Pearson coefficient of 0.90, while the average Pearson coefficient between the RF weights and the absolute values of LR and SVM weights is 0.20 and 0.14, respectively.

Despite the fact that RF does not consider negative relevance values, the distribution of the importance of each SNP is different to that of the other methods (see Figure 8). On the other hand, the sign of the relevance learnt by SVM and LR provides more information regarding the class. That is, positive relevance is related to diabetic persons, while negative relevance is related to healthy persons.

However, depending on the variability of the relevance values obtained, we can observe that SVM and LR incur a high variability, demonstrating a high degree of instability. On the other hand, RF standard deviations are the lowest.

In the other scenarios, the methods show similar results regarding relevance learning.

4. Discussion

RF is a simple method that does not require too many parameters nor underlying assumptions about the domain in order to obtain SNP's relevance while obtaining a high predictive power. Moreover, the accuracy of RF is better than that of other existing machine learning methods.

Two of the important benefits of using the RF method is its capacity to manage missing information, and the stability of the feature relevance set obtained. Missing data is inherent when handling genetic data. On the other hand, stability is important regarding result reproducibility [14].

In this regard, we have explored how the SNP relevance values could be transferred to other decision support tools. In doing so, we have used k-NN

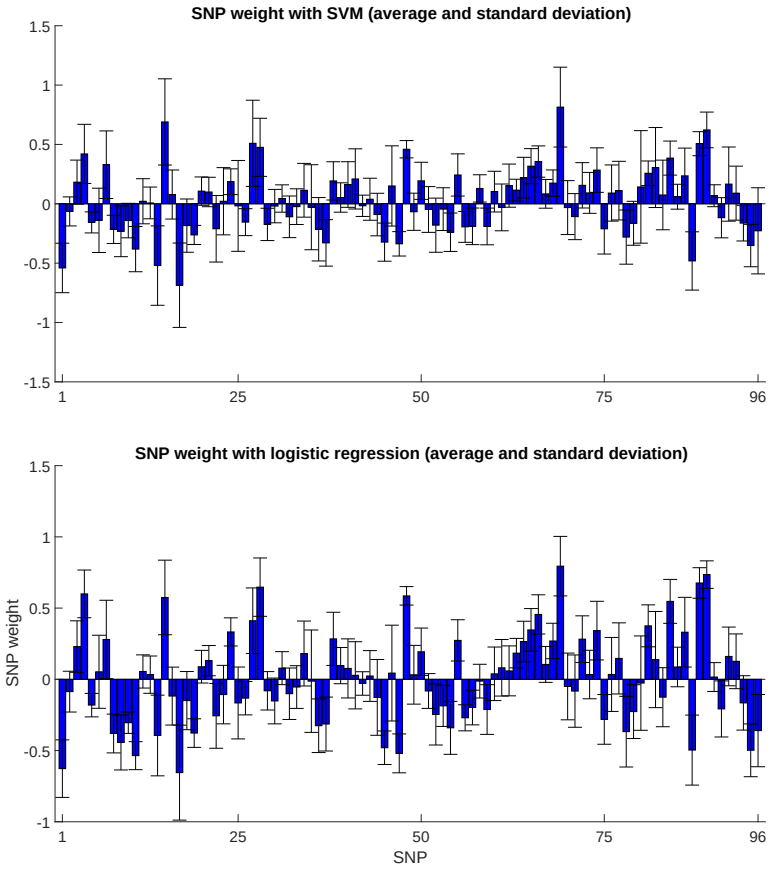

Figure 7: SNP relevance obtained with SVM (top) and LR (bottom)

methods [26], which classify new samples q according to a weighted average similarity measure. The weight of the k-NN method have been set to the SNP's relevance learnt using RF. The accuracy results however, differs from those of the RF ones (83.95%), as they decrease about 5 points down to 78.75%. This means that the aggregation of the Gini importance performed when extracting the SNP's relevance loses some information that is inherent in the RF regarding their predictive power.

On the other hand, some works such as [14] suggests the use of handling different subsets of features in a same domain, instead of a single, averaged feature set. The problem in this scenario will be to decide which context determines

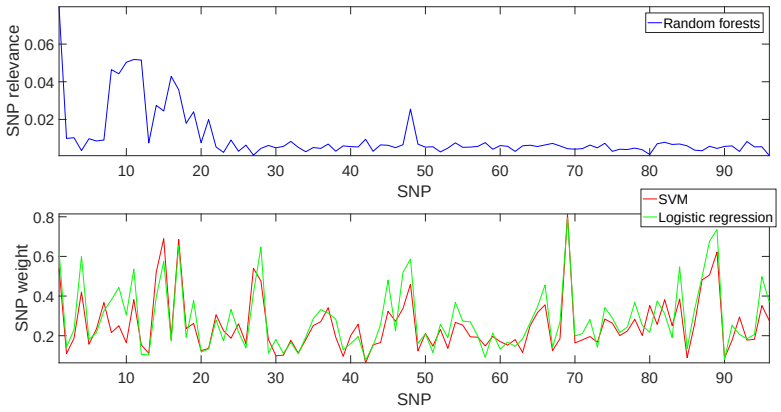

Figure 8: SNP relevance obtained with RF (top) and SVM and LR (bottom). Lines are drown for highlighting the shapes of the relevance learnt.

the identification of one subset or another.

There are several avenues of future work that could improve the RF model presented here. First of all, it could be interesting to quantify the uncertainty of the RF predictor following the work proposed in [10], which could be extended to detect covariate features. Moreover, the sparsity constraint, as described in [27], helps to manage missing information. This constraint allows to define that only some features are informative. It can be used to appropriately discard features instead of the threshold cut applied in this work. Other alternative methods to better manage missing data are proposed in [28], by providing alternative importance measures. Finally, regarding imbalanced data, other measures than AUC are proposed in [29] which are more robust towards class imbalance regarding ensemble importance measures.

5. Conclusions

SNP's data are complex data which presents the dimensionality problem (low samples with regard to the number of SNPs) with usually a lot of missing information. Some machine learning methods suffer the risk of overfitting when trying to build a predictive model. Moreover, they can also suffer from

some instability regarding the SNP's relevance. The RF technique is a machine learning technique that naturally handles the complexity of SNP datasets, while providing SNP's relevance with a certain degree of stability as shown in this paper, with the application of RF to obtain a predictive model for type 2 diabetes prediction and the relevance of the SNPs that conforms to such a model.

Acknowledgements

Funding: This work was supported by the European Unions Horizon 2020 research and innovation programme [grant number 689810, PEPPER]; the University of Girona [grant number MPCUdG2016]; and the Spanish MINECO [grant number DPI2013-47450-C21-R].

Work developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016).

This work was supported by Fondo Europeo de Desarrollo Regional (FEDER) funds.

References

[1] M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics, Nature Reviews Genetics 16 (6) (2015) 321–332. doi:10.1038/nrg3920.

[2] B. A. McKinney, D. M. Reif, M. D. Ritchie, J. H. Moore, Machine learning for detecting gene-gene interactions: a review, Applied bioinformatics 5 (2) (2006) 77–88.

[3] R. Bellazzi, B. Zupan, Towards knowledge-based gene expression data mining., Journal of biomedical informatics 40 (6) (2007) 787–802. doi:10.1016/j.jbi.2007.06.005.

[4] S. Shalev-Shwartz, S. Ben-David, Understanding machine learning: from theory to algorithms, Cambrbidge University Press, 2014.

[5] S. Okser, et al., Regularized Machine Learning in the Genetic Prediction of Complex Traits, PLoS Genetics 10 (11) (2014) e1004754. doi:10.1371/journal.pgen.1004754.

[6] E. Capobianco, Ten challenges for systems medicine, Frontiers in Genetics 3 (2012) 193. doi:10.3389/fgene.2012.00193.

[7] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[8] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, Journal of Machine Learning Research 15 (2014) 3133–3181.

[9] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics 10 (1) (2009) 213. doi:10.1186/1471-2105-10-213.

[10] L. Mentch, G. Hooker, Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, Journal of Machine Learning Research 17 (26) (2016) 1–41.
URL http://jmlr.org/papers/v17/14-168.html

[11] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering 40 (1) (2014) 16–28. doi:10.1016/j.compeleceng.2013.11.024.

[12] L. Molina, L. Belanche, A. Nebot, Feature selection algorithms: a survey and experimental evaluation, in: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., IEEE Comput. Soc, 2002, pp. 306–313. doi:10.1109/ICDM.2002.1183917.

[13] S. Loscalzo, L. Yu, C. Ding, Consensus group stable feature selection, in: Proceedings of the 15th ACM SIGKDD international conference on Knowl-

edge discovery and data mining - KDD '09, ACM Press, New York, New York, USA, 2009, p. 567. doi:10.1145/1557019.1557084.

[14] Z. He, W. Yu, Stable feature selection for biomarker discovery, Computational Biology and Chemistry 34 (4) (2010) 215–225. doi:10.1016/j.compbiolchem.2010.07.002.

[15] B. López, R. Viñas, F. Torrent-Fontbona, J. M. Fernández-Real, Handling Missing Phenotype Data with Random Forests for Diabetes Risk Prognosis, in: Proc. ECAI Workshop on Artificial Intelligence for Diabetes, The Hague, NL, 2016, pp. 39–42.

[16] Z. Yu, D. J. Schaid, Methods to impute missing genotypes for population data, Human Genetics 122 (5) (2007) 495–504. doi:10.1007/s00439-007-0427-y.

[17] M. Saar-Tsechansky, F. Provost, Handling Missing Values when Applying Classification Models, The Journal of Machine Learning Research 8 (2007) 1623–1657.

[18] M. Robnik-Šikonja, Improving random forests, in: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 359–370. doi:10.1007/978-3-540-30115-8_34.

[19] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, Curran Associates Inc., 2013, pp. 431–439.

[20] J. Rogers, S. Gunn, Identifying feature relevance using a random forest, in: Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop, SLSFS 2005, Bohinj, Slovenia, February

23-25, 2005, Revised Selected Papers, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 173–184. doi:10.1007/11752790_12.

[21] S. Li, et al., Random KNN feature selection - a fast and stable alternative to Random Forests, BMC Bioinformatics 12 (1) (2011) 450. doi:10.1186/1471-2105-12-450.

[22] C. N. Hales, D. J. P. Barker, Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis, Diabetologia 35 (7) (1992) 595–601. doi:10.1007/BF00400248.

[23] R.-N. Feng, C. Zhao, C.-H. Sun, Y. Li, Meta-Analysis of TNF 308 G/A Polymorphism and Type 2 Diabetes Mellitus, PLoS ONE 6 (4) (2011) e18480. doi:10.1371/journal.pone.0018480.

[24] Y. Wang, K. Xiang, T. Zheng, W. Jia, K. Shen, J. Li, [The UCSNP44 variation of calpain 10 gene on NIDDM1 locus and its impact on plasma glucose levels in type 2 diabetic patients]., Zhonghua yi xue za zhi 82 (9) (2002) 613–6.

[25] O. Ali, Genetics of type 2 diabetes., World journal of diabetes 4 (4) (2013) 114–23. doi:10.4239/wjd.v4.i4.114.

[26] D. T. Larose, k-Nearest Neighbor Algorithm, in: Discovering Knowledge in Data, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005, pp. 90–106. doi:10.1002/0471687545.ch5.

[27] E. Scornet, G. Biau, J. P. Vert, Consistency of random forests, Annals of StatisticsarXiv:1405.2881, doi:10.1214/15-AOS1321.

[28] A. Hapfelmeier, T. Hothorn, K. Ulm, C. Strobl, A new variable importance measure for random forests with missing data, Statistics and Computing 24 (1). doi:10.1007/s11222-012-9349-1.

[29] S. Janitza, G. Tutz, A. L. Boulesteix, Random forest for ordinal responses: Prediction and variable selection, Computational Statistics and Data Analysisdoi:10.1016/j.csda.2015.10.005.