Multifactor Dimensionality Reduction for the Analysis of Obesity in a Nutrigenetics Context

Katerina Karayianni[1], Ioannis Valavanis[2],
Keith Grimaldi[3], and Konstantina Nikita[1]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str., 15780, Zografos, Athens, Greece
[2] Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Vas. Konstantinou 46, 11635, Athens, Greece
[3] Biomedical Engineering Laboratory, Institute of Communication and Computer Systems, National Technical University of Athens, Polytechniou Str., 15780, Zografos, Athens, Greece
kkarayian@biosim.ntua.gr, ivalavan@eie.gr,
keith.grimaldi@gmail.com, knikita@ece.ntua.gr

Abstract. The current work aims to study within a nutrigenetics context the multifactorial trait beneath obesity. To this end, the use of parallel Multifactor Dimensionality Reduction (pMDR) is investigated towards the identification of i) factors that have an impact to obesity onset solely or interacting with each other and ii) rules that describe the interactions among them. Data have been obtained from a large scale nutrigenetics study and each subject, characterized as normal or overweight based on Body Mass Index (BMI), is featured a 63-dimensional vector describing his/her genetic variations and nutritional habits. pMDR method was used to reduce the initial set of factors into subsets that can classify a subject into either normal or overweight with a certain accuracy and are further used by corresponding prediction models. Results showed that pMDR selected factors associated to obesity and constructed predictive models showing a good generalization ability. Rules describing interactions of the selected factors were extracted, thus enlightening the classification mechanism of the constructed model.

Keywords: nutrigenetics, obesity, Multifactor Dimensionality Reduction, prediction model.

1 Background

Obesity has been found to have a positive relationship with cardiovascular disease (CVD) mortality in various large scale studies [1-2]. Although obesity onsets mostly as a result of certain environmental exposures, e.g. the high in sugars and fats westernized nutrition, or the lack of physical exercise, it can be also studied as an interactive effect among the genetic profile of a person and its exposure to environmental factors. Such interactive effects, studied by the relatively new field of nutrigenetics [3], can provide insights on the development of obesity and may trigger

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 231–238, 2012.

forming new strategies for the control of obesity and CVD, not limited solely to environmental exposures.

The importance of indentifying gene-gene interactions towards the study of diseases has been highlighted in [4], where it is emphasized that epistasis (gene-gene interactions) is ubiquitous in human diseases. Apart from complex gene-gene interactions, which can reveal more biological information than individual gene analysis , various studies have focused on the additional importance of analyzing how gene and environmental factors interact and have an impact towards the development of disease [5]. Various gene-environment interactions that are related to obesity have been identified in [6], where authors conclude the importance of accounting for gene-environment interactions towards the understanding and treatment of obesity. Choosing to study gene-environment interactions in obesity has also be shown to be a justifiable approach in [7], in which the relationship of genetic and environmental factors in a person's BMI was studied, though without reporting over specific genetic variations. The contribution of genes and environmental factors towards BMI has been also studied in [8], where the statistical analysis conducted identified two polymorphisms that in synergy with fats intake contribute to the modulation of waist circumference.

Studies of gene-gene and gene-environment interactions have some inherent difficulties. On one hand, there is a relative difficulty in collecting the appropriate environmental and genetic factors for a significant number of subjects. On the other hand, there is difficulty in analyzing a problem of such complexity due to the high dimensionality of the data [9]. Various advanced computational methods have been proposed and applied to identify interactions among genetic and environmental factors that may trigger perturbations into biological pathways and contribute to the development of diseases [10]. Multifactor Dimensionality Reduction (MDR) is a popular non-parametric, model-free method for detecting gene-gene and gene-environment interactions developed by Ritchie et al. [11]. It has been used for the study of gene-gene interactions in various diseases, e.g. hypertension [12], type 2 diabetes [13] and breast cancer [14].

In the current work, the method chosen to analyze the available data towards the identification of causal interactions beneath obesity is a more recent algorithm developed by Ritchie et al., i.e. parallel Multifactor Dimensionality Reduction (pMDR) [15]. The particular implementation offers various advantages in relation to the prior Dimensionality Reduction method (MDR), as it scales to handle big datasets. In addition, it allows constructing rules that describe the interactions among the selected subset of factors, providing more valuable information that enlightens the interplay of the involving features. This study employs a large scale dataset of more than 2300 subjects and targets i) associations of obesity and interaction rules from a pool of 63 features describing gender, various nutritional elements and genetic variations that have been previously individually associated with obesity and

cardiovascular health [16] and ii) corresponding predictive models that can distinguish subjects characterized as normal or overweight based on Body Mass Index (BMI). Regarding previous findings, it comes to take one step further our previous work in [9], in which ANN-based methods were used select the most informative features from the same nutrigenetics data and construct predictive models for obesity status.

2 Dataset

The dataset employed comes from a previous large scale nutrigenetics study [16]. It includes data for 2341 white people, and for each subject a total of 38 nutrition measurements have been collected, e.g. daily intake of cholesterol, supplements of metals, in addition to the recording for 24 genetic variations (Single Nucleodite Polymorphisms-SNPs, or Insertions/Deletions) that have been found to have an influence on daily requirements of various nutritional elements for improving various CVD health aspects [16]. Nutritional elements, genetic variations and additionally gender complete the set of 63 input factors. Each subject is characterized as overweight or normal, according to his/her BMI (calculated as weight (Kg)/height^2 (m^2)). 1464 subjects were labeled as overweight (BMI > 25), and the remaining 877 as normal (BMI $\leq$ 25). Before analyzing the data by pMDR, it was necessary to do a pre-processing of the data to convert them into categorical values. Nutritional factors from intake of supplements were classified into four classes, which are bottom 33.3%, middle 33.3% and top 33.3% of non-zero values and zero, while the rest of the nutritional factors were classified according to the quartiles. Gene variations are categorical variables by themselves, e.g. a SNP corresponds to a three state categorical feature (AA/GG/AG). For each categorical variable, numerics (e.g. 1,2 and 3 for a three class variable) were used when importing data to pMDR. An extensive description of the dataset used can also be found in [9].

3 Methods

The current work investigates the use of the pMDR algorithm as a computational method to derive prediction models from the factors measured for each subject in the dataset and identify interactions among the elements of the models. The method uses a new algorithm in relation to the previous implementation of MDR that is able to analyze and identify interactions among factors from large datasets. The implementation is done with the Message Passing Interface (MPI) that enables parallel processing into multiple processors, which can make possible the handling high-complexity problems. An additional advantage of the method is that it can extract rules of interaction, capturing how the various combinations among the

categorical values of the reduced features can predict a two-state result. The pMDR algorithm can be used for extremely large datasets of individuals and with many variable states, making feasible the analysis of small order interactions in very large datasets. The analysis of higher-order interactions in large datasets is also feasible, but is demanding in machine computational power and running time [11].

The pMDR algorithm is a non-parametric and model-free method that uses cross-validation to derive results. Firstly, data are divided into the training and testing set The desired number of factors to comprise the reduced model is also specified. For all possible groups with this number of elements from the initial set of factors, the algorithm derives the various combinations of states among them. For example, for the two-factor combinations, the model consisting of a nutritional factor with four categorical states and a gene factor with three different types of variations has twelve different states of combinations. Then, each individual from the subjects is grouped to the combination that matches its characteristics. In this way, it can be calculated for each combination the ratio of cases to controls (in our case, obese to non-obese subjects). Then, this ratio is compared to the general ratio of cases to controls for the whole data set. If it is higher or equal to the general ratio, then the particular combination is characterized as high risk, else as low risk. Given the true labeling of subjects, sensitivity and specificity are calculated for each model and are used to compare the performance amongst the various models. The so-called balanced accuracy is the average of sensitivity and specificity of the model and it is calculated for each model for both the training and testing set in each cross-validation step. In addition, for each model there is an estimation of the prediction error in the testing set, based on the proportion of mislabeled subjects by the model. In the final step of the algorithm, the single best model across all combinations and up to the maximum model order selected, i.e. maximum number of factors included, is selected based on the highest cross-validation consistency and prediction accuracy [15]. pMDR allows the configuration of various execution parameters, which can have an impact on both the robustness of the method as well as the computational cost. For example, increasing the number of cross-validations augments the computational cost, yet having a sufficient number of cross-validations is necessary to ensure the validity of results. pMDR was used here in a cluster of two processing nodes of specification Intel(R) Xeon(TM) CPU 3.00GHz dual-core.

4 Results

In this part we present the results obtained by the pMDR method. The particular execution was configured to include models consisting up to seven factors (maximum order: 7) from the total 63 included in the dataset. In addition, it was selected to perform five-fold cross-validation (at each cross-validation step the best model is also kept in order to measure cross validation consistency, see following paragraph).

Table 1. Selected models (of order:1,..,7), based on five-fold cross-validation. For each model, average predicted balanced accuracy, average prediction error (correspond to measurements in testing sets) and CV consistency are presented.

Factors	Average Prediction Balanced Accuracy (%)	Average Prediction Error (%)	CV consistency
Gender (1)	59.34	43.83	5
Saturated Fat-Food Only, Gender (2)	60.57	37.99	5
Vitamin B6-Food Only, Saturated Fat-Food Only, Gender (3)	61.90	38.30	4
Vitamin B6, Vitamin A, Saturated Fat, Caffeine (4)	52.87	46.49	1
Vitamin C-Food Only, Omega 3, Cholesterol, Caffeine, Calcium (5)	55.27	43.72	3
Vitamin B12-Food Only, Vitamin A-Food Only, Refined Carbohydrate, Folic Acid-Supplement Only, Cruciferous, Caffeine (6)	51.00	47.16	1
Vitamin D - Food Only, Vitamin C-, Food Only Refined Carbohydrate, Omega 3, Cholesterol, Caffeine, Calcium - Supplement Only (7)	53.08	44.53	1

Table 2. Interactions among factors included in the best model (For Saturated Fat – Food Only, 0 corresponds to the lowest intake and 3 to the highest intake)

IF Saturated Fat – Food Only = 0 AND Gender = Male THEN STATUS = Overweight
IF Saturated Fat – Food Only = 0 AND Gender = Female THEN STATUS = Normal
IF Saturated Fat – Food Only = 1 AND Gender = Male THEN STATUS = Overweight
IF Saturated Fat – Food Only = 1 AND Gender = Female THEN STATUS = Normal
IF Saturated Fat – Food Only = 2 AND Gender = Male THEN STATUS = Overweight
IF Saturated Fat – Food Only = 2 AND Gender = Female THEN STATUS = Normal
IF Saturated Fat – Food Only = 3 AND Gender = Male THEN STATUS = Overweight
IF Saturated Fat – Food Only = 3 AND Gender = Female THEN STATUS = Overweight

After completing all cross-validation steps the algorithm evaluates the cross-validation consistency (CV consistency: number of occurrences as best model in the cross-validations) and the average values for balanced accuracy and prediction error for each model. Final results are shown in Table 1. The method identifies the single best model based on the highest CV. In case two or more models have the same CV value the single best model is determined based on the highest average predicted balanced accuracy and least prediction error. The single best model obtained here uses two factors corresponding to saturated fat – food only (factor 20, corresponds to saturated fat contained in food and not in supplements, see [9], [16] as well) and gender (factor 1). pMDR outputs in form of rules the identified interactions among

the factors that comprise each model. For the single best model [Saturated Fat - Food Only, Gender] the resulting rules are shown in Table 2. The combinations of factors states that have not resulted into a classifiable status (Normal/ Overweight) have been omitted.

5 Discussion

The average prediction balanced accuracy obtained in testing sets, which is the selected measure for the evaluation of the models, is for the single best model almost 61%. The average prediction error for the same model is about 38%. In addition, the best model is consistently the best in all five cross-validations steps that took place. These values are satisfactory given that dimensionality reduction analysis has been conducted and that models are accompanied by rules that enlighten the classification mechanism of the selected model.

Interpreting the rules generated by the algorithm for the selected model and reported here, we can derive that high values of saturated fat (categorical value 3) can be associated with obesity in both genders, while lower values (categorical value 1) seem to affect males the most.

Comparing the results of the pMDR algorithm with the previous methods of PDM-ANN and GA-ANN used in [9], it is noted that the balanced accuracy of the best model is comparable to the mean accuracy of training with the ANN-based methods, although these do not match directly, since the methods do not use exactly the same fitness measures. The seven-order model consists of factors that are included in the results of methods PDM-ANN (apart from caffeine) and GA-ANN. In addition, method PDM-ANN (when 5 factors are used) has three factors in common with the 5-dimensional model obtained by pMDR. Thus, there is partial accordance of the pMDR results with the other methods, yet it has to be evaluated for higher-order models to confirm the relevance of results. It's noted here that the 7-order model obtained by pMDR comprised environmental factors only, without any genetic factors being included and did not highlight gene-environment interactions. This has to be further examined by obtaining higher order models by pMDR method, which needs further available computational power (e.g. running on a Grid). On the contrary, the methods used in our previous study [9], stochastic based or serially selecting features, could give higher order models, in which genetic variations were included, too.

The computational needs of the algorithm are high. An increase by one of the number of factors per combination increases disproportionately the necessary running time. When the algorithm was set to run for eight factors the running would took more four weeks to complete in the available cluster machine. Thus, the algorithm seems satisfactory to reveal low order interactions in such a large dataset, while demands very powerful computers to run for higher order models.

Future work shall include using pMDR to find higher-order models together with the relevant rules of interactions. This would enable the comparison of the results with the subset of factors derived from the PDM-ANN and GA-ANN methods in our previous work.. This is a very computationally intensive task and could be become

possible by executing the pMDR algorithm in parallel processes in a high-performance computer cluster or Grid, appropriate for large-scale bioinformatics applications. Future work may also combine the approach of the current work with the previous ones conducted on the same dataset using the ANN-based methods. The most important factors identified by the latter could be passed into pMDR, in order to construct rules of interaction among them and gain more information.

6 Conclusions

In this study, the pMDR algorithm was used to analyze a large set of data from a nutrigenetics study on almost 2300 people, for all of which 63 genetic/nutritional factors, gender and BMI were recorded. The dimensionality reduction that was feasible with the available computing processing power reduced the 63 factors into seven. From the models derived, the one showing the greatest accuracy and least prediction error consisted of two factors, namely saturated fat intake and gender. The information from the remaining higher order models is also informative, as it gives insights about the interactions among them towards the development of obesity. The higher order models showed some consistency in the factors included with the ones identified by the previously ANN-based methods applied to the same dataset.

Acknowledgement. KK would like to thank the A.G. Leventis Foundation.

References

1. McGee, D.: Diverse Populations Collaboration: Body mass index and mortality: A meta-analysis based on person-level data from twenty-six observational studies. Ann. Epidemio. 15, 87–97 (2004)
2. Wilson, P., D'Agostino, R., Sullivan, L., Parise, H., Kannel, W.: Overweight and obesity as determinants of cardiovascular risk: The Framingham experience. Arch. Intern. Med. 162, 1867–1872 (2002)
3. Ordovas, J., Mooser, V.: Nutrigenomics and nutrigenetics. Curr. Opin. Lipidol. 15, 101–108 (2004)
4. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56, 73–82 (2003)
5. Hunter, D.J.: Gene-environment interactions in human diseases. Nature Reviews Genetics 6(4), 287–298 (2005)
6. Andreasen, C.H., Andersen, G.: Gene-environment interactions and obesity–further aspects of genomewide association studies. Nutrition 25(10), 998–1003 (2009)
7. Karnehed, N., Tynelius, P., Heitmann, B.L., Rasmussen, F.: Physical activity, diet and gene-environment interactions in relation to body mass index and waist circumference: the Swedish young male twins study. Public Health Nutr. 9, 851–858 (2006)
8. Robitaille, J., Pérusse, L., Bouchard, C., Vohl, M.C.: Genes, Fat Intake, and Cardiovascular Disease Risk Factors in the Quebec Family Study. Obesity 15, 2336–2347 (2007)

9. Valavanis, I., Mougiakakou, S., Grimaldi, K., Nikita, K.: A multifactorial analysis of obesity as CVD risk factor: Use of neural network based methods in a nutrigenetics context. BMC Bioinformatics 11, 453 (2010)
10. Heidema, A., Boer, J., Nagelkerke, N., Mariman, E., van der, A.D., Feskens, E.: The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet. 7, 23 (2006)
11. Hahn, L., Ritchie, M., Moore, J.: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376–382 (2003)
12. Williams, S., Ritchie, M., Phillips, J., Dawson, E., Prince, M., Dzhura, E., Willis, A., Semenya, A., Summar, M., White, B., Addy, J., Kpodonu, J., Wong, L., Felder, R., Jose, P., Moore, J.: Multilocus analysis of hypertension: a hierarchical approach. Hum. Hered. 57, 28–38 (2004)
13. Cho, Y., Ritchie, M., Moore, J., Park, J., Lee, K., Shin, H., Lee, H., Park, K.: Multifactor-dimensionality reduction shows a two locus interaction associated with Type 2 diabetes mellitus. Diabetologia. 47, 549–554 (2004)
14. Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F., Moore, J.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 128–147 (2001)
15. Bush, W., Dudek, S., Ritchie, M.: Parallel Multifactor Dimensionality Reduction: a tool for the large scale analysis of gene-gene interactions. Bioinformatics 22, 2173–2174 (2006)
16. Arkadianos, I., Valdes, A., Marinos, E., Florou, A., Gill, R., Grimaldi, K.: Improved weight management using genetic information to personalize a calorie controlled diet. Nutr. J. 6, 29 (2007)