A Hybrid Approach for Mining Metabolomic Data

Dhouha Grissa[1,3], Blandine Comte[1], Estelle Pujos-Guillot[2], and Amedeo Napoli[3]

[1] INRA, UMR1019, UNH-MAPPING, F-63000 Clermont-Ferrand, France,
[2] INRA, UMR1019, Plateforme d'Exploration du Métabolisme, F-63000 Clermont-Ferrand, France
[3] LORIA, B.P. 239, F-54506 Vandoeuvre-lès-Nancy, France

Abstract. In this paper, we introduce a hybrid approach for analyzing metabolomic data about the so-called "diabetes of type 2". The identification of biomarkers which are witness of the disease is very important and can be guided by data mining methods. The data to be analyzed are massive and complex and are organized around a small set of individuals and a large set of variables (attributes). In this study, we based our experiments on a combination of efficient numerical supervised methods, namely Support Vector Machines (SVM), Random Forests (RF), and ANOVA, and a symbolic non supervised method, namely Formal Concept Analysis (FCA). The data mining strategy is based on ten specific classification processes which are organized around three main operations, filtering, feature selection, and post-processing. The numerical methods are mainly used in filtering and feature selection while FCA is mainly used for visualization and interpretation purposes. The first results are encouraging and show that the present strategy is well-adapted to the mining of such complex biological data and the identification of potential predictive biomarkers.

Keywords: hybrid knowledge discovery, Random Forest, SVM, ANOVA, Formal Concept Analysis, feature selection, discrimination, predication, visualization

1 Introduction

Metabolomics allows the analysis of a biological system by measuring metabolites, i.e. small molecules, present and accessible in the system. Usually different techniques are necessary for such an analysis. In particular, one challenge of metabolomics is to identify, among thousands of features, predictive biomarkers, i.e. a measurable indicator of some biological status, of a disease development [7]. This can be viewed as a hard data mining task as data generated by metabolomic platforms are massive, complex and noisy. In the present study, we aim at identifying predictive metabolic biomarkers of future T2D –type 2 diabetes– development, a few years before occurrence, in an homogeneous population considered

as healthy at the time of the analysis. The datasets include a limited number of individuals, a large number of variables or attributes, e.g. molecules or fragments of molecules, and one binary target variable, i.e. developing or not the disease a few years after the analysis.

One important problem here is to distinguish between discriminant and predictive features. A feature is said to be "discriminant" if it separates individuals in distinct classes, such as for example healthy vs not healthy. A feature is said to be "predictive" if it enables predicting the evolution of individuals towards the disease a few years later. However, the most discriminant features are not necessarily the best predictive features. Thus, it is essential to compare different feature selection methods and to evaluate their capabilities to select relevant features for further use in prediction.

Accordingly, we propose a knowledge discovery process which is based on a combination of numeric-symbolic techniques for different purposes, such as noise filtration for avoiding overfitting –which occurs when the analysis describes random error or noise instead of the underlying relationships–, feature selection for reducing dimension, and checking the relevance of selected features w.r.t. prediction. FCA [3] is then applied to the resulting reduced dataset for visualization and interpretation purposes. More precisely, this hybrid data mining process combines FCA with several numerical classifiers including Random Forest (RF) [1], Support Vector Machine (SVM) [8], and Analysis of Variance (ANOVA) [2]. RF, SVM and ANOVA are used to discover discriminant biological patterns which are then organized and visualized thanks to FCA.

Actually, the initial problem statement is based on a data table *individuals* $\times$ *features*. The data preparation step involves filtering methods based on the correlation coefficient ("Cor") and mutual information ("MI") to eliminate redundant and dependent features, to reduce the dimensions of the data table and to prepare the application of RF, SVM and ANOVA. The initial data table *individuals* $\times$ *features* is transformed into a binary data table *features* $\times$ *classification-process* in the following way. Ten different classifications processes (CPs hereafter) are defined and applied to the initial data table. Every CP provides a ranking of features. Then, the N best classified features are kept for being processed by FCA. Actually, a feature is selected when it is ranked among the six first features. This leads us to a selection of $N = 48$ features and to a binary data table, 48-*features* $\times$ 10-*CPs*, which is in turn considered as a formal context for the application of FCA and the construction of concept lattices. These $N = 48$ features are shared by all CPs and are interpreted as potential biomarkers of disease development.

Meanwhile, biological experts want to classify the selected features as potential predictive biomarkers, i.e. biomarkers able to predict the disease development a few years before occurrence. Predictive biomarkers can be detected thanks to ROC curves [6]. In the current study, such an analysis produces a short list of the best predictive features which are selected as a core set of biomarkers. Finally, FCA is used again to build the best ranking within this core set of biomarkers and for visualization and interpretation purposes. This is one originality of

this paper to present a combination of numerical data mining methods based on RF and SVM with FCA, which in turn is mainly used for interpretation and visualization.

The paper is organized as follows. Section 2 presents preparation and mining of the data for discovering the potential predictive features. Then Section 3 describes experiments performed on real-world metabolomic data set. A discussion and a conclusion complete the paper.

2 The preparation and the mining of metabolomic data

All experiments in the following were carried out on a Dell machine running Ubuntu GNU/Linux 14.04 LTS, a 3.60 GHZ $\times$ 8 CPU and 16 GB RAM. The data analysis methods are taken from the RStudio software environment (Version 0.98.1103, R 3.1.1). Rstudio is freely available and offers a selection of packages suitable for many different types of data[1].

2.1 The dataset description

The dataset which is analyzed is based on a case-control study from the GAZEL French population-based cohort (20000 subjects). This set includes numeric and symbolic data about 111 male subjects (54-64 years old) free of T2D at baseline. 55 subjects developed T2D at the follow-up belong to class "1" (non healthy or diabetic subjects) while 56 subjects belong to class "-1" (controls or healthy subjects). Three thousand features are generated after carrying out mass spectrometry (MS) analysis. After noise filtration, 1195 features are remaining for describing every subject.

The reference dataset is composed of homogeneous individuals considered healthy at the beginning of the study. The binary variable describing the two target classes, i.e. healthy and not healthy, is based on the health status of the same individuals at another time, actually five years after the initial analysis. Meanwhile, some individuals developed the disease. Thus, discriminant features which enable a good separation between target data classes (healthy vs. not healthy) are not necessarily the best features predicting the disease development five years later.

2.2 Data preprocessing

Only a few features allow a good separation between the target classes. Therefore, it is necessary to reduce data dimension to select a small number of relevant features for further use in prediction. Reducing the dimensionality of the data is a challenging step, requiring a prior filtering of the initial data. Metabolomic data contain highly correlated features, which can have an impact on feature selection and data mining [4]. Thus, two filtering methods are chosen, namely the

[1] https://www.rstudio.com/

coefficient of correlation ("Cor") and mutual information ("MI"). Both filtering methods are used to discard correlated features and dependent features.

Figure 1 describes the global classification workflow. At the beginning, the filtering methods "Cor" and "MI" eliminate highly correlated features. Afterward, two reduced subsets are generated: a first subset contains 963 features (after "Cor" filtering) while the second subset contains 590 features (after "MI" filtering). Both reduced subsets are used as input for the application of RF and SVM classifiers.

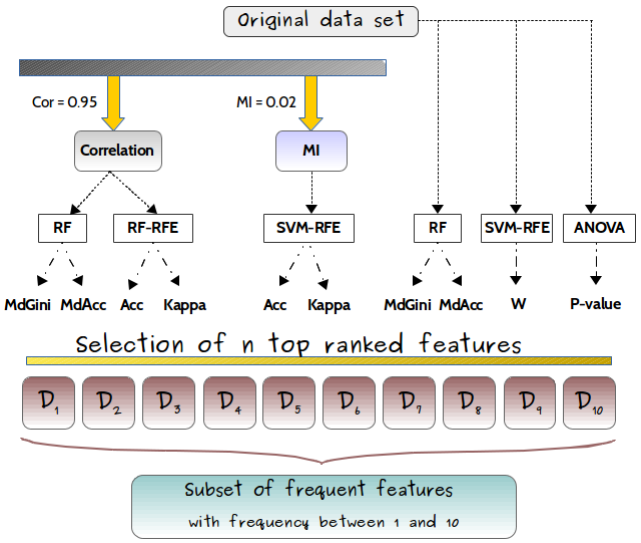

Fig. 1. Feature selection and dimensionality reduction process.

2.3 Feature selection

Two main classification techniques are then applied to the resulting filtered subsets of features, namely RF and SVM. Moreover, for improving the process, RF and SVM are combined with RFE ("Recursive Feature Elimination"), which is a backward elimination method used for feature selection [5]. Finally, three different classification processes are defined: (i) RF applied to data filtered with "Cor", (ii) RF+RFE applied to data filtered with "Cor", (iii) SVM+RFE applied to data filtered with "MI".

In parallel, we apply the the ANOVA method directly on the original dataset as this is a common practice in metabolomics (without any filtering process). This time, SVM+RFE, RF and ANOVA are directly applied to the original dataset.

The measure of the importance of each selected feature in the output of the classification process is the purpose of post-processing. There, several measures of interest (accuracy metrics) enable the ranking of the features, namely MdGini, MdAcc, Accuracy and Kappa. MdGini stands for "Mean decrease in Gini index" is used as an impurity function. MdAcc stands for "Mean decrease in accuracy" and measures the importance/performance of each feature in the classification. Kappa is a statistical measure comparing observed accuracy with expected accuracy. The general idea about the use of these metrics is to measure the decrease in accuracy after permutation of the values of each variable. The scores given by these metrics allow to rank the features (highest discriminative power) within each classification process.

On the same basis, when no filtering is applied, post-processing is based on MdGini and MdAcc for RF, on the weight magnitude of features "W" for SVM+REF, and on the "p-value" for ANOVA. The p-value determines the statistical significance of the results when a hypothesis test is performed.

Based on these different processes, various forms of results, e.g. feature ranking and feature weighting, and as well multiple sets of ranked features are produced. Actually, 10 sets are generated, corresponding to the different CPs and ranking scores. They are denoted by $D_i, i = 1, \ldots, 10$ in Figure 1.

For each classification process, a corresponding name is created which describe the set of operations the process is based on. We have the ten following processes: (1) "Cor-RF-MdAcc", i.e. filtering with "Cor", feature selection with RF and post-processing with MdAcc, (2) "Cor-RF-MdGini", (3) "Cor-RF-RFE-Acc", (4) "Cor-RF-RFE-Kap", (5) "MI-SVM-RFE-Acc", (6) "MI-SVM-RFE-Kap", (7) "RF-MdAcc", (8) "RF-MdGini", (9) "SVM-RFE-W" and finally (10) "ANOVA-pValue".

To select the most important features, we retain the 200 first ranked features, except for "ANOVA-pValue" where we only selected 107 features with a reasonable p-value for filtering purposes (lower than 0.1). Finally, ten reduced sets of ranked features, i.e. $D_i, i = 1, \ldots, 10$, are obtained and should be analyzed for discovering the "best features". Then, the visualization of these "best features" is carried out thanks to FCA.

2.4 Visualization and interpretation with FCA

In this section, we show how to compare the highly ranked features in the reduced subsets $D_i, i = 1, \ldots, 10$. For this purpose, a binary data table *features* $\times$ *CPs* is built (see Table 1), where objects in rows correspond to features and attributes in columns correspond to the 10 classification processes (CPs). The presence of 1 in a cell of the data table means that the feature in the row is ranked for the CP in the column. Every feature has a support, i.e. the number of 1 in the row, which should be at least of 6/ This means that every feature appears among the best ranked features with a frequency between 6 and 10. This leads us to consider $N = 48$ such features. The new binary data table 48-*features* $\times$ 10-*CPs* is presented in Table 1. Starting from the initial data table 111-*individuals* $\times$

1195-*features* we finally get a binary data table 48-*features* $\times$ 10-*CPs*. The "m/z" label of features stands for "mass per charge".

Applying FCA on the 48-*features* $\times$ 10-*CPs* data table considered as a context produces a concept lattice with 272 concepts (Figure 2). This concept lattice illustrates the combination of numerical classification methods with FCA, and allow an interpretation of the relations between features and classification processes, and further on the discriminative and predictive powers of the features. Four features "m/z 383", "m/z 227", "m/z 114" and "m/z 165" have a maximal support of 10 (see the maximum rectangle full of 1 in Table 1). There are strong relationships among the 44 remaining features, especially involving "m/z 284", "m/z 204", "m/z 132", "m/z 187", "m/z 219", "m/z 203", "m/z 109", "m/z 97" and "m/z 145". Moreover, among the 48 features, 39 are significant w.r.t. ANOVA (with a p-value < 0.05).

The concept lattice highlights the potential of the feature selection approach for analyzing metabolomic data. It enables discriminating direct and indirect associations, e.g. highly linked metabolites belonging to the same concept. The links between the concepts in the lattice can be interpreted as interdependency between concept and metabolites.

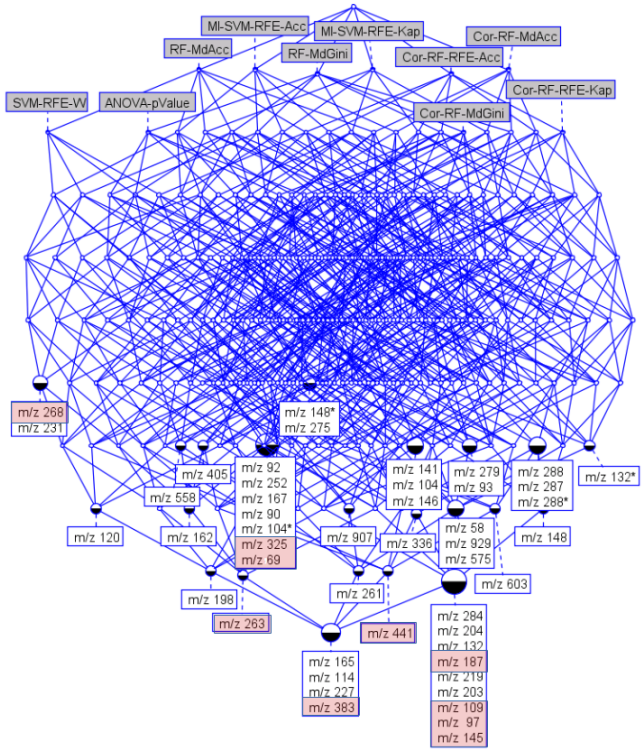

Fig. 2. The concept lattice derived from the 48 $\times$ 10 binary table (Table 1).

Table 1. Input binary table describing the 48 frequent features with the 10 CPs.

Features	Cor-RF-MdGini	Cor-RF-MdAcc	Cor-RF-RFE-Acc	Cor-RF-RFE-Kap	RF-MdGini	RF-MdAcc	MI-SVM-RFE-Acc	MI-SVM-RFE-Kap	SVM-RFE-W	ANOVA-pValue
m/z 383	1	1	1	1	1	1	1	1	1	1
m/z 227	1	1	1	1	1	1	1	1	1	1
m/z 114	1	1	1	1	1	1	1	1	1	1
m/z 165	1	1	1	1	1	1	1	1	1	1
m/z 145	1	1	1	1	1	1	1	1		1
m/z 97	1	1	1	1	1	1	1	1		1
m/z 441	1	1	1	1	1	1	1		1	1
m/z 109	1	1	1	1	1	1	1	1		1
m/z 203	1	1	1	1	1	1	1	1		1
m/z 219	1	1	1	1	1	1	1	1		1
m/z 198	1		1	1	1	1	1	1	1	1
m/z 263	1	1	1		1	1	1	1	1	1
m/z 187	1	1	1	1	1	1	1	1		1
m/z 132	1	1	1	1	1	1	1	1		1
m/z 204	1	1	1	1	1	1	1	1		1
m/z 261	1	1	1	1	1	1	1	1	1	
m/z 162	1	1			1	1	1	1	1	1
m/z 284	1	1	1	1	1	1	1	1		1
m/z 603	1	1	1	1	1	1			1	1
m/z 148	1	1	1	1	1	1		1		1
m/z 575	1	1	1	1	1	1	1	1		
m/z 69	1	1			1	1	1	1		1
m/z 325	1	1			1	1	1	1		1
m/z 405	1	1			1	1	1		1	1
m/z 929	1	1	1	1	1	1	1	1		
m/z 58	1	1	1	1	1	1	1	1		
m/z 336	1	1	1	1	1	1	1			1
m/z 146	1		1	1	1		1	1		1
m/z 104	1		1	1	1		1	1		1
m/z 120			1	1	1	1	1	1	1	1
m/z 558	1	1			1	1	1	1	1	
m/z 231					1	1	1	1	1	1
m/z 132*	1	1	1	1	1		1	1		
m/z 93		1	1	1	1	1	1	1		
m/z 907		1	1	1		1	1	1	1	
m/z 279		1	1	1	1	1	1	1		
m/z 104*	1	1			1	1	1	1		1
m/z 90	1	1			1	1	1	1		1
m/z 268					1	1	1	1	1	1
m/z 288*	1	1	1	1	1	1				1
m/z 287	1	1	1	1	1	1				1
m/z 167	1	1			1	1	1	1		1
m/z 288	1	1	1	1	1	1				1
m/z 252	1	1			1	1	1	1		1
m/z 141	1		1	1	1		1	1		1
m/z 275			1	1	1		1	1		1
m/z 148*			1	1	1		1	1		1
m/z 92	1	1			1	1	1	1		1

3 Evaluation and discussion

Considering the 48 most frequent features previously identified, we evaluate their predictive capabilities using ROC curves (Figure 3). This analysis was carried out

using the ROCCET tool (`http://www.roccet.ca`), with calculation of the area under the curve ("AUC") and confidence intervals (CI), calculation of the true positive rate (TPR), where $TPR = TP/(TP+FN)$, and the false discovery rate (FDR), where $FDR = TN/(TN+FP)$. The p-values of these relevant features are also computed using t-test.

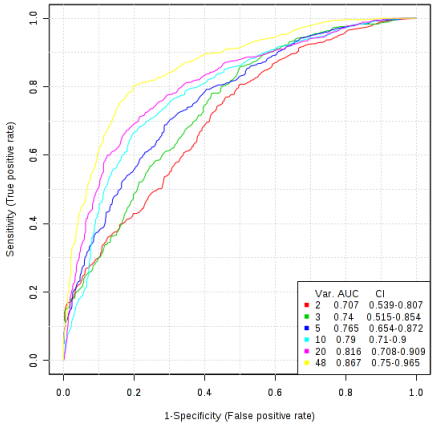

Fig. 3. The ROC curves of at least 2 and at most 48 combined frequent features based on RF model and AUC ranking.

The analysis based on ROC curves is considered as being one of the most objective and statistically valid method for biomarker performance evaluation [6]. ROC curves are commonly used to evaluate the prediction performance of a set of features, or their accuracy to discriminate diseased cases from normal cases.

Since the number of features to propose as predictive biomarkers should be rather small (because of clinical constraints), we rely on the ROC curves of 2, 3, 5, 10, 20 and 48 of features ranked w.r.t. AUC values. The ROC curves enable identifying this best combination of predictive features. Figure 3 shows that the best performance is given to the 48 features all together (with $AUC = 0.867$). But a predictive model with 48 features is not usable for clinical purposes. The set of best features with the smallest p-values and the highest accuracy values is selected and yields a short list of "potential biomarkers". For the ten first features in Table 2, we have $AUC = 0.79$ and $CI = 0.71 - 0.9$. For the four first features, we have $AUC = 0.75$. These high AUC values are witness of a good predictive behavior.

Then we selected as "potential biomarkers" the 10 first features with an AUC greater than 0.74 and significantly small t-test values ($< 10E-5$) (Table 2). We compare this subset with the four most frequent features (features whose frequency is 10 in Table 1) and we find only one feature in common, namely "m/z

383". This confirms that the most frequent features are not the best predictive ones, as biologically suspected, because the metabolomic analysis is performed 5 years before disease occurrence. Moreover, these best "predictive features" or "potential biomarkers" are not belonging to the same concept.

Figure 2 shows that the best predictive biomarkers are lying in different concepts, depicted by red squares in the lattice. For example, the features "m/z 145", "m/z 97", "m/z 109" and "m/z 187" are in the extent of a concept whose intent includes all CPs but "SVM-RFE-W". By contrast, the feature "m/z 268" belongs to another concept whose intent includes 6 CPs, namely "RF-MdGini", "RF-MdAcc", "MI-SVM-RFE-Acc", "MI-SVM-RFE-Kap", "SVM-RFE-W", "ANOVA-pValue". Here again, the direct visualization through the concept lattice shows the position of the predictive features among the discriminant ones and their associations with CPs. This information is very interesting for the domain experts for choosing the best combinations of feature selection methods that can identify the predictive biomarkers.

Name	AUC	T-tests
m/z 145	0.79	1.4483E-6
m/z 383	0.79	5.0394E-7
m/z 97	0.78	1.5972E-6
m/z 325	0.77	2.2332E-5
m/z 69	0.76	1.2361E-5
m/z 268	0.75	4.564E-6
m/z 441	0.75	9.0409E-5
m/z 263	0.75	5.996E-6
m/z 187	0.74	9.0708E-6
m/z 109	0.74	2.6369E-5

Table 2. Table of performance of the best 10 AUC ranked features.

4 Conclusion and future work

In this paper, we presented a hybrid approach for the identification of predictive biomarkers from complex metabolomic dataset. The nature of metabolomic data, i.e. highly correlated and noisy, leads us to build and analyze reduced datasets for identifying important features to be interpreted as potential biomarkers. Moreover, the present hybrid approach is based on an original combination of numerical supervised classification methods (mainly RF, SVM and ANOVA) and a symbolic unsupervised method such as FCA. This study shows the interest of such a combination to reveal hidden information in such high dimensional datasets and how FCA can be used for visualization and interpretation purposes. Based on the resulting lattice, experts in biology will be able to lead a deeper interpretation. Finally, additional experiments on different metabolomic datasets are required to confirm the success of this hybrid approach.

References

1. L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.
2. H. W. Cho, S. B. Kim, and M. K. Jeong et al. Discovery of metabolite features for the modelling and analysis of high-resolution nmr spectra. *International Journal of Data Mining and Bioinformatics*, 2(2):176–192, 2008.
3. B. Ganter and R. Wille. *Formal concept analysis - mathematical foundations.* Springer, 1999.
4. PS. Gromski, H. Muhamadali, DI. Ellis, Y. Xu, E. Correa, ML. Turner, and R. Goodacre. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Anal Chim Acta.*, 879:10–23, 2015.
5. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46:389–422, 2002.
6. Xia J, Broadhurst DI, Wilson M, and Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9(2):280–99, 2013.
7. M. Mamas, WB. Dunn, L. Neyses, and R. Goodacre. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch Toxicol.*, 85(1):5–17, 2011.
8. V.N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, John Willey & Sons, 1998.