A Hybrid Data Mining Approach for the Identification of Biomarkers in 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 an approach for analyzing complex biological data obtained from metabolomic analytical platforms. Such platforms generate massive and complex data that need appropriate methods for discovering meaningful biological information. The datasets to analyze consist in a limited set of individuals and a large set of attributes (variables). In this study, we are interested in mining metabolomic data to identify predictive biomarkers of metabolic diseases, such as type 2 diabetes. Our experiments show that a combination of numerical methods, e.g. SVM, Random Forests (RF), and ANOVA, with a symbolic method such as FCA, can be successfully used for discovering the best combination of predictive features. Our results show that RF and ANOVA seem to be the best suited methods for feature selection and discovery. We then use FCA for visualizing the markers in a suggestive and interpretable concept lattice. The outputs of our experiments consist in a short list of the 10 best potential predictive biomarkers.

Keywords: hybrid knowledge discovery, random forest, SVM, ANOVA, formal concept analysis, feature selection, biological data analysis, lattice-based visualization

1 Introduction

In the analysis of biological data, one of the challenges of metabolomics[1] is to identify, among thousands of features, predictive biomarkers[2] of disease development [13]. However, such a mining task is difficult as data generated by metabolomic platforms are massive, complex and noisy. In the current study,

[1] Metabolomics is the characterization of a biological system by the simultaneous measurement of metabolites (small molecules) present in the system and accessible for analysis. Data obtained are provided from different techniques and different analytical instruments.

[2] A biomarker, or biological marker, generally refers to a measurable indicator of some biological status or condition.

ISBN 978-5-600-01454-1, National Research University Higher School of Economics, 2016.

we aim at identifying from a large metabolomic dataset, predictive metabolic biomarkers of future T2D (type 2 diabetes) development, a few years before occurrence, in an homogeneous population considered healthy at the time of the analysis. The datasets include a rather limited number of individuals and a quite large set of variables. Specific data processing is required, e.g., feature selection. Accordingly, we propose a knowledge discovery process based on data mining methods for biomarker discovery from metabolomic data. The approach focuses on evaluating a combination of numeric-symbolic techniques for feature selection and evaluates their capacity to select relevant features for further use in predictive models. Actually, we need to apply feature selection for reducing dimension and avoid over-fitting[3]. The resulting reduced dataset is then used as a context for applying FCA [5] for visualization and interpretation. More precisely, we develop a hybrid data mining process which combines FCA with several numerical classifiers including Random Forest (RF) [3], Support Vector Machine (SVM) [16], and the Analysis of Variance (ANOVA) [4]. The dataset relies on a large number of numerical variables, e.g. molecules or fragments of molecules, a limited numbers of individuals, and one binary target variable, i.e. developing or not the disease a few years after the analysis. RF, SVM and ANOVA are used to discover discriminant biological patterns which are then organized and visualized thanks to FCA. Because it is known that the most discriminant[4] features may not be necessarily the best predictive[5] ones, it is essential to be able to compare different feature selection methods and to evaluate their capacity to select relevant features for further use in predictive models. The initial problem statement based on a data table of $individuals \times features$ is transformed into a binary table $features \times classification\ process$. Data preparation for feature selection is carried out using filter methods based on the correlation coefficient and mutual information to eliminate redundant/dependent features, to reduce the size of the data table and to prepare the application of RF, SVM and ANOVA.

A comparative study of the best k features from the combination of these different classification process (CP) –10 combinations of CP are considered– is performed. Then a binary data table is built consisting of $N\ features \times 10\ CP$. This binary table is considered as a formal context and as a starting point for the application of FCA and the construction of concept lattices. The features shared by all CP combinations can be interpreted as potential biomarkers of disease development. However, it is essential for biological experts to evaluate and compute the performances of the proposed biomarkers in models predicting the disease development a few years before occurrence. The performance of prediction models can be assessed using different methods. One classical method used by biologists for binary outcomes is the receiver operating characteristic

[3] The problem of over-fitting occurs when a statistical model describes random error or noise instead of the underlying relationship.

[4] A feature is said to be discriminant if it separates individuals in distinct classes (as, healthy vs not healthy).

[5] A feature is said to be predictive if it enables predicting the evolution of individuals towards the disease a few years later.

(ROC) curve [11], where the TPR (True positive rate) is plotted in function of the FDR (False discovery rate) for different cut-off points. A short list of the best predictive features is selected as the core set of biomarkers. Based on this selection, FCA is used to identify the top list of feature selection methods that provide the best ranking of these core set of biomarkers. This additional visualisation is essential for experts to discover the few best predictive biomarkers from the massive metabolomic dataset.

The remainder of this paper is organized as follows. Section 2 provides a description of related works. Section 3 presents the proposed approach and explains the methodological analysis of biomarker identification. Section 4 describes the experiments performed on a real-world metabolomic data set and discusses the results, while section 5 concludes the paper.

2 State of the art

In [14], the authors discuss the main research topics related to FCA and focus on works using FCA for knowledge discovery and ontology engineering in various application domains, such as text mining and web mining. They also discuss recent papers on applying FCA in bio-informatics, chemistry and medicine. Bartel et al. [1] are one of the first papers which apply FCA in chemistry. They use FCA to analyze the structure-activity relationships to predict the toxicity of chemical compounds. Gebert et al. [6] use an FCA-based model to identify combinatorial biomarkers of breast cancer from gene expression values. Since, the structure of gene expression data (GED) differs from metabolomic data, we can approve according to literature that FCA is never applied on metabolomic data. Indeed, the GED data tables include genes which are more or less expressed. Each gene is represented by a vector of values that explain the relative expression of the gene. This is totally different from metabolomic data where input data tables contain samples in rows and thousands of metabolites (small molecules) or feature in columns expressed as signal intensities. The goal is to identify metabolites that predict the evolution towards a clinical outcome. The processing of such metabolomic data is usually performed within different supervised learning techniques, such as PLS-DA (partial least squares discriminant analysis), PC-DFA (Principal component discriminant function analysis), LDA (Linear discriminant analysis), RF and SVM. Standard univariate statistical methodologies (as ANOVA or Student's t-test[6]) are also frequently used to analyze the metabolomic data [10]. In [8], authors show that there is no universal choice of method which is superior in all cases, even if they show that PLS-DA methods outperform the other approaches in terms of feature selection and classification. In a more detailed study [7], authors compare different variable selection approaches (LDA, PLS-DA with Variable Importance in Projection

[6] t-test or Student's t-test is a statistical hypothesis test which can be used to determine if two sets of data are significantly different from each other. If the p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis is rejected in favor of the alternative hypothesis.

(VIP), SVM-Recursive Feature Elimination (RFE), RF with Accuracy and Gini scores) in order to identify which of these methods are ideally suited to analyze a common set of metabolomic data, capable of classifying the Gram-positive bacteria Bacillus. They conclude that RF with its feature ranking techniques (mean decrease gini/accuracy) and SVM combined with SVM-RFE [9] as a variable selection method display the best results in comparison to other approaches. All these studies show that the choice of the appropriate algorithms is highly dependent on the dataset characteristics and the objective of the data mining process. In the field of biomarker discovery, SVM and RF algorithms prove to be robust for extracting relevant chemical and biological knowledge from complex data, in particular in metabolomics [7]. RF is a highly accurate classifier, based on a robust model to outlier detection (a sample point that is distant from other samples). Its main advantage [2] includes essentially its power to deal with overfitting and missing data, as well as its capacity to handle large datasets without variable elimination in terms of feature selection. Nevertheless, it generates unstable and volatile results, contrary to SVM which delivers a unique solution. These alternative approaches may be useful for data dimensionality reduction and feature selection purposes, and may be suitable to combine with FCA.

3 Design approach for Metabolomic data analysis

In this study, we design a hybrid data mining strategy based on the combination of numerical classifiers including RF, SVM, the univariate analysis ANOVA with the symbolic method FCA, to discover the best combination of biological features. In this work, we aim to find, from a large dataset, predictive metabolomic biomarkers of future T2D development.

We evaluate the proposed approach from a performance point of view. For this, we use Dell machine with ubuntu 14.04 LTS, a 3.60 GHZ $\times$8 CPU and 15,6 GBi RAM. We perform all data analyses using the RStudio software (Version 0.98.1103, R 3.1.1) environment. Rstudio is available for free and offers a selection of packages suitable for different types of data.

3.1 Dataset description and pre-processing

Dataset description: we use a biological data set obtained from a case-control study within the GAZEL French population-based cohort (20 000 subjects). The data set includes the measurements (signal intensities) of 111 male subjects (54-64 years old) free of T2D at baseline. It consists in continuous numerical (semi quantitative) data which represent measurements performed on for each individual. Cases (55 subjects) who developed T2D at the follow-up belong to class '1' (diabetes) and are compared to Controls (56 subjects) which belong to class '-1' (healthy controls). A total of about three thousand features is generated after carrying out mass spectrometry (MS) analysis. But after noise filtration, each subject is described by 1195 features. In the rest of this paper, we consider this new filtered dataset of 1195 features, the original dataset.

The obtained dataset is then the result of an analysis performed on homogeneous individuals considered healthy at that time. However, the binary target variable describing the data classes is introduced based on the health status of the same individuals five years after the first analysis. Some of these individuals developed the disease at the follow-up. For this reason, we can not consider the discriminant features as the predictive ones, since features enabling a good separation between data classes (healthy vs not healthy) are not necessarily the same that predict the disease development a few years later.

Data pre-processing: the metabolomic database contains thousands of features with a wide intensity value range. A data preprocessing step is mandatory for adjusting the importance weights allocated to the features. Thus, before applying any FS method, except ANOVA, data are transformed using a Unit-Variance scaling method. It divides each feature value by its standard deviation; so that all features have the same chance to contribute to the model as they have an equal unit variance. The transformed dataset of 1195 features is used as input for all FS methods, except for ANOVA.

3.2 Feature selection for data dimensionality reduction

Only a few features (a small part of the original dataset) allow a good separation between data classes. Therefore, it is necessary to reduce data dimension to select a small number of relevant features for further use in predictive models. Reducing the dimensionality of the data is a challenging step, requiring a careful choice of appropriate feature selection techniques [15]. Filter and embedded methods are used for this purpose. We discarded wrapper approaches since they are greedy in computational cost.

The metabolomic data contain highly correlated features, which may impact the calculation of feature importance and ranking features [8]. To overcome this problem, we use two filter methods, the coefficient of correlation (Cor) and mutual information (MI). The first filter (Cor) is used to discard very highly correlated features, and the second filter (MI) is used to remove very dependent features. As embedded methods [12], we retain two FS techniques that are widely used on biological data, which are RF and SVM.

Figure 1 describes the feature selection workflow we propose to obtain a reduced set of relevant features. This workflow considers at the beginning the filter methods 'Cor' and 'MI' to eliminate redundant/dependent features. In order to limit the loss of information, very highly correlated features are discarded (one feature per group of correlated ones is kept) to keep a reasonable number of features to work with. All the features whose MI average values are smaller than the threshold are selected, since it is known that high mutual information is indicating a large reduction of uncertainty [17]. We then set correlation and mutual information thresholds to 0.95 and 0.02, respectively. Consequently, two reduced subsets are generated: the first subset contains 963 features after 'Cor' filter, and the second one contains 590 features after 'MI' filter. When we fix

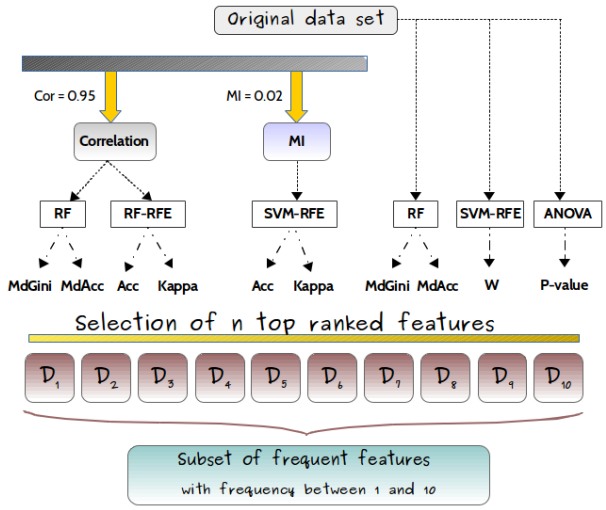

Fig. 1. Feature selection and dimensionality reduction process.

a lower threshold of correlation, we remove a lot of features since the original dataset is very correlated. When we set the MI threshold to a lower value, we keep only a small number of features and consequently we may loose a lot of information.

Both reduced subsets are used as input for the application of RF and SVM classifiers. Nonetheless, as correlation values between variables are still high, we furthermore adapt the RFE[7] approach with RF and SVM. To cover various possible classification results, we apply the embedded methods RF, RF-RFE and SVM-RFE on both filtered subsets. We also apply the ANOVA method on the original data set (not transformed) since it is commonly applied on metabolomic data. Three different classification models are respectively obtained. The first model is built from the application of RF on data filtered with Cor. The second classification model is fitted according to RF-RFE also on the subset of data filtered with 'Cor'. The third model is built from the application of SVM-RFE on the subset of data filtered with 'MI'. Based on these three classification models, we use several accuracy metrics to measure the importance of each feature in the overall result. These measures include MdGini[8], MdAcc[9], Accuracy, and

[7] Recursive Feature Elimination (RFE) is a backward elimination method, originally proposed by Guyon et al. [9] for binary classification. This is one of the classical embedded methods for feature selection with SVM.

[8] Mean decrease in Gini index (MdGini) provides a measure of the internal structure of the data.

[9] Mean decrease in accuracy (MdAcc) measures the importance/performance of each feature to the classification. The general idea of these metrics is to permute the values of each variable and measure the decrease in the accuracy of the model.

Kappa[10]. The scores given by these metrics enable ranking the features by means of the classification models already built.

When no filter is used, three feature selection techniques (SVM-RFE, RF and ANOVA) are applied directly to the original dataset using the feature weight values 'W' (i.e. the weight magnitude of features), p-value[11], MdGini and MdAcc scores to sort the features and identify those with the highest discriminative power. Various forms of results (feature ranking, feature weighting, etc.) and multiple (sub)sets of ranked features are obtained as output. In total, 10 (sub)sets are generated, corresponding to the different CP and ranking scores (Figure 1). For each CP, we give a corresponding name that well describe the whole classification process. The first CP is called 'Cor-RF-MdAcc', which means that we apply firstly the correlation coefficient 'Cor', then we apply RF on the obtained set and rank features according to MdAcc. We follow the same logic to name the other CP: (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 (10) 'ANOVA-pValue'. To preserve only important features, we retain the 200 first ranked ones from each of the 10 (sub)sets, except the set 'ANOVA-pValue' from which we select only 107 features that have a reasonable p-value (lower than 0.1). Ten reduced sets of ranked features are consequently obtained, named D_i, where $i \in \{1, \ldots, 10\}$. Then, to analyze the relative importance of individual features and to enable a comprehensive interpretation of the results, these reduced sets of ranked features are combined for comparison.

3.3 Visualization with FCA

This section focuses on comparing all the reduced sets (D_i, where $i \in \{1, \ldots, 10\}$) of highly ranked features (Figure 1). The combination of these subsets resulting from different CP, enables covering several possible results and yields to a stable unique reduced output. For the comparison propose, a binary table of *features* $\times$ *CP* is built (e.g., Table 1), where the objects (rows) are the features and the variables (columns) are the 10 CP. We put '1' if the feature exists in the reduced set of a corresponding CP; otherwise, we put '0'. Each feature has then a support[12] calculated from the obtained binary table, where the most frequent features are those existing in all the reduced sets (support =10). Nevertheless, since we are looking for frequent features according to the different CP, a subset of features common to at least 6 techniques is selected (i.e., features belonging to D_i, where $i \in \{1, \ldots, 10\}$ and identified by at least 6 CP), and a new subset of 48 frequent features is obtained. The choice of this value (6) is not random,

[10] Cohens Kappa (Kappa) is a statistical measure which compares an Observed Accuracy with an Expected Accuracy (random chance)

[11] A p-value helps determining the statistical significance of the results when a hypothesis test is performed.

[12] The support is the number of times we have '1' in each row, according to the binary table.

but it enables obtaining results from complementary FS methods. It ensures the selection of some relevant features that could have been removed by filters, while keeping a reasonable dataset size (48 features). A new binary table of the form 48 $features \times 10\,CP$ is obtained and presented in Table 1. It describes features in rows by the CP in columns and transforms then the initial problem statement from a data table of 111 $individuals \times 1195\ features$ to 48 $features \times 10\,CP$. The labels of the features start with the word 'm/z' which corresponds to the mass per charge value.

From this (48×10) binary table, we apply FCA with the help of ConExp tool [18]). Two seventy six concepts are obtained from the derived concept lattice (Figure 2). The combination of FCA with the results of the numerical methods and the transformation of the problem statement bring new light to the generated data. Four features 'm/z 383', 'm/z 227', 'm/z 114' and 'm/z 165' of the subconcept are identified as the most frequent (maximum rectangle full of 1 in Table 1). Most of the 44 remaining features highlight strong relationships between each others, such as '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'. Among the 48 frequent features, 39 are significant w.r.t. ANOVA (have a pvalue$<$0.05). The generated lattice highlights then the potential of the proposed feature selection approach for analyzing metabolomic data. It enables discriminating direct and indirect associations: highly linked metabolites belonging to the same concept. The links between the concepts in the lattice represent the degree of interdependencies between concept and metabolites belonging to the same concept. These 48 frequent features are then proposed as candidate for prediction.

4 Evaluation and discussion

4.1 Predictive performance evaluation and interpretation

Considering the 48 most frequent features previously identified, we would like to evaluate their predictive capacities. Accordingly, we start the performance evaluation using the ROC curves (Figure 3) of the 48 features with associated confidence intervals. These analyses are performed 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.

ROC curve is a non-parametric analysis, which is considered to be one of the most objective and statistically valid method for biomarker performance evaluation [11]. They 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 biomarkers requires to be quite limited (because of clinical constraints), we rely on the ROC curves of the top 2, 3, 5, 10, 20 and 48 of important features ranked based on their AUC values. These small sets of features are used to build the RF classification models based on the

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

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

cross validation (CV) performance. The ROC curves enable identifying this best combination of predictive features. Figure 3 shows that the best performance is

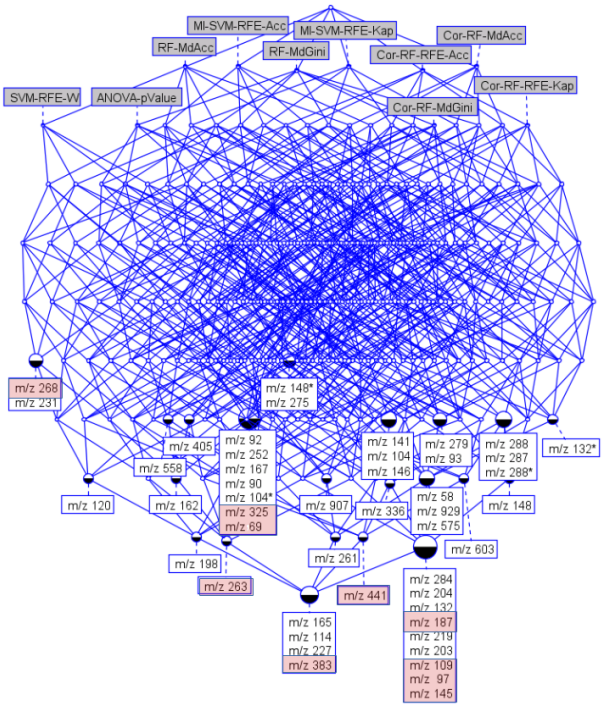

Fig. 2. The concept lattice derived from the 48 × 10 binary table (Table 1).

given to the 48 features together (AUC=0.867). But a predictive model with 48 metabolites is not useable in clinical practices. The set of best features with the smallest p-values and the highest accuracy values is selected to finally obtain a short list of potential biomarkers. When we select the ten first features (Table 3), we have an AUC equals to 0.79, and a CI=0.71-0.9. When we select the first four features, we obtain an AUC close to 0.75. These high AUC values show a good predictive performance.

In sight of these results, it is more advisable to select the 10 first features which have an AUC greater than 0.74 and a significant small t-test values (Table 3) as potential biomarkers. We compare this subset of 10 best predictive features with the four most frequent features (features with full of '1' in Table 1), we find that only one feature is in common, 'm/z 383'. We conclude that the core set of most frequent features is not the best predictive set, as expected biologically because the metabolomic analyses are performed 5 years before disease occurrence. Moreover, these best predictive features (or potential biomarkers) are not belonging to the same concept. Figure 2 highlights this conclusion and shows that the best predictive biomarkers have different extents and belong to concepts with different intents. They are depicted by the red squares in the lat-

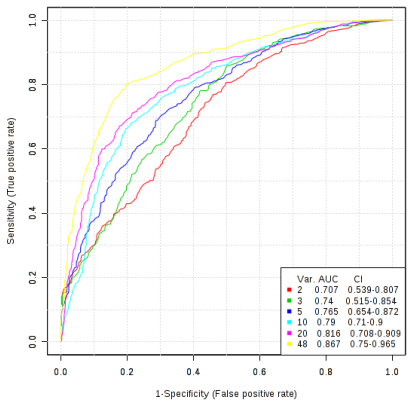

Fig. 3. The ROC curves of at least 2 and max 48 combined frequent features based on RF model and AUC ranking.

tice. For example, the features 'm/z 145', 'm/z 97', 'm/z 109' and 'm/z 187' are part of the intent of a concept including all the CP, except 'SVM-RFE-W', in extent. By contrast, the feature 'm/z 268' belongs to another concept including 6 CP in extent (*'RF-MdGini', 'RF-MdAcc', 'MI-SVM-RFE-Acc', 'MI-SVM-RFE-Kap', 'SVM-RFE-W', 'ANOVA-pValue'*). Here again, the simple visualization of the lattice comes to highlight the position of the predictive features among the discriminant ones and shows the associations with selection methods. This information is interesting for the expert domain since this visualization allows choosing the best combination of feature selection methods.

4.2 Selection of the best FS method(s)

As some feature selection methods do not retain the ten best predictive ones as their highly ranked, it remains essential to identify the methods that provide the best selection from metabolomic data. Here again, FCA comes to highlight and to assist information retrieval and visualization of the results. We then retain only the subset of ten best features (*'m/z 145', 'm/z 441', 'm/z 383', 'm/z 97', 'm/z 325', 'm/z 69', 'm/z 268', 'm/z 263', 'm/z 187' and 'm/z 109'*) identified previously due to the ROC curve, and apply FCA another time on their corresponding binary Table 2. A new concept lattice is generated (Figure 4) showing a superconcept with 4 feature selection methods, 'ANOVA-pValue', 'MI-SVM-RFE-Acc', 'RF-MdAcc' and 'RF-MdGini', verified by all features.

This is a very interesting result which needs a deeper interpretation before validation. We then consider these 4 methods and look for their ranking w.r.t. the 10 best predictive features (Table 3). Table 4 shows that RF-based techniques and Anova provide a good ranking to the 10 features contrarily to 'MI-SVM-RFE-Acc'. For example, 'm/z 145' is ranked first according to 'RF-MdAcc', 'RF-

Table 2. Input binary table describing the 6 best predictive features with the 10 CP.

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 145	1	1	1	1	1	1	1	1		1
m/z 97	1	1	1	1	1	1	1	1		1
m/z 263	1	1	1		1	1	1	1	1	1
m/z 325	1	1			1	1	1	1		1
m/z 268					1	1	1	1	1	1

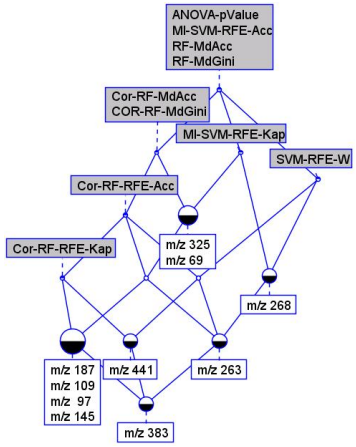

Fig. 4. The concept lattice of the 10 best predictive variables.

MdGini', second according to 'ANOVA-pvalue' and hundredth within 'MI-SVM-RFE-Acc'. The feature 'm/z 441' is ranked 6th according to 'RF-MdAcc', 8th within 'RF-MdGini', 172th within 'MI-SVM-RFE-Acc', and 11th according to 'ANOVA-pvalue'. Consequently, the toplist methods for biomarker identification from metabolomic data are RF-based and ANOVA.

5 Conclusion and future works

In this paper, we presented a new approach for the identification of predictive biomarkers from complex metabolomic dataset. Due to the nature of metabolomic data (highly correlated and noisy), the results highlighted the importance of working on reduced datasets to identify important variables related to the observed discrimination between case and control subjects and candidate for pre-

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

Feature	RF-MdAcc	RF-MdGini	MI-SVM-RFE-Acc	ANOVA-pValue
m/z 145	1	1	100	2
m/z 383	3	3	40	1
m/z 97	2	2	63	3
m/z 325	5	5	38	8
m/z 69	4	4	65	7
m/z 268	9	6	168	4
m/z 441	6	8	172	11
m/z 263	8	7	28	5
m/z 187	14	10	27	6
m/z 109	7	9	37	9

Table 4. Ranking of the 10 features with respect to 4 CP.

diction. Indeed, a combination of numerical (supervised) and symbolic (unsupervised) methods remains the best approach, as it allows combining the strengths of both techniques.

In this study, we used machine learning methods, RF and SVM, that we combined with FCA, to select a subset of good candidate biological features for prediction diseases. Our results showed the interest of this association to reveal subtle effects (hidden information) in such high dimensional datasets and how FCA highlighted the relationship between the best predictive features and the selection methods. RF-based methods as well as ANOVA gave the toplist of relevant features that best predict the disease development. With this help, the experts in biology will go deeper in interpretation, attesting the success of the knowledge discovery process. Additional experiments on other metabolomic datasets are required to attest the success of the proposed approach.

References

1. Bartel, H.G., Brüggemann, R.: Application of formal concept analysis to structure-activity relationships. Fresenius' Journal of Analytical Chemistry 361(1), 23–28 (1998)
2. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. 13(1), 1063–1095 (2012)
3. Breiman, L.: Random forests. In: Machine Learning. pp. 5–32 (2001)
4. Cho, H., Kim, S.B., Jeong, M.K., Park, Y., Miller, N.G., Ziegler, T.R., Jones, D.P.: 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)

5. Ganter, B., Wille, R.: Formal Concept Analysis – Mathematical Foundations. Springer (1999)
6. Gebert, J., Motameny, S., Faigle, U., Forst, C., Schrader, R.: Identifying Genes of Gene Regulatory Networks Using Formal Concept Analysis. Journal of Computational Biology 2, 185–194 (2008)
7. Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L., Goodacre, R.: A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. Analytica Chimica Acta 829, 1–8 (2014)
8. Gromski, P., Muhamadali, H., Ellis, D., Xu, Y., Correa, E., Turner, M., Goodacre, R.: 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)
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422 (2002)
10. Jansen, J., Hoefsloot, H., van der Greef, J., Timmerman, M., Westerhuis, J., Smilde, A.: Asca: analysis of multivariate data obtained from an experimental design. Journal of Chemometrics 19(9), 469–481 (2005)
11. Jianguo Xia, David I. Broadhurst, M.W., author, D.S.W.: Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics 9(2), 280–99 (2013)
12. Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A.: Feature Extraction: Foundations and Applications, chap. Embedded Methods, pp. 137–165. Springer Berlin Heidelberg (2006)
13. Mamas, M., Dunn, W., Neyses, L., Goodacre, R.: The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Arch Toxicol. 85(1), 5–17 (2011)
14. Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Formal concept analysis in knowledge processing: A survey on applications. Expert Systems with Applications 40(16), 6538 – 6560 (2013)
15. Saeys, Y., Inza, I., Larraaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
16. Vapnik, V.: Statistical Learning Theory. Wiley-Interscience, John Willey & Sons (1998)
17. Wang, H., Khoshgoftaar, T.M., Wald, R.: Measuring Stability of Feature Selection Techniques on Real-World Software Datasets. Information Reuse and Integration in Academia and Industry, pp. 113–132. Springer Vienna (2013)
18. Yevtushenko, S.A.: System of data analysis "concept explorer". In: Proceedings of the 7th national conference on Artificial Intelligence. pp. 127–134. KII'2000 (2000)