Prognostic Prediction through Biclustering-Based Classification of Clinical Gene Expression Time Series

André V. Carreiro [1]*, Orlando Anunciação [1], João A. Carriço [2], Sara C. Madeira [1]*

[1]Instituto Superior Técnico, Technical University of Lisbon, Portugal
and Knowledge Discovery and Bioinformatics (KDBIO) group, INESC-ID, Lisbon, Portugal

[2]Molecular Microbiology and Infection Unit, IMM and Faculty of Medicine, University of Lisbon, Lisbon, Portugal

Summary

The constant drive towards a more personalized medicine led to an increasing interest in temporal gene expression analyzes. It is now broadly accepted that considering a temporal perpective represents a great advantage to better understand disease progression and treatment results at a molecular level. In this context, biclustering algorithms emerged as an important tool to discover local expression patterns in biomedical applications, and CCC-Biclustering arose as an efficient algorithm relying on the temporal nature of data to identify all maximal temporal patterns in gene expression time series. In this work, CCC-Biclustering was integrated in new biclustering-based classifiers for prognostic prediction. As case study we analyzed multiple gene expression time series in order to classify the response of Multiple Sclerosis patients to the standard treatment with Interferon-β, to which nearly half of the patients reveal a negative response. In this scenario, using an effective predictive model of a patient's response would avoid useless and possibly harmful therapies for the non-responder group. The results revealed interesting potentialities to be further explored in classification problems involving other (clinical) time series.

1 Background

In the last decade, several techniques to evaluate gene expression became available, such as microarrays, which measure, instantly, the expression level of up to thousands of genes (possibly, all genes in a genome), and more recently RNA seq technologies for transcriptomic profiling. Gene expression experiments were, until more recently, limited to a static analysis, in which only a snapshot of the gene expression for a set of samples was available. However, the last years have witnessed an increase in time-course gene expression experiments and analysis. In fact, being able to study the temporal dynamics of gene expression is now enabling the study of complex biomedical problems, such as disease progression and drug response, from a different perspective. Studying these data is however challenging, both from a computational and biomedical point of view [1, 2].

Biclustering algorithms have been recognized as an important tool for the discovery of local expression patterns [3]. Given the known importance of discovering local temporal patterns of expression in both biological and clinical applications [1, 2, 4, 5], recent biclustering approaches have addressed this problem in the specific case of time series expression data [6]. In

* To whom correspondence should be addressed. Email: {smadeira,acarreiro}@kdbio.inesc-id.pt

fact, recent research on biclustering algorithms showed that although most of the biclustering problems are NP-hard, when working with expression time series the interesting biclusters are those with coherent evolutions on the columns [6]. This restriction leads to a tractable problem.

In this work, we use CCC-Biclustering [6], which finds all maximal contiguous column coherent (CCC) biclusters (subsets of genes with coherent expression patterns in contiguous subsets of time-points) by analyzing a discretized version of the expression matrix using efficient string processing techniques based on suffix trees. We propose a set of new biclustering-based classifiers, where CCC-Biclusters are used as the class discriminative features. We use as case study the prognostic prediction in patients with Multiple Sclerosis (MS) in response to the Interferon (IFN)-β standard treatment, using data from Baranzini et al. [7].

The main advantage of the proposed biclustering-based classifier lies on the interpretability of the results. In this case, the class-discriminant features are the biclusters, identifying subsets of genes coherently expressed over a subset of contiguous time-points, thus pinpointing potentially relevant biological processes related to disease progression and/or drug response. This information can be directly analyzed, while this is not possible or straightforward for other classifiers using gene expression time series [4, 5, 7]. The most limiting aspect of the proposed biclustering-based supervised learning approach might be the required discretization step, which conceals a risk of losing important information. However, a possibility to overcome this drawback might reside in an increase in the number of symbols used in the discretization step yielding more and thus more specific biclusters, while maintaining the efficiency of biclustering and the interpretability of the classification results. We note that the discretization step is key to guarantee the completeness and the efficiency of the biclustering algorithm, which would otherwise have to rely on heuristics.

1.1 Multiple Sclerosis

MS can be defined as a chronic inflammatory disease, characterized by a demyelinating disorder of the central nervous system (CNS) [8]. Although its etiology remains, to date, still far from total understanding, the interrelation of both genetic and environmental factors is believed to be crucial to the development of MS. A major factor to be considered in this regard is the phenotypic heterogeneity in MS, where different pathologic patterns may indicate differences in the pathogenic mechanisms [8]. Moreover, the search for single candidate genes that could account for the disease development is still unfruitful. The main conclusion is that MS is genetically complex, and thus it is not possible to select single genes to explain a person's susceptibility, since this might be a result of the interaction of several altered genes [8]. Consequent to the heterogeneity of the disease, the treatment response, even for one stage of MS only (relapsing-remitting (RR) MS), presents a high variability, suggesting different responses at the molecular level, leading to diverse clinical outcomes as the inhibition of the CNS inflammation [9]. Nevertheless, the treatment of RR-MS patients has routinely been carried with the use of recombinant human IFN-β [7]. However, up to half the patients show no benefits from this treatment, and negative side effects, such as flu-like symptoms and tissue damage, have to be considered [5]. Thus, the main goal of a time-course profiling of the treatment response of MS patients rests, as can be anticipated, in the possibility of accurately predicting a given patient's response, avoiding useless and possibly harmful treatments.

1.2 Related Work

Biclustering has been recently used in several data mining tasks, such as collaborative filtering [10] and bioinformatics, as is the case of miRNA based tumor classification [11]. Nevertheless, to our knowledge, biclustering was not used before in classification problems involving clinical expression time series. In what concerns the dataset used in this work, it has already been studied by other authors over the last years, whose work we briefly describe below.

Baranzini et al. [7] collected a dataset containing the profiling of MS patients subjected to IFN-β therapy. These authors proposed a quadratic analysis-based integrated Bayesian inference system (IBIS) to analyze it. They chose the best discriminative triplets of genes, obtaining a prediction accuracy up to 86% for a gene triplet consisting of Caspase 2, Caspase 10 and FLIP. We note, however, that in this work only the first time-point was considered. Lin et al. [4] proposed a new classification method, based on Hidden Markov Models (HMMs) with discriminative learning (using both positive and negative examples). In this work, the analysis was preceded by a feature selection step, to eliminate the least discriminative genes. The main results of applying this method to the MS dataset for two to seven time points were a prediction accuracy of up to 88%, and most importantly, the consideration and identification of patient-specific response rates. Finally, Costa et al. [5] introduced the concept of constrained mixture estimation of HMMs and applied it to the MS dataset. The constraints were positive when two patients were forced to be associated in the same group, or negative when they were not allowed to be grouped together. A preprocessing feature selection step was also performed. The main results include a prediction accuracy over 90% and the possibility of subgroup classification (two subgroups of good responders). This method also suggested the existence of one mislabeled patient, which was confirmed by Baranzini et al. [7].

2 Methods

In this section, we present the new biclustering-based classification strategies developed in this work: k-Nearest Neighbors (kNN) with different similarity measures and a meta-profiles classifier. We start by explaining CCC-Biclustering, an algorithm specifically proposed to analyze expression time series, used as part of the proposed classifiers. Figure 1 shows their workflow.

2.1 CCC-Biclustering

Let A' be an N_G rows by N_T columns gene expression matrix defined by its set of rows (genes), G, and its set of columns (time-points), T. In this context, A'_{ij} represents the expression level of gene i in time-point j. We address the case where gene expression levels in matrix A' can be discretized to a set of symbols, Σ, that represent distinctive activation levels. After the discretization process, matrix A' is transformed into matrix A, where $A_{ij} \in \Sigma$ represents the discretized value of the expression level of gene i in time-point j. In Figure 2 a three symbol alphabet $\Sigma = \{D, N, U\}$ was used, where 'D' (down), 'N' (no-change), and 'U' (up) mean that the expression levels decreased, didn't change or increased between consecutive time-points T_j and T_{j+1}. When T_j, T_{j+1} or both are missing, the symbol '$-$' is used.

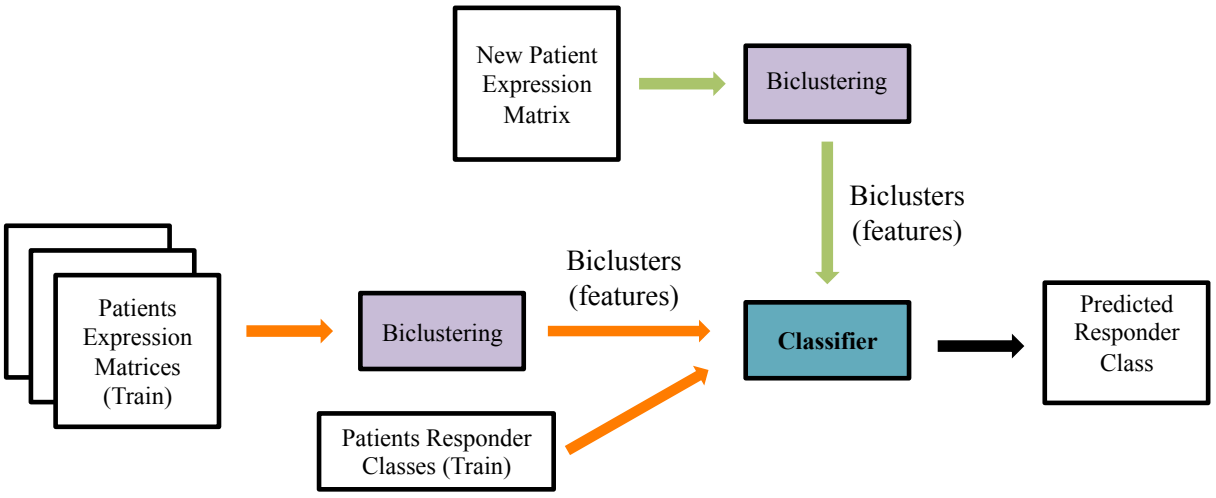

Figure 1: Basic workflow of a biclustering-based classifier.

The goal of biclustering algorithms is to identify a set of biclusters $B_k = (I_k, J_k)$ such that each bicluster satisfies specific characteristics of homogeneity [3]. For time series gene expression data analysis, Madeira et al. [6] defined the concept of CCC-Bicluster as follows: A *contiguous column coherent bicluster* (CCC-Bicluster) A_{IJ} is a subset of rows $I = \{i_1, \ldots, i_k\}$ and a subset of **contiguous** columns $J = \{r, r+1, \ldots, s-1, s\}$ such that $A_{ij} = A_{lj}, \forall i, l \in I$ and $\forall j \in J$. A CCC-Bicluster defines a string, common to every row in I for the columns in J. A CCC-Bicluster A_{IJ} is **maximal** if no other CCC-Bicluster exists that properly contains A_{IJ}, that is, if for all other CCC-Biclusters A_{LM}, $I \subseteq L \wedge J \subseteq M \Rightarrow I = L \wedge J = M$.

Consider now the matrix obtained by preprocessing matrix A using a simple alphabet transformation, that appends the column number to each symbol in the matrix and the generalized suffix tree built for the set of strings corresponding to each row in A. CCC-Biclustering is a linear time biclustering algorithm that finds and reports all maximal CCC-Biclusters based on their relationship with the nodes in the generalized suffix tree (see [6] for details and Figure 2).

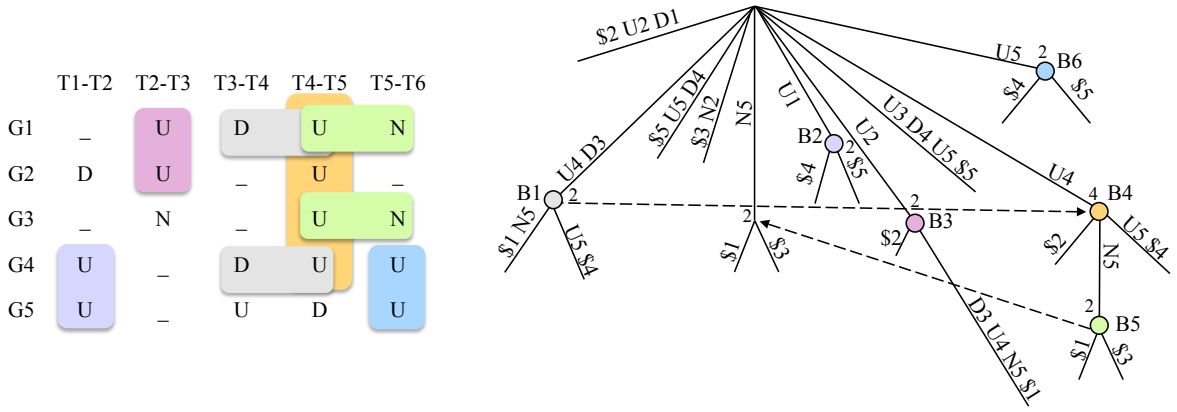

Figure 2: Maximal CCC-Biclusters in the discretized matrix and related nodes in the suffix tree.

2.2 Biclustering-based k-Nearest Neighbors

The kNN algorithm is a simple supervised learning method, whose goal is to classify an object based on the k closest training instances. The k parameter is a positive integer, usually small, chosen empirically with the help of some cross validation schemes, for example. In order to favor the best scoring instances, a distance-weighted algorithm can be used. The weights can be a function of their rank (with a weight of $1/d$, d being the rank or the distance to the test object). Algorithm 1 shows the biclustering-based kNN algorithm used to classify patient prognosis.

Algorithm 1: Biclustering-based kNN

Input : Score matrix between patients: S
Output: $predictedClass$
foreach *test patient* **do**
  build list with k highest scoring train patients $\rightarrow kPatients$
  from $kPatients$, compute the sum of the scores for each of the N classes: $scores_n$, $n \in 0, ..., N$
  $predictedClass$ = $arg_{max}(scores_n)$, $n \in 0, ..., N$

2.2.1 Score Matrix based on Biclusters Similarities

The matrix that represents the relationship between test and train patients, from which the k most similar train patients are selected to classify each test patient, is the score matrix (the higher the score, the higher the degree of similarity between the patients). It has as many rows as the number of train patients, and a number of columns equal to the number of test patients. To reduce the effects of unbalanced data, as is the case in the case study used in this work, a penalty/weight can be included in the computation of this matrix, a fraction which alters the actual score between the two patients: $InitialScore \times penalty = WeightedScore$.

The entry (i, j) of the score matrix between patients represents the degree of similarity between a bicluster B_i from the set of biclusters of a test patient, and a bicluster B_j from the set of biclusters of a train patient. This similarity can be computed from the fraction of common elements between the two biclusters, using an adapted version of the Jaccard Index, used by Madeira et al. [6], where we include the information of the genes expression variation, since we are comparing biclusters from different patients. However, it is necessary to transform this measure of similarity between two patients into a single score value, to proceed with kNN classification. This transformation is performed as follows:

$$S(P_{test}, P_{train}) = \frac{\sum_{i=1}^{\#B_{test}} max(Sim(B_i, B_j), j \in 1, ..., \#B_{train})}{\#B_{test}} \quad (1)$$

where $Sim(B_i, B_j)$ is the similarity between biclusters B_i and B_j, and $\#B_{test}$ and $\#B_{train}$ represent the number of biclusters of, respectively, the test and train patient, P_{test} and P_{train}.

2.2.2 Score Matrix based on Profiles Similarities

Another strategy to compute the score matrix between the test patients and the training set relies on the fact that each CCC-Bicluster is represented by a pattern of symbols, a profile,

representative of the coherent evolution in the expression of the genes in the bicluster along the bicluster time-points. A profile is said to be shared between patients if it identifies a similar expression pattern and represents biclusters which have the required minimum number of genes and/or time-points in common, a parameter determined empirically.

The score matrix between patients is computed, such that an entry (i, j) represents the number of profiles shared between train patient i and test patient j. Instead of the sum of shared profiles between patients, the entry (i, j) of the score matrix can also be computed with a polynomial kernel (a quadratic kernel in general). The idea is to penalize the patients with a larger number of biclusters, since a higher number of profile matches could be due to random events.

Filtering Non-Discriminative Biclusters based on Profiles

A given profile is kept in the filtered set, if and only if it contributes more to the discrimination than to the confusion between classes, that is, a profile in a train patient's set of profiles is maintained if and only if it is shared by more patients of the same class than of the other class. A minimum number of shared genes and/or time-points can also be used and fine-tuned.

2.2.3 Score Matrix based on Symbol Pairing with Time-Lags

As one might expect, even when the same genes are involved in a given mechanism in different patients, the expression evolution pattern for one patient might be delayed when compared to others'. As such, the possibility of time-lags in gene expression should be taken into account, as it is a consequence of the patient-specific response rate, and shown to be of particular importance in previous time series expression studies [4]. In this approach, all biclusters (or filtered ones) of the test patient are analyzed and a parameter for a maximum time-lag (number of time-points to consider in the delay) is defined. Then, for each test bicluster, a comparison is made between the discretized symbols, computing the number of perfect matches, considering translations in the time-points, from 0 (the original position) to the maximum time-lag, and its symmetric, thus allowing translations in both directions along the time axis (Figure 3). The time-lag returning the highest score is chosen, and the binary submatrix resulting from that specific comparison is written in a final matrix. The sum of this matrix represents the score between the two patients, the entry (i, j) of the score matrix for the whole set of patients.

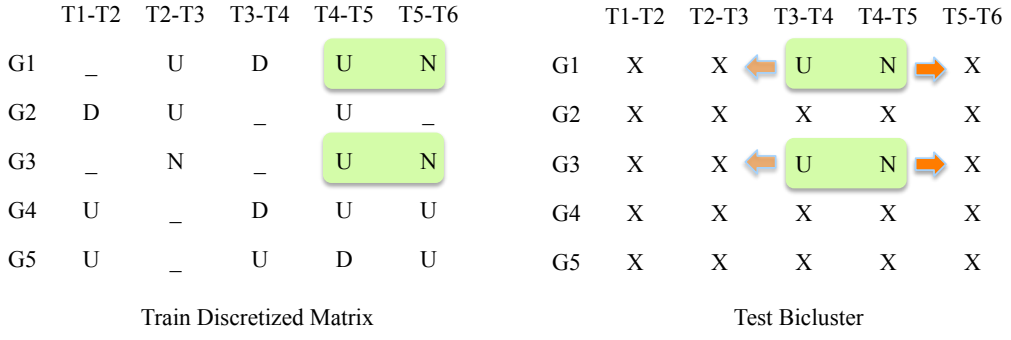

Figure 3: Symbolic comparison between a test bicluster (on the right) and the train discretized expression matrix (on the left). In this case, the highest scoring time-lag is +1 corresponding to shifting the test bicluster one position to the right, where a perfect symbol match is achieved with the symbols shaded green on the left, identifying a delayed expression profile between the patients.

2.3 Meta-Profiles Classification

Having explored different strategies to combine biclustering and kNN classification, we now present a new classification approach following the biclusters computation. It is based on the mentioned fact that each bicluster identifies a pattern of temporal evolution in terms of gene expression, which is represented by a profile. A meta-profile represents a set of similar profiles. In this approach, presented in Algorithm 2, the goal is to analyze if a given profile is shared between more patients of one of the classes.

For example, if a train profile is shared only between good responders, then if a test patient shows an equivalent expression profile, the probability of this patient being a good responder increases. In this method, the class proportions for each profile of a test patient contribute for the patient classification, in a weighted-voting scheme. Due to the difference in the class distributions, a penalty can also be introduced here to soften the classification.

Algorithm 2: Meta-Profiles Classification

Input : Meta-Profiles space: list with the profiles of all computed biclusters
Output: $predictedClass$
foreach *meta-profile* m **do**
  foreach *train patient* tp **do**
    $TrainIndexes \leftarrow \{\}$
    if *meta-profile* $m \in$ *set of profiles of* tp **then**
      add tp to $TrainIndexes$
  compute class proportions of meta-profile m for each of the N classes: $Proportions_n$, $n \in 0, ..., N$
foreach *test patient* i **do**
  foreach *test profile* p **do**
    if $p \in$ *meta-profiles space* **then**
      associate respective class proportions to p
  compute the sum of class proportions for all test profiles: $sumProportions_n$, $n \in 0, ..., N$
  $predictedClass$ = $arg_{max}(sumProportions_n)$, $n \in 0, ..., N$

In the case study, the performed tests revealed that the best discriminative criterion was that the patients with more balanced class proportions were classified as good responders (class 1). This corresponds to the following classification criterion, when two classes are considered:

if $sumProportions_1 \times penalty < sumProportions_0$ **then** $predictedClass = 1$;
else $predictedClass = 0$;

3 Results and Discussion

In this section, we present and discuss the specificities of the case study used in this work, including the dataset description and preprocessing, as well as the evaluation strategies used. We discuss the main results obtained with the proposed classification approaches, as an indicator of the classifiers' performance, in different contexts. This includes an evaluation on the effect of feature selection and an analysis of the classifiers' performance when different time-points are used. We also compare the biclustering-based classifiers with state-of-the-art classifiers, using both the real-valued and a discretized version of the dataset. Finally, and in the case of

the meta-profiles classifier, we discuss the biomedical interpretation of the results by analyzing the genes which occur more often in the most class-discriminant biclusters.

3.1 Dataset

The dataset used as case study in this work was collected by Baranzini et al. [7]. Fifty two patients with RR-MS were followed for a minimum of two years after the treatment initiation. After that time, patients were classified according to their response to the treatment, as good or bad responders. Thirty two patients were considered good responders, while the remaining twenty were classified as bad responders to IFN-β therapy. Seventy genes were pre-selected based on biological criteria, and their expression profile was measured in seven time-points (initial point, and three, six, nine, twelve, eighteen and twenty-four months after treatment initiation), using one-step kinetic reverse transcription PCR [7].

3.1.1 Preprocessing

In order to apply CCC-Biclustering [6] as part of the proposed biclustering-based classifiers, we normalized and discretized the expression data. The discretization was performed by computing variations between time-points as carried by Madeira et al. [6], thus resulting in patterns of temporal gene expression evolution with three symbols: decrease ('D'), no change ('N') and increase ('U'). In the case of standard classifiers, not able to deal with missing values directly, these were filled with the average of the closest neighboring values, after data normalization. In the biclustering-based classifiers proposed in this work, the CCC-Biclustering algorithm is able to handle missing values directly. However, for comparison purposes, the results shown were obtained with filled missing values, even for the biclustering-based classifiers.

3.2 Evaluation

Since we are dealing with a classification task to predict a given patient's response to a MS treatment, the evaluation of any method must be based on the algorithm capability to predict the response class of a new MS patient, based on similar data used to train the classifier. For a finite dataset, the prediction accuracy is defined as the percentage of correctly predicted instances. However, to extract more information about the performance of a classifier, we use the confusion matrix, which accounts for the number of correctly classified instances for each class, allowing the computation of two important ratios: *True Positive Rate* - the ratio of correctly classified instances of class 1 (good responder), and *False Positive Rate* - the ratio of incorrectly classified instances of class 0 (bad responder). Furthermore, using these two ratios as data points, we are able to construct the Receiver Operating Characteristics (ROC) curve. This provides information about a classifier's performance for all misclassification costs, and all possible class ratios. When only a few (FP, TP) pairs are available, we use an interpolated curve corresponding to an approximate ROC curve.

3.2.1 Cross-validation

A commonly used strategy to evaluate accuracy estimation is the k-fold cross validation (CV) scheme, where dataset D is randomly partitioned into k mutually exclusive subsets, of (approximately) equal size : $D_1, ..., D_k$. Then, for each fold (1 to k), the classifier is trained with the $k-1$ remaining subsets, and tested with subset D_k. This is repeated k times, and the overall prediction accuracy estimate is the mean prediction accuracy for all k folds. When k equals the number of instances in the dataset, the term Leave-one-out (LOO) CV is used, since one instance is left out to test the classifier, while all the others constitute the train set. When the class proportions in the original dataset are maintained in the subsets or folds, the term stratified CV is used. In this work, we use both LOO and 5 repetitions of stratified 4-fold CV.

3.3 Performance

Table 1 summarizes the prediction accuracies obtained for the proposed biclustering-based classifiers. Figure 4 shows the confusion matrix and approximate ROC curve for the meta-profiles classification method (sum criterion with penalty = 61.4%, using the best 90% of computed biclusters, in terms of p-value, as computed in [6]). This classifier achieved a prediction accuracy of 94.23% for LOO CV. We note, however, that with 5 x 4-fold CV, the prediction accuracy dropped below 68%. This fact might be justified by an optimistic value of LOO CV due to data overfitting, or to the small size of the dataset. In fact, in a 4-fold CV with 52 patients, the test set has 13 patients, significantly reducing the training set. Thus, important information to accurately compute the profile proportions between classes might be lost.

Table 1: Prediction accuracies for biclustering-based classifiers using LOO and 5 x 4-fold CV. Abbreviations: BS (Biclusters Similarities), PS (Profiles Similarities), K (Kernel), SP (Symbol Pairing), TL (Time-Lags), NF (Not Filtered), F (Filtered), SPEN (sum penalty).

	kNN BS	kNN PS		kNN PSK		kNN SPTL		Meta-Profiles
		NF	F	NF	F	=0	=1	SPEN=61.4%
LOO	59.62	50.00	63.46	46.15	63.46	57.69	69.23	**94.23**
5 x 4-fold	61.92	52.69	51.92	51.92	57.31	60.77	68.46	**67.31**
	±5.83	±2.92	±3.99	±2.36	±4.59	±7.01	±2.19	**±4.90**

At this point it is important to emphasize particular characteristics of this dataset. Besides class unbalance, we can find considerable differences between what is shared between the patients of the two response classes: good responders have a significant number of similar biclusters in common with other good responders, but also with the bad responders. These shared similar biclusters might include characteristic disease expression signatures, common to all RR-MS patients, a fact that shall be further investigated. Bad responders, however, show evidences of having few similar biclusters in common, beside the ones also shared with the good responders group. This fact suggests that there are different expression signatures associated to a poor response to IFN-β treatment or an absence of signature present in good responders, a probable result of differences in the fragile balance of several pathways associated to the disease and/or treatment response. This might justify the used criterion in the meta-profiles method, which led to a prediction accuracy up to 94.23%.

		Classified as	
		0	1
Real Class	0	17 (85.00%)	3 (15.00%)
	1	0 (0.00%)	32 (100.00%)

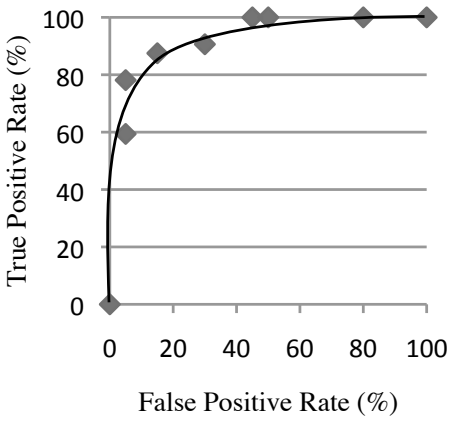

Figure 4: (Left): Confusion matrix for meta-profiles classification (with penalty = 61.4%). (Right): Approximate ROC curve constructed with variations of the penalty in the sum criterion for meta-profiles classification. Class 0 and 1 correspond to bad and good responders, respectively.

Even in this scenario, some of the proposed classifiers revealed potentialities which are worth exploring. We highlight the computation of the score matrix using symbol pairing with time-lags, since simply considering time delays improved significantly the prediction accuracy (p-value $=0.0025$, paired t-test). The best predictor was the biclustering-based meta-profiles classifier, where the used criterion was opposite to the generally expected. This might be a result of data specificities, and should be further analyzed, especially regarding its lower specificity: a permutation test, shuffling the class labels 1000 times, returned a drop in the prediction accuracy of approximately 15%.

It was not possible to reproduce previous results on the MS dataset [5], due to serious difficulties in getting access and running the classifiers, or even to obtain the test/train sets. This led to the choice of comparing our results only with standard classifiers. This was performed using Weka (`www.cs.waikato.ac.nz/ml/weka`). Instead of a matrix per patient, the dataset used in this situation consists of a single matrix where each row represents the data for a single patient, with 7 blocks (number of time-points) of 70 expression values each (number of genes). This means that the expression values of all genes for a given time-point are grouped together in the matrix, and each of its entries (P,J), with $J = GT$ represents the expression value of a gene G in a time-point T for a patient P. Nonetheless, with the exception of meta-profiles, the prediction accuracies obtained by the other proposed biclustering-based classifiers are lower than the results reported (but not reproducible) in the literature for this dataset [4, 5, 7]. We note, however, that all these approaches used feature selection by first selecting a small set of genes. We also performed experiments with real-valued gene expression matrices using feature selection as a preprocessing step achieving a 5 x 4-fold cross validation accuracy rate of 95.77% with a SVM classifier using a quadratic kernel. We consider this approach, using feature selection as a preprocessing step, to be biased since it uses information of data that is going to be part of test sets in the cross validation strategy. Therefore the results in Table 2 were obtained using a strategy in which we applied feature selection in each step of cross validation and only using training data.

Table 2 shows the best prediction accuracies obtained for the different state of the art classifiers in the real-valued expression data, together with the best prediction accuracies in a discretized version of the expression data. We aimed to assess the influence of the discretization process used in the biclustering-based classifiers. We show that standard classifiers tested on

Table 2: Prediction accuracies obtained from standard classifiers using the real-valued dataset (Top) and a discretized version (Bottom), when LOO (Leave-One-Out) and 5 x 4-fold cross validation are used. We use the following abbreviations: DT (Decision Tree), kNN (k-Nearest Neighbors), SVM (Support Vector Machines), LR (Logistic Regression), MLP (Multilayer Perceptron) and RBFN (Radial Basis Function Network).

	DT	kNN	SVM	LR	RBFN	MLP	Data
LOO	71.15	86.54	92.31	80.77	88.46	86.54	real-valued
5 x 4-fold	70.77 ±7.86	82.31 3.70	85.00 ±2.51	80.38 ±4.59	83.85 ±3.22	86.15 ±3.70	
LOO	51.92	55.77	59.62	40.38	57.69	46.15	discretized
5 x 4-fold	54.61 ±3.75	49.62 ±10.30	53.08 ±5.86	45.77 ±10.21	56.15 ±11.08	57.95 ±10.18	

the real-valued dataset outperformed most of the biclustering-based classifiers, excluding the meta-profiles method. It is also possible to observe that the standard decision trees prediction accuracy is not significantly higher (p-value > 0.16, paired t-test) than the one of the biclustering-based kNN based on symbol pairing with a maximum time-lag of one time-point in each direction (mean prediction accuracy of 69.23% (LOO) and 68.08% (5 x 4-fold)). It is also shown that the use of a discretized version of data lowers the classifiers' performance significantly (p-value < 0.05 for all classifiers, paired t-test). These evidences suggest that this kind of classifiers cannot deal well with discretized data of this type (especially with the significant differences in what is shared between patients of different classes). In fact, the biclustering-based kNN classifier based on symbol pairing with time-lags outperformed significantly all these standard classifiers (p-value < 0.05, paired t-test) when acting upon discretized data. The superiority of the meta-profiles classifier was even more evident.

Although the precision accuracies obtained for the MS dataset are not as high as desired, we highlight that IFN-β therapy is, currently, the standard treatment for MS. Therefore, if a classifier is able to correctly predict the patients response in a percentage higher than the proportion of good responders in the population, then it presents a significant advantage. However, we must contrast the case of false positives (bad responders classified as good responders, thus receiving the treatment) and false negatives (good responders missing the treatment). Given the negative side effects and the arising of alternative therapies, the classification should favor the bad response classification. This means that we should minimize the false positive rate, thus avoiding useless and possibly harmful treatments, allowing for the patients to an earlier change to different forms of treatment, potentially more effective for their particular situation.

3.3.1 Feature Selection

The dataset characteristics points to an overfitting of the data, since there are many more genes than time-points. In this scenario, feature selection can be used to reduce the number of genes used in the classification in attempt to minimize overfitting. Thus, we resorted to the reduced set of genes found by Costa et al. [5], which consists of 17 genes out of all the 70 initial ones. We note, however, that the algorithm used to compute this reduced set was built to find the genes which would return the best prediction accuracies for those specific classifiers. This subset of genes can be found in Table 4. The prediction accuracies are summarized in Table 3.