A-iBic: reconstruction of gene expression data by aggregation of imputed biclusters with coherent evolution

Alan Ramponi
University of Trento
Via Sommarive 9, 38123 - Povo (TN), Italy
alan.ramponi@studenti.unitn.it

ABSTRACT

The presence of missing values in a dataset is a problem of great importance in the analysis and the subsequent interpretation of data. Handling them in an approximate or an improper way can lead to inaccurate results, and then to infer misleading conclusions. Thanks to the advent, in the last decade, of a wide range of missing data imputation techniques, it is now possible to ensure that the process of data analysis is as accurate as possible. Although for some dataset a simple approach is enough, gene expression datasets derived from microarray and RNA-seq experiments need a greater effort. In this paper is accurately described the problem and the proposed solution, along with the performance evaluation methodology used in order to choose the best approach to adopt for this very purpose. First, the dataset was cleaned coherently to the distribution of missing values of outdated platforms. Second, it was splitted into chunks using a biclustering approach that preserves the evolution behaviour of genes. Third, different imputation techniques are computed in each of them. Finally, biclusters were aggregated into a unique fully imputed dataset.

Categories and Subject Descriptors

[**Applied computing**]: Life and medical sciences—*bioinformatics*; [**Information systems**]: Data mining—*clustering, nearest-neighbor search, data cleaning*; [**Computing methodologies**]: Machine learning—*motif discovery, supervised learning by regression, validation*.

Keywords

Data mining, missing data, data cleaning, data imputation, biclustering, big data, biological data, gene expression data.

1. INTRODUCTION

In the field of statistics and data analysis is recurring the presence, within a dataset, of a large amount of missing values. In the case of data gathered from microarray and RNA-seq experiments, the presence of missing data is due to the hardware limitations of some old platforms which lead to a less thorough analysis. The robustness and the accuracy of the conclusion drawn after the analysis of the data, especially in the life sciences, is crucial. Therefore, is necessary, as much as possible, to reconstruct the missing data through appropriate algorithms that infer the missing value from within the very dataset. By forming an ensemble of imputed biclusters with coherent evolution, it was possible to obtain a full imputed dataset that minimizes as much as possible the error rate between expected and predicted condition contrasts for each gene.

This paper is structured as follows: (Section 2) briefly offers a motivating example that introduces the reader to the problem, (Section 3) analyzes it and presents the *Vitis Vinifera*[1] dataset that was used during all the process for testing purposes, (Section 4) discusses the solution by dividing it into two parts: the data cleaning step and the data imputation step. (Section 5) presents the related work, results are presented in (Section 6) with an accurate analysis of the performance evaluation methodology and finally (Section 7) summarizes the obtained results.

Project report, Department of Information Engineering and Computer Science, University of Trento, Italy. Course: Laboratory of Biological Data mining. Academic Year: 2015/16. Author: Alan Ramponi (ID: 179850, e-mail address: alan.ramponi@studenti.unitn.it).

2. MOTIVATING EXAMPLE

Some valid motivating examples are represented by the challenge of finding drugs in order to discover new treatments to diseases and to predict health problems in support of human expertise in clinical decisions. Usually, medical datasets have huge proportions of missing data caused by information that haven't been requested by assessors or that the patients neglect to share [6]. The way of handling missing features can determine the outcome of a vital prediction. Thus, it becomes necessary to address the problem in the right way.

3. PROBLEM STATEMENT

The dataset taken into account in this paper is an organism-specific matrix of expression values of *Vitis Vinifera* derived from publicly available microarray and RNA-seq experiments which are homogenized to make them comparable. The rows correspond to all the known genes of the organism and the columns correspond to condition contrasts[2]. There are a total of 58 experiments (41 were retrieved from GEO and 14 were retrieved from ArrayExpress) and 1744

[1] *V. vinifera* is a *Vitis* species, the common grape vine native to the Mediterranean area, central Europe, and s.w. Asia.

[2] We refer to it as condition contrasts because they don't represent experimental conditions, but the difference between a test and reference condition.

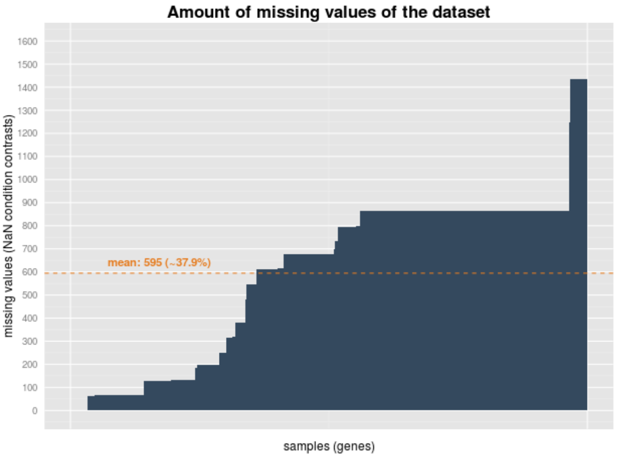

Figure 1: Amount of missing data of the whole dataset (x axis: samples, y axis: features).

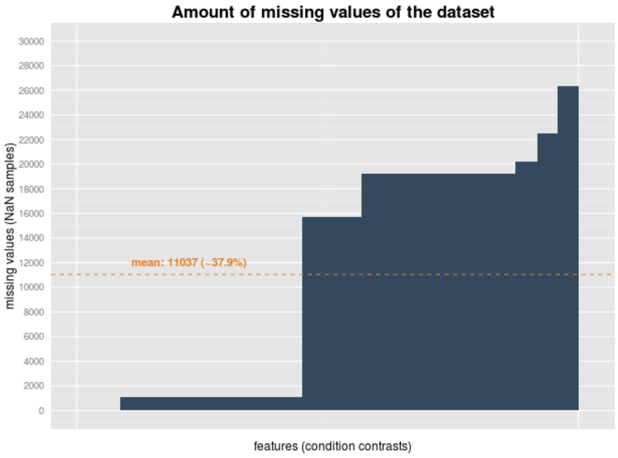

Figure 2: Amount of missing data of the whole dataset (x axis: features, y axis: samples).

samples measured on 19 different platforms. In details, the dataset is composed as follows:

- **samples**: 29090 (i.e. the *genes*);
- **features**: 1564 (i.e. the *condition contrasts*).

The main issue with this dataset is that it has a large amount of missing values. In particular, a missing values analysis gave as a result that both the samples and the features have an average of 37.9% of missing values, that is something unacceptable (see Figure 1 and Figure 2). Some genes and some condition contrasts have also almost only missing data. Therefore, it is crucial to decide how to treat these NA (*not available* data) properly without compromising the amount and the quality of information contained into the dataset.

4. SOLUTION

The solution to the problem of missing data consists of two main steps: the first one concerns the removal of those samples or features that are not informative at all, and the second one concerns the filling of the remaining missing data through the use of a combination of robust and effective state of the art algorithms.

4.1 Data cleaning

After the analysis of the distribution of missing data in the dataset (for both the samples and the features) it has been possible to identify genes and condition contrasts that are less informative. Discarding all the rows and the columns that have missing data may result to an excessive reduction in the size of the original dataset, so it can strongly affect the performances. On the other hand, the definition of an heuristic or a specific percentage that says what and how many rows and columns to remove from the dataset represents a very rough approach. It was therefore decided to define the amount of data to be removed according to the very distribution of missing values within the dataset. As it is possible to see in Figure 3 and Figure 4, data was removed according to the step-like distribution of missing values.

Using this approach it was possible to extract an outstanding information: all the last three "steps" in the ordered missing values distribution of the condition contrasts were extracted from the very same outdated platforms (*cDNA*, *2channel* and *unknown*[3]), so this justifies the distribution and makes this methodology of data cleansing very trustworthy.

The cleaning of the dataset has led to drastically reduce the presence of missing values for both the samples and the features, leading some genes, among the other things, to have no missing values. The resulted dataset after the dimensionality reduction is shown in Table 1 and the resulting amount of missing values is shown in the Appendix A (Figure 9 and Figure 10).

	Original dataset		**Cleaned dataset**	
	abs. #	Percentage	abs. #	Percentage
Samples	29090	100%	28044	96.4%
NA mean	595	37.9%	410	26.1%
Features	1567	100%	1370	87.4%
NA mean	11037	37.9%	8384	28.8%

Table 1: A comparison between the original dataset and the cleaned dataset as regards the number of samples and features and their missing values mean.

4.2 Data imputing

Once the data cleaning operation was finished, it was possible to focus the attention on the remaining missing values in order to compute the imputing on them. A first analysis that was made concerns the nature of missing values. In literature, missing data problems are classified into three categories called *Missing completely at random* (MCAR), *Missing at random* (MAR) and *Missing not at random* (MNAR) [10]. Briefly, the first one occurs when the probability of being missing is the same for all the cases, the second one occurs when the probability of being missing is the same only within groups defined by the observed data and the

[3] This name refers to an unknown platform, according to the given metadata information.

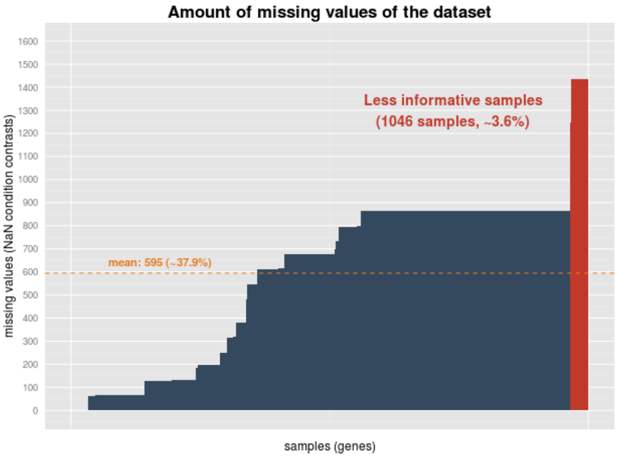

Figure 3: In red are represented the genes removed from the original dataset (x axis: samples, y axis: features).

last one occurs if neither MCAR nor MAR holds. The results obtained are that compendia that aggregate biological experiments from a large number of platforms fall into the second category. In fact, missing values are not randomly distributed across all the cases.

After that, a new approach to use imputation was designed, in order to exploit the current state of the art in a better way by applying it to this specific domain.

4.2.1 *Biclustering of genes with coherent evolution*

In order to both preserve as much as possible the behaviour of each single gene across its condition contrasts and to prevent computational issues, the original dataset was splitted into many biclusters. Biclustering is a technique which allows simultaneous clustering of the rows and columns of a matrix [11]. Given a set of m rows in n columns, the biclustering algorithm generates some biclusters (i.e. subsets of rows which exhibit similar behavior across a subset of columns, or vice versa) according to some similarity conditions. Two different biclustering algorithms were tested for this very purpose:

- **Plaid biclustering**: based on the well-known plaid model of Lazzeroni and Owen [7], this approach finds biclusters with constant rows or columns. This leads to a set of overlapping biclusters with coherent values;
- **xMotifs biclustering**: based on the Xmotifs algorithm of Murali and Kasif [8], it searches for conserved gene states firstly by discretizing the data and then by choosing a random column n times, and for each of them it performs the following steps:
    1. Choose a subset of columns m times and collect all rows with equal state in this subset, including the above column;
    2. Collect all columns where these rows have the same state;
    3. Return the bicluster if it has the most rows of all the bicluster found and if it is also larger than an α fraction of the data.

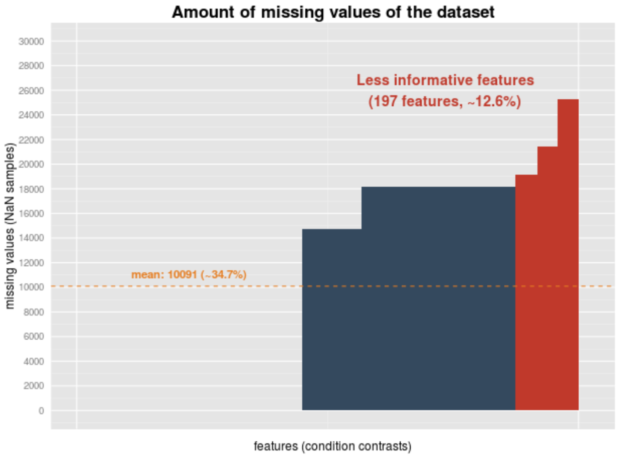

Figure 4: In red are represented the cond. contrasts removed from the original dataset (x axis: features, y axis: samples).

After an accurate analysis of the results, α was setted to 0.00001, n was setted to 100 and m was setted to 50.

The first algorithm was discarded after a single computation because it generated biclusters of genes with constant values across all the columns, forgetting the intrinsic behaviour of genes. Thus, the second one was used in all the process since it generated biclusters taking into account the evolution of each single observation (see Figure 5[4] for an example). In this way, it was possible to infer accurate results based on the most similar genes in the dataset.

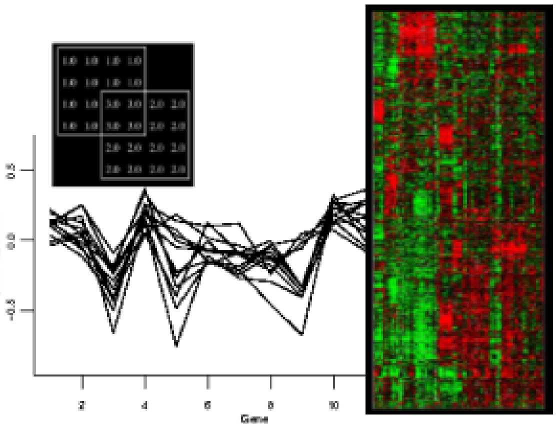

Figure 5: An example of biclustering based on coherent evolution of genes. In the left it is possible to see the behaviour of genes across the condition contrasts, and in the right it is possible to see the heatmap generated after the process of biclustering.

[4] An image taken from the web for illustrative purposes only.

4.2.2 Imputation of each bicluster

After a thorough analysis of the current imputing state of the art, it was possible to analyze several methods in order to choose the most suitable for this task. The approaches explored with their critical analysis are the following:

- **Raw mean imputation**: it was confirmed that the raw mean of each row (i.e. each gene) used to infer a missing condition contrast about that particular gene produces unreliable results. In particular, the mean performs poorly because it does not make use of the underlying correlation structure of the data [9]. Despite the raw mean as a single imputation technique is not very robust, it was decided to use it for reasons of a subsequent comparison in terms of performances with the other more accurate techniques;

- **Multivariate imputation by chained equations**: multiple imputation is one of the great ideas in statistical science that performs a single imputation multiple times until the convergence. The chained equation process can be broken down into these steps [1]:

  1. A simple imputation is performed for every missing value in the dataset. These imputations can be thought of as *place holders*;
  2. The *place holder* imputations for one variable *var* are set back to missing;
  3. The observed values from *var* in (2) are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. Thus, *var* is the dependent variable in a regression model and all the other variables are independent variables in it;
  4. The missing values for *var* are then replaced with predictions from the regression model. When *var* is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used;
  5. Steps 2-4 are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one cycle. At the end of one cycle all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data;
  6. Steps 2-4 are repeated for a (specified) number of cycles, with the imputations being updated at each cycle. At the end the final imputations are retained, resulting in one imputed dataset.

  Some univariate imputation methods that can be used are *predictive mean matching*, *Bayesian linear regression*, *non-Bayesian linear regression* and *2-level linear model*. However, after an accurate analysis of this method, it was discovered that it was not suitable for the dataset taken into account due to its assumption that missing data is MCAR.

- **Weighted k-nearest neighbors (WkNN)**: this well-known technique is simple and elegant. It is simple because it fills the holes in the data with plausible values, it is elegant because the uncertainty about the unknown data is coded in the data itself. The input consists of the k closest examples to the target gene and the output is the average of the values of the k nearest neighbors, weighted by their distance from the target;

- **Singular value decomposition (SVD)**: with this technique first missing values are filled using the mean of the column. Then a low, rank-k approximation of the input matrix is computed and the missing values are filled again from the rank-k approximation. The rank-k approximation is recomputed with the imputed values and missing values are filled again, for n times;

- **Approx. singular value thresholding (aSVT)**: as in the previous case, first missing values are filled using the mean of the column. Then, the SVD of the matrix is computed and a penality λ is subtracted from each of the singular values, with a threshold at 0. Thus, imputation works by multiplying back out the augmented SVD;

- **Singular value thresholding (SVT)**: this is a recent algorithm for matrix completion by Cai, Candes and Shen [2]. The algorithm minimizes the nuclear norm of a matrix, subject to certain types of constraints, and it becomes so powerful in the task of recovering a large matrix from a small subset of its entries. The algorithm is iterative: it produces a sequence of matrices $\{\boldsymbol{X}^k, \boldsymbol{Y}^k\}$, and at each step mainly performs a soft-thresholding operation on the singular values of the matrix $\boldsymbol{Y}^k$.

Thus, this analysis leads to discard the multivariate imputation by chained equations from the whole process due to its MCAR assumption of the missing values, and then to proceed with the evaluation of all the other techniques presented in this paper.

4.2.3 Aggregation of imputed biclusters

As a last step of the process of data imputing, all the biclusters previously generated and then imputed were ensembled in order to reconstruct the whole original dataset filled with values as accurate as possible. It is important to note that, given a matrix M with row indexes i and column indexes j, if there is more than one bicluster with an M_{ij} imputed an average of the predicted values was computed in order to refine the imputation. This scenario mainly happens in the case of using a biclustering technique that allows overlapping biclusters (e.g. Plaid biclustering).

The whole process of the solution explained in this section is represented in Figure 6.

5. RELATED WORK

Over the past decade, many researchers conducted studies on the biological data mining field and on the techniques to handle missing values. However, in literature there isn't any work that exploits the power of biclustering along with the problem of missing data by maintaining coherent evolution of genes, and then by aggregating the obtained biclusters in the way explained in this paper. Among the most interesting research works that can be found there are surely [3]

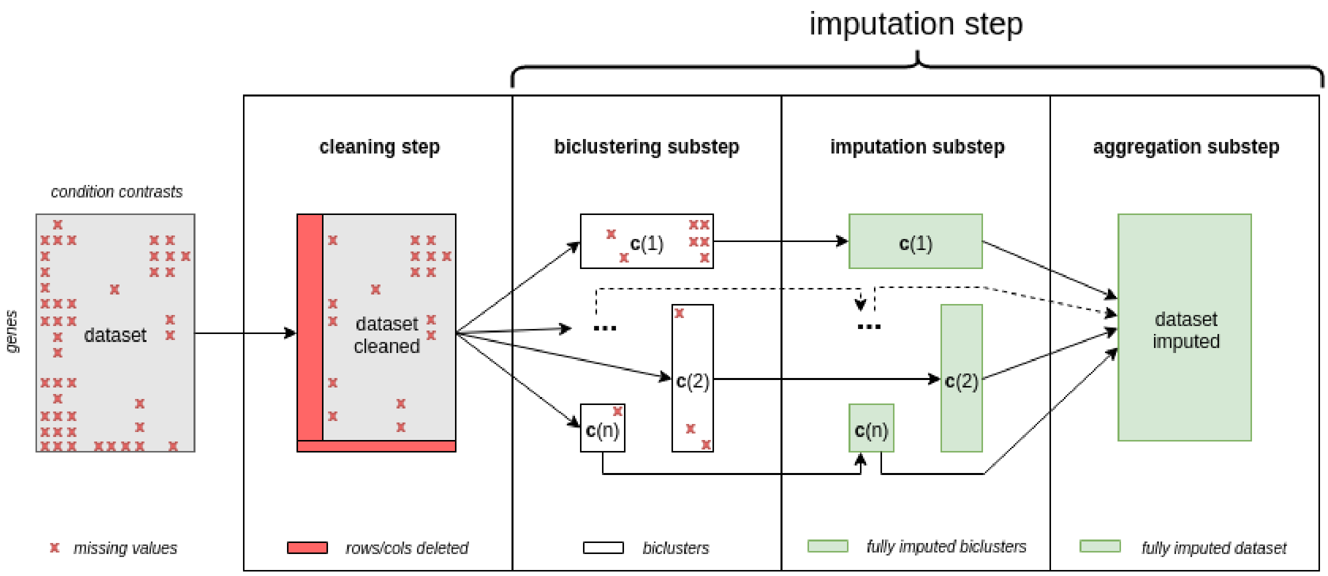

Figure 6: The workflow from the original dataset to the fully imputed one. As it can be seen, the imputation step involves three substeps: the biclustering substep, the biclusters imputing substep and the aggegation substep.

as regards biclustering of gene expression data and [4] that explains a slightly different solution to the problem.

6. EXPERIMENTAL EVALUATION

In order to determine the best way to approach the problem, an accurate experimental evaluation of the imputing techniques was done. The biclustering approach used wasn't evaluated because it wasn't the purpose of this research; however, a performance evaluation between the several biclustering techniques can be found in [5].

Before doing the evaluation, the dataset was divided into three parts: a *training set*, a *test set* and a *validation set*. The validation set was setted aside before the biclustering step, in order to evaluate the performances of each imputation technique in a correct and more robust fashion, allowing to assess the model in a later stage. Thus, two steps were involved in the experimental evaluation: the performance evaluation step and the model assessment step. More precisely:

- **Performance evaluation**: in this step only the training data in each bicluster was used by the imputing algorithms in order to predict the test data. In this way, validation data wasn't involved in the whole process;
- **Model assessment**: in this step the training data and the test data in each bicluster were used by the imputing algorithms in order to predict the validation data, allowing to validate our model.

The evaluation of interest mainly concerns the imputation substep. Thus, four biclusters of different dimensions were chosen at random from all the biclusters generated by the *xMotifs* algorithm. For each of them, the following steps were performed:

1. **Imputation**: all the imputing techniques that were chosen before (*raw mean*, *WkNN*[5], *SVM*, *aSVT*, *SVT*) were applied in order to predict the missing values;
2. **MSE calculation**: an error measure (*Mean Squared Errors*) between the predicted values and the expected values was computed in order to determine the best imputing technique.

After that, an average of the MSE of each imputation technique was calculated in order to choose the best imputation technique for this task.

It is important to note that the missing values inserted had the same proportion of the missing values of the whole dataset ($\sim 30\%$), and that the distribution of missing values was preserved in order to better simulate the real situation of the data.

6.1 Performance evaluation

The methodology explained previously led to the results in Figure 7, also represented in Table 2. As it can be seen, the worst imputing technique is represented by the raw mean, with an average MSE of 0.085. On the other hand, SVT is the technique that performs better in this kind of tasks. In fact, it produced the lowest MSE (with an average of 0.078). The other techniques produced results mainly between them without substantial differences.

6.2 Model assessment

In this evaluation step, each algorithm was tested again in order to assess the models used. This was possible by predicting the validation data using all the remaining data.

[5] After a preliminary testing to evaluate the best k to choice, it was setted to the half of the total genes of the bicluster.

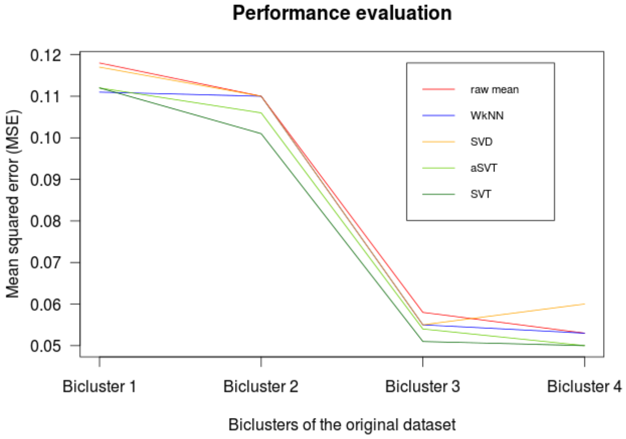

Figure 7: Performance evaluation of the five different imputation techniques in predicting values on four biclusters.

Thus, the predictions were compared with the expected values. The results (showed in Figure 8) confirmed the accuracy of the predictions of the SVT, with slightly different performances as regards WkNN and SVD.

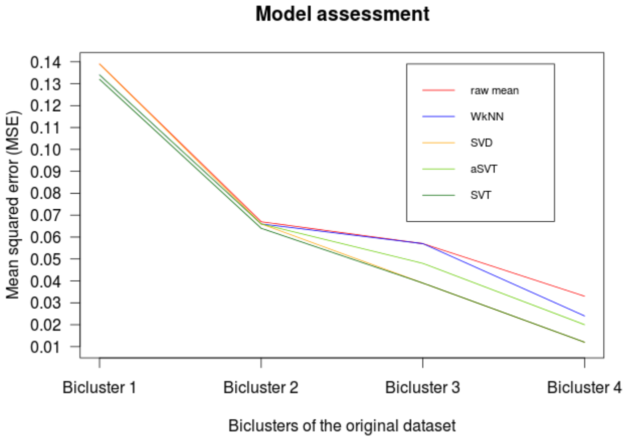

Figure 8: Results assessment of the five different imputation techniques in predicting values on four biclusters.

In the end it is possible to state that, among all the algorithms that were tested, the SVT represent the best approach and the raw mean represents the worst imputation technique.

It is important to note that the randomly chosen biclusters had different sizes in rows and columns and in the total number of values. This allowed us to evaluate in the best way all the process explained before.

7. CONCLUSION

The findings resulted from the evaluation phase allow to better manage the missing values within a gene expression dataset and then to choose the best imputing technique for this kind of tasks: the singular value thresholding (using a threshold of 100). It was proved that it represents a very good approach that is directly applicable to large problems of this kind with a large percentage unknown entries [2]. Despite this novel process works really good, a planned future work concerns the parallelization of the imputing substep (Figure 6), in order to reduce even more the overall computational time of the whole process. This research highlighted, among the other things, the fact that biclustering and imputing must be done in the same time in order to get the best predictions, and that the preprocessing step represents a crucial phase that involves all the subsequent phases. In the end, this research underlined the difficulties that occour in a real scenario[6], allowing to analyze them and to handle them in a proper way.

	Evaluation		**Results**
	test avg	validation avg	overall results
raw mean	0.085	0.074	**0.080**
WkNN	0.082	0.070	**0.076**
SVD	0.085	0.064	**0.075**
aSVT	0.080	0.067	**0.074**
SVT	0.078	0.062	**0.070**

Table 2: A comparison between the various imputation techniques as regards the performance evaluation. Each represented number in test and validation columns is an average of the results of each bicluster test / validation.

8. REFERENCES

[1] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 2011.

[2] J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *Society for Industrial and Applied Mathematics*, 2010.

[3] Y. Cheng and G. M. Church. Biclustering of expression data. *Proceedings: International Conference on Intelligent Systems for Molecular Biology*, 2000.

[4] F. O. de Franca, G. P. Coelho, and F. J. Von Zuben. Predicting missing values with biclustering: A coherence-based approach. *Pattern Recognition*, 2013.

[5] K. Eren, M. Deveci, O. Kucuktunc, and U. V. Catalyurek. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 2012.

[6] M. Grana. Handling varying amounts of missing data when classifying mental-health risk levels. *Innovation in Medicine and Healthcare*, 2014.

[7] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica, Stanford University*, 2002.

[8] T. M. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*, 2003.

[9] P. Schmitt, J. Mandel, and M. Guedj. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 2015.

[6] The dataset of *V. Vinifera* was granted by the Edmund Mach Foundation. Reference: http://www.fmach.it/eng.

[10] S. van Buuren. *Flexible Imputation of Missing Data.* Chapman & Hall/CRC, 2012.

[11] Wikipedia. Biclustering page. `https://en.wikipedia.org/wiki/Biclustering`, 2015.

APPENDIX

A. OTHER PLOTS

In this section some other plots are presented in order to give a complete view of the amount of missing values remained after the cleaning step. In Figure 9 is represented the amount of missing values with the genes on axis x and with the condition contrasts on axis y. In Figure 10 is represented the amount of missing values with the condition contrasts on axis x and with the genes on axis y.

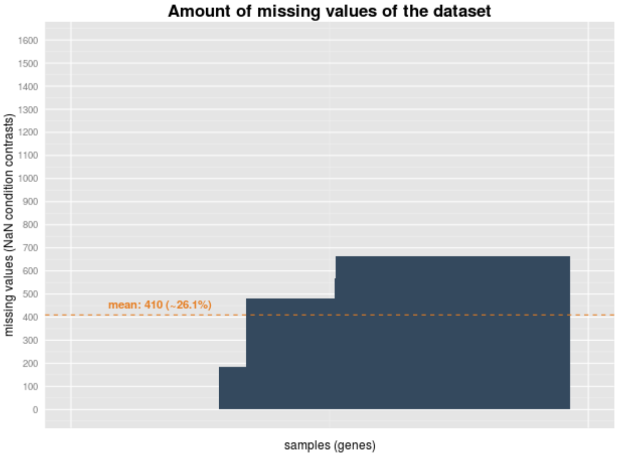

Figure 9: Amount of missing data of the cleaned dataset (x axis: samples, y axis: features).

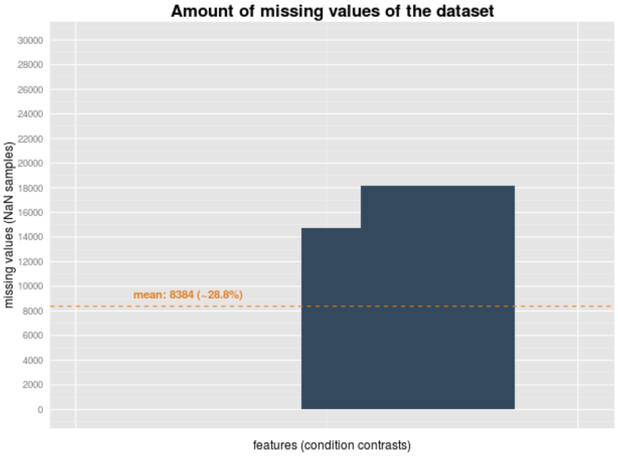

Figure 10: Amount of missing data of the cleaned dataset (x axis: features, y axis: samples).