On the impact of missing value imputation methods for multiple kernel learning on bipartite graphs

Victor Vidal[1], **Tássia Bastos**[1], **Rafael Ferreira Mello**[1,2],
Péricles Miranda[1], **André C. A. Nascimento**[1,2]

[1]Departamento de Computação – Universidade Federal Rural
de Pernambuco (UFRPE), Recife, Brazil

[2]Cesar School, Recife, Brazil

{victor.vidal, andre.camara}@ufrpe.br

Abstract. *In the last decade, the study of pharmacological networks has received a lot of attention, given its relevance to the drug discovery process. Many different approaches for predicting biological interactions have been proposed, especially in the area of multiple kernel learning (MKL). Such methods comprise integrative approaches that can handle heterogeneous data sources in the form of kernels, but can suffer from the missing data problem. Techniques to handle missing values in the base kernel matrices can be used, usually based on simpler techniques, such as imputing zeroes, mean and median of the kernel matrix. In this work, techniques for handling missing values were evaluated in the context of bipartite networks. Our analyses showed that depending on the amount of missing data, k-NN and Singular Value Decomposition (SVD) techniques performed much better than the other techniques, bringing encouraging results, while zero-fill showed the worst performance in relation to all other evaluated methods.*

1. Introduction

With the constant growth and aging of the world population, large health challenges such as combating several types of cancer and infectious diseases, diabetes and neurodegenerative diseases are in great need for innovations. Despite this context, the rapid and economic development of new drugs is far from meeting this demand [Peter Csermely 2013]. The slow pace in drug development is due to the large amount of risks involved, these risks end up causing an excess of caution in the pharmaceutical industry. [C. Chong 2007]

The analysis of the evolution of structure-activity relationship patterns and topology of drug-target networks have demonstrated a pattern in which more than 80% of new drugs tend to bind to targets, which are also connected to other drugs in an organism biological network [Murat Cokol 2005, Peter Csermely 2013]. Thus, an excellent way to mitigate the risks associated with the development of new drugs is to use previous knowledge. In this context, it is possible to understand the notoriety that drug-protein networks have received in recent years[Andre Nascimento 2016]. However, the techniques that use these networks suffer in terms of feasibility, especially in the presence of missing data [R Rivero 2017]. Multiple factors can contribute to the occurrence of missing values in biological data, including experimental factors, laboratory equipment limitations, or the high cost of data acquisition [Jin et al. 2021].

In data science literature, the simplest technique used to heandle missing data, in cases where plenty of data is available, is simply the removal of incomplete instances. However, when missing data is deleted, the size of the sample space is reduced, which can lead to a considerable loss of statistical power. Other techniques known in the literature such as imputation with zero and mean also have their demerits [R Rivero 2017]. Hence choosing the right technique is essential.

Previous works have explored the impact of different missing value imputation methods in the context of metabolomics network data [Wei 2018]. However, this study was limited to an unipartite graph scenario, i.e., the underlying network is composed by only a single type of node (e.g., proteins). In this work, we extend previous studies to the bipartite network context. Besides, we also consider the impact of different data imputation methods on a Multiple Kernel Learning (MKL) setting. MKL is a form of intermediate data integration [Andre Nascimento 2016] method, that combines kernels from multiple sources with a data-oriented approach, this makes it possible to use different notions of similarity and improve accuracy[M. Gonen 2011, F Aiolli 2015]. Thus, we perform a systematic analysis of the effect that the simpler absent value imputation techniques such as zero, mean and median, as well as more complex techniques such as SVD (Singular-Value Decomposition) and imputation by k-NN (k-Nearest Neighbor) have in the performance of prediction methods of drug-target interactions. The effects of imputation are analyzed in a kernel-based learning algorithm, the pairwiseMKL, originally proposed by [A. Cichonska 2018]. This algorithm has a better performance compared to traditional MKL methods, as its learning step is performed without the explicit calculation of paired matrices. .

This paper is organized as follows: Section 2 presents some relevant previous works. Section 3 describes the imputation techniques considered as well as the pairwiseMKL algorithm. Section 4 presents the dataset used on this work and Section 5 details the experimental setting. Section 6 presents and discusses the experiment results. Finally, section 7 presents the conclusions that were found.

2. Related work

Kernel methods in computational biology have a lot of potential to facilitate data integration from a myriad of heterogeneous sources. However, information contained in these biological databases is often incomplete or even missing. Some commonly adopted solutions include the removal of instances whose information is not complete, which takes to a decrease in the data set and consequently in the predictive power of that sample. Some recent studies that use MKL techniques in problems of unipartite have invested in complementing missing values in kernel matrices.

According to [Kumar et al. 2013], the problem of deriving a kernel array from a set of incomplete arrays can be worked around by filling in the missing values. The imputation can be done with the average or simply filling with zeros [Andre Nascimento 2016, R Rivero 2017]. Given the above, the handling of missing data in kernel matrices can be greatly improved, considering recent advances in research into methods of imputing missing values in MKL problems [Liu et al. 2019, Kumar et al. 2013].

Previous work have limited the evaluation of missing data imputation methods in the context of unipartite networks. In the study by [Wei 2018], a comprehensive comparison was performed between eight methods of imputing missing values in the context of metabolomics data based on mass spectrometry, namely: zero, half of the minimum (HM), mean, median, as well as other, machine learning based methods, such as Random Forest (RF), Improved Singular Value Decomposition (iSVD), k Nearest Neighbors Imputation (k-NN) and Quantile Regression Imputation of Left-Censored data (QRILC).

In the context of MKL studies, [R Rivero 2017] proposes the Mutual kernel Matrix Completion (MKMC) algorithm, which exploits the Expectation Maximization (EM) algorithm to minimize the Kullback-Leibler divergence between the base kernel matrices. The results indicate that as the proportion of missing data increases, the algorithm increases its advantage over simpler strategies such as zero or mean. Recent approaches have sought to integrate the combination of kernels and the treatment of incomplete kernels in a single step, in order to reduce the computational cost of treating the two problems in separate steps. The Multiple Kernel Clustering - MKC [Liu et al. 2019] algorithm is proposed that treats missing positions in kernel arrays as auxiliary variables to be optimized, also obtaining encouraging results. This work is extended in the study by [Zhu et al. 2018], in which a localized version of the algorithm is proposed, requiring only to analyze the local neighborhood (k-neighbors) of a sample to estimate the missing values. Recently, the authors proposed an extension incorporating a matrix-induced regularization term to handle the correlation among base kernels[Li et al. 2021].

However, all the works mentioned above developed their experiments to fill in missing values in unipartite networks, and no evidence was observed for the bipartite context. Thus, this research emphasizes the evaluation of techniques for imputing missing values in the bipartite context with the objective of studying the effect that different techniques have on the chosen MKL based predictive model as well as in the learned kernel weights.

3. Methods

In this section, we present the learning algorithm (pairwiseMKL) considered in the experiments, as well as the imputation methods used to fill the missing values in the kernel matrices.

3.1. Pairwise MKL

The pairwiseMKL algorithm is a kernel based method, that can handle diverse, heterogeneous data sources in the form of kernels. Let k_c and k_d be the two kernel functions such that they produce positive semidefinite kernel matrices $K_d \in R^{n_d \times n_d}$ and $K_c \in R^{n_c \times n_c}$ for drugs and cell lines, respectively, where n_d is the number of drugs and n_c the number of cell lines in the dataset. Let also $K = K_d \otimes K_c$ the Kronecker product of such kernels. The learning method is based on the kernel ridge regression (KRR), in which the objective function is defined based on the total quadratic loss associated with an L2-norm regularizer. The combinations of all kernels in the KRR setup can be defined as:

$$\left(\left(\mu_1 K_d^{(1)} \otimes K_c^{(1)} +, ..., +\mu_P K_d^{(P)} \otimes K_c^{(P)}\right) + \lambda I\right) \alpha = y, \qquad (1)$$

where $\mu_{(i)}$ is the weight associated to the i-th pairwise kernel combination. The solution the system of linear equations above proposed by [A. Cichonska 2018] uses the conjugate gradient (CG - Conjugate Gradient) method, iteratively, until reaches convergence.

In summary, pairwiseMKL, initially performs a centralized kernel alignment procedure in order to avoid the explicit calculation of several — usually large — arrays of pairs in the selection of the mixing weights of the input pair kernels. For this, [A. Cichonska 2018] did a new decomposition of Kronecker from the centering operator to the kernel pair by pair. That is, the algorithm generates a measure of matrix similarity between the final kernel and the ideal kernel, derived from label values (Gaussian response kernel), from a convex combination of kernels in input pairs. That is, this approach makes the method suitable for solving problems in the context of large paired spaces, which is the case of drug bioactivity prediction.

3.2. Imputation Methods

In this work, three single-value imputation techniques (mean, median and zero) and two supervised imputation techniques (k-NN and Improved Singular Value Decomposition, iSVD) were evaluated. The choice of techniques was determined by the high rate of use in other studies and for having validated results in the literature [Wei 2018].

The first three techniques used correspond to the imputation of simple values — mean, median and zero. For the mean technique, the average of each matrix was calculated and each result was used to fill in the missing values of the matrices, respectively. The median technique has a similar development process to the average technique, but in this case using the median value of the matrix to fill in the missing spaces [Wei 2018]. The third and final simple imputation technique comprises filling in the gaps directly with the number zero [Tuikkala 2008].

The k-NN algorithm is one of the most popular machine learning methods, given its simplicity and good results in diverse learning tasks. Its use as an imputation method is rather common, specially in the context of microarray data[Wei 2018]. For each position with missing values, the algoritms finds the k nearest values based on the Euclidean metric, and missing values are replaced by the average of its neighbors. In this work we used the k-NN implementation available in the *fancyimpute* library [Alex Rubinsteyn] with 3 as the k value.

The Singular Value Decomposition (SVD) algorithm — is considered to be the basis of the most accurate methods when the objective is to solve least squares problems, and especially to determine the null space of matrices. SVD is the most reliable matrix decomposition/factoring method, but its use requires a longer execution time [Yuan et al. 2019]. Aiming to improve the performance of SVD, [Kurucz et al. 2007] brings a modification in the traditional implementation of Lanczos code, which allows the imputation of missing data as well as the handling of very large input databases. In this work, we used the improved SVD (iSVD) technique proposed by [Kurucz et al. 2007] and also implemented in [Alex Rubinsteyn].

4. Dataset

The dataset used for evaluation was extracted from an anticancer drug response database from the GDSC (Genomics of Drug Sensitivity in Cancer) project, originally proposed by [Yang 2012], and used in [A. Cichonska 2018]. The data is constituted by the responses of 124 human cancer cell lines to 124 drugs, thus, 15,376 measures of sensitivity are available in the form ln(IC50), in nano molar values. [Ammad-Ud-Din 2016]

The histogram, in Figure 1, presents the distribution of bioactivity values. It is possible to observe that the data follows a normal distribution, where the highest concentration of data is in the affinity range 0 to 5.

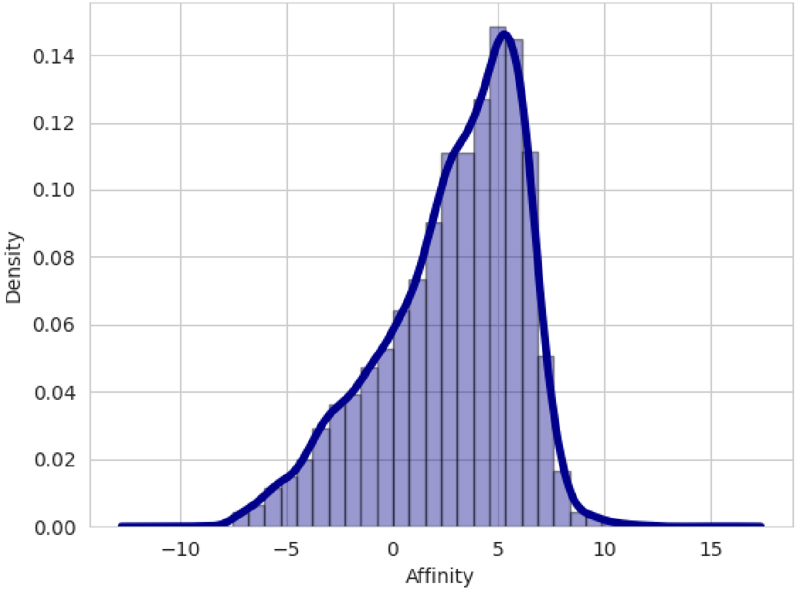

Figure 1. Interaction affinity histogram.

The inputs to the pairwiseMKL algorithm consists of a set of distinct drug and cell line kernels. In this work, we used the same kernels considered in [A. Cichonska 2018], with different artificially created missing values. A brief description of the kernels are given below (Table 1), for more information, see [A. Cichonska 2018].

- **Drug Kernels:** A total of 10 distinct drug kernels were considered and all were calculated over fingerprint molecular descriptors
- **Cell Line kernels:** The construction of cell line kernels was based on the calculation of Gaussian kernels, over measurements of copy number variation of 43,255 genes (Kc-cn-XXX), basal gene expression measurements of 13,321 genes (Kc-exp-XXX), methylation levels of 482,892 CpG islands and actual value profiles of 12,366 somatic mutations (Kc-mut-XXX) [A. Cichonska 2018], totaling 12 cell line kernels.

5. Experimental Setting

In order to evaluate the performance of the learning algorithm [A. Cichonska 2018] in the context of bipartite networks, a systematic procedure was carried out to evaluate the effectiveness of the method when using an incomplete heterogeneous biological dataset as input. The evaluation experiment can be described in 3 phases: generation of missing data, imputation of missing values and the model's training/prediction. The experiment was carried out on a Debian SO in a machine with 8 CPUs, 2 GB of RAM and 3.7.13

Type	Kernel Name	Feature description and kernel type
Drug	Kd-circular	Extended Connectivity 1024-bit fingerprint (ECFP6).
Drug	Kd-estate	79-bit fingerprint corresponding to the 'Estate' substructures.
Drug	Kd-ext	Path-based 1024-bit block fingerprint, taking ring systems into account.
Drug	Kd-graph	1024-bit block fingerprint based on path, considering connectivity.
Drug	Kd-hybr	1024-bit block fingerprint based on path, considering hybridization states.
Drug	Kd-kr	4860-bit fingerprint [Klekota and Roth 2008].
Drug	Kd-maccs	166-bit fingerprint based on MACCS structural keys.
Drug	Kd-PubCh	881-bit fingerprint defined by PubChem.
Drug	Kd-sp	1024-bit fingerprint based on the shortest paths between atoms, taking into account ring and charge systems.
Drug	Kd-std	1024-bit block fingerprint based on path.
Cell Line	Kc-cn-146	Copy number data, with Gaussian kernel ($\sigma = 146$).
Cell Line	Kc-cn-270	Copy number data, with Gaussian kernel ($\sigma = 270$).
Cell Line	Kc-cn-417	Copy number data, with Gaussian kernel ($\sigma = 417$).
Cell Line	Kc-exp-147	Gene expression data, with Gaussian kernel ($\sigma = 147$).
Cell Line	Kc-exp-163	Gene expression data, with Gaussian kernel ($\sigma = 163$).
Cell Line	Kc-exp-177	Gene expression data, with Gaussian kernel ($\sigma = 177$).
Cell Line	Kc-met-176	Methylation data, with Gaussian kernel ($\sigma = 176$).
Cell Line	Kc-met-210	Methylation data, with Gaussian kernel ($\sigma = 210$).
Cell Line	Kc-met-252	Methylation data, with Gaussian kernel ($\sigma = 252$).
Cell Line	Kc-mut-57	Somatic mutations data, with Gaussian kernel ($\sigma = 57$).
Cell Line	Kc-mut-71	Somatic mutations data, with Gaussian kernel ($\sigma = 71$).
Cell Line	Kc-mut-132	Somatic mutations data, with Gaussian kernel ($\sigma = 132$)

Table 1. Different configurations of cell line and drug kernels considered. Source: [A. Cichonska 2018]

python version. Figure 2 shows a graphic representation of the imputation stage of the experiment.

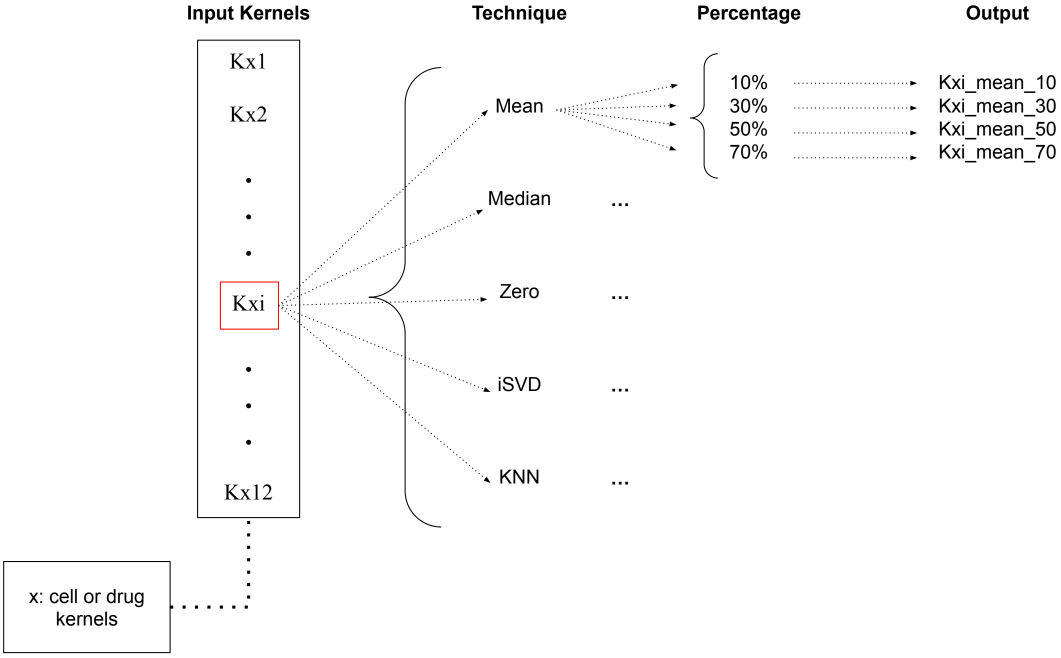

Figure 2. Simulation of missing data methodology.

The first phase of the experiment consisted of generating missing data in the previously described kernel matrices. With this purpose, an algorithm was implemented that received as input a complete kernel matrix and a percentage of missing values that should be generated, then the algorithm replaced randomly selected values from the matrices by missing values for distinct proportions (10%, 30%, 50% and 70% of the data in each kernel matrix) until the desired proportion was reached (Figure 2). It is worth noting that the properties inherent to the kernel matrices were maintained after the algorithm was executed. The described algorithm was executed for each of the 22 kernel matrices existing in the database and each of the percentages of missing values chosen, totaling 88 executions. It is important to notice that the imputation of missing values in all kernels was done randomly, that is, no matrix has exactly the same deleted positions,

In this work, three single-value imputation techniques (zero, mean and median) and two supervised imputation techniques (iSVD and kNN with k=3) were evaluated. Some of these methods were also evaluated in previous data imputation studies [Wei 2018].

Then, the training and prediction process was performed using the original algorithm proposed by [A. Cichonska 2018], with a small modification. The number of inner folds used in the cross-validation process was reduced from 3 to 1. This modification was carried out in order to reduce the execution time of the algorithm. The training and cross-validation process was performed separately on each set of 22 imputed kernels using each of the chosen techniques and percentage of missing data. The result of the process described previously were 3 text files, one for each chosen evaluation metric, in which each line represents the value of the metric in question for each outer fold of the cross-validation performed in the combination technique-percentage. Hence this step resulted in 3x4x5=60 result files.

5.1. Evaluative metrics

In order to assess the impact of the missing data in this setting, we adopted the evaluation metrics considered in [A. Cichonska 2018]. Those were: F1-score, Pearson's correlation coefficient (r) and root mean squared error (RMSE). F1 can be defined as the harmonic mean of the recall and precision of the [Dalianis 2018] model predictions.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Pearson's correlation coefficient (r)corresponds to the degree of linear association between two quantitative variables [Liu 2020]. Correlation analysis, in general, starts with the graphical representation of the relationship of data pairs through the use of a scatter diagram. Pearson's correlation coefficient can be defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

The coefficient corresponds to a dimensionless index and the values range from -1 to +1, reflecting the strength of a linear relationship between two sets of data. That is, positive values indicate a tendency for one variable to increase or decrease together with another, and negative values indicate a tendency for the increase in the values of one variable to decrease the other. Values close to zero indicate low association. [Kirch 2008]

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2}$$

The root mean squared error (RMSE) corresponds to the square root of the mean square of all errors. RMSE is widely used and considered a great general-purpose error metric for numerical predictions. [S. P. Neill 2018]

6. Results and discussion

In this section, the results obtained by each imputation method will be discussed. Initially, experiments were performed with the pairwiseMKL with the complete kernels. The results obtained in the original scenario can be used as a baseline model to evaluate the performance of the techniques used. The closer to the original metrics, the better the imputation method. Table 2 presents the evaluation metrics for each technique-percentage combination. One can note that the imputation by zero, presented the lowest F1-score and Pearson's coefficient value accross all missing data percentages. The imputation by zero also obtained the highest RMSE in all iterations. It is also possible to observe that the mean and median imputation techniques maintained similar values in their metrics in all iterations, but it is interesting to observe that these two techniques presented values slightly higher than the supervised imputation techniques when 70% of the kernel matrices were missing.

Then, it is possible to notice that the iSVD algorithm achieved better performance than the the other supervised technique (k-NN) when the percentage of missing values were 30% and 50%. However, it is interesting to note that the k-NN technique ($k = 3$) presented considerably better results than the iSVD when a lower percentage of missing values is observed, but there was a gradual degradation in its performance as the percentage of missing numbers increased. Such behavior may be due to the low number of the chosen parameter k. The evolution of the metric values according to the increase in the percentage of missing data is presented in Figure 3 and 4. It is possible to state that the second-greatest degradation belongs to the supervised technique k-NN.

The analysis of the kernel weights assigned by the pairwiseMKL algorithm can be used to verify the method's ability to correctly identify the most relevant information sources. The approach adopted here was a simple and individual analysis of the average weights of each drug kernel-cell lines combination, accross all folds, expressed in the form of heatmaps (Figure 5).

Regarding the analysis of the weights, given that the choice of deleted positions in all kernels was done randomly, there is an impediment in the direct comparison between the weights of the applied techniques. However, it is possible to observe that the distribution of weights in the original scenario was very similar to the distribution presented by most imputation methods, with the exception of the imputation by zero. Such a similar

Table 2. Comparative analysis of metrics for each technical-percentage combination

Technique	Percentage	F1-score	Pearson	RMSE
Baseline	-	0.6303	0.8576	1.6816
Zero	10%	0.562628	0.752361	2.24138
Mean		0.604825	0.804185	2.79702
Median		0.597916	0.786369	2.59705
iSVD		**0.627187**	**0.859277**	**1.67534**
KNN		0.618058	0.846715	1.74128
Zero	30%	0.534463	0.67539	2.70722
Mean		0.559874	0.701951	2.73056
Median		0.581671	0.753536	4.00589
iSVD		**0.625167**	**0.858552**	**1.67919**
KNN		0.595823	0.830259	1.82613
Zero	50%	0.500251	0.57906	3.40646
Mean		0.574567	0.731764	4.6214
Median		0.579967	0.750462	3.61004
iSVD		**0.594521**	0.770161	2.96752
KNN		0.562707	**0.771577**	**2.15227**
Zero	70%	0.476902	0.46762	4.35748
Mean		**0.597821**	**0.816953**	**1.90404**
Median		0.581574	0.763762	2.35843
iSVD		0.562527	0.772375	2.14511
KNN		0.521095	0.663076	2.78747

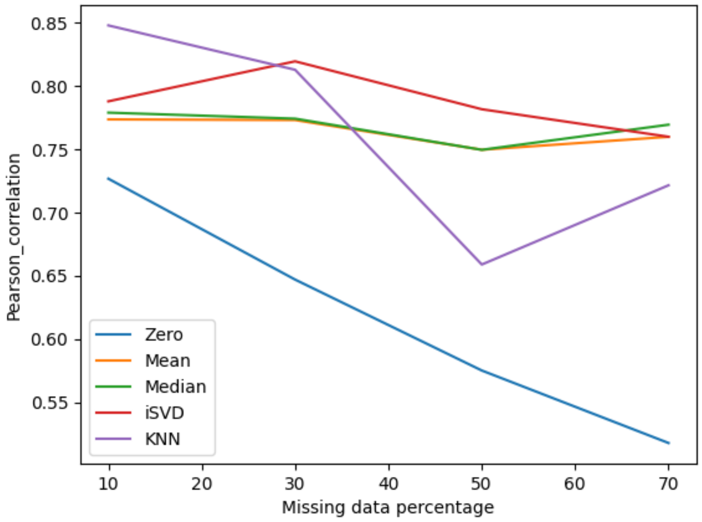

Figure 3. Impact on Pearson correlation of the missing data percentage

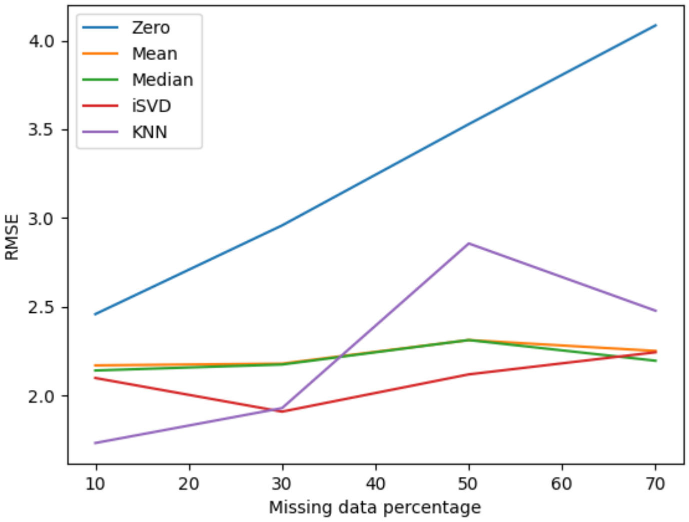

Figure 4. Impact on RMSE correlation of the missing data percentage

distribution may have influenced the superior performance of the k-NN technique under the 10% missing data scenario, when compared to the original baseline setup.

For more than 10% of missing values, the iSVD technique presents very similar results in all applied percentages, being considered for this data set the most efficient technique among the others. As iSVD-based estimation is essentially a method of linear regression in a lower dimensional space, this performance degradation is not surprising for non-time series data, where a clear expression pattern is often not present.

According to the literature, the averaging technique is one of the most used for imputing missing values [A. Cichonska 2018, Wei 2018, Zhang 2016]. Although it presented better results in relation to replacing the missing values by zero, the average technique yielded a lower precision than that presented by the iSVD technique. It is important to highlight the results obtained with the median imputation strategy, in the most challenging 70% missing data setting. However, such setting is rather extreme, and probably a more profound investigation is needed to address the specificity under this scenario.

The fact is that simple value imputation techniques present inferior performance in the 30% and 50% settings, when compared to other more sophisticated techniques, as is the case of iSVD. Therefore, more sophisticated methods like iSVD coupled with pairwiseMKL provide more accurate ways to estimate missing values in the considered drug bioactivity interaction dataset. The iSVD technique, applied in the context of bipartite networks, presented a much superior performance in relation to the simplest solutions, taking advantage of the correlation structure of the data to estimate the missing expression values.

7. Conclusion

The research presented in this article aims to study the use of different techniques for imputing missing values in the context where kernel methods are used to predict drug-protein interactions. Through the experiment carried out, it was possible to obtain a deeper knowl-

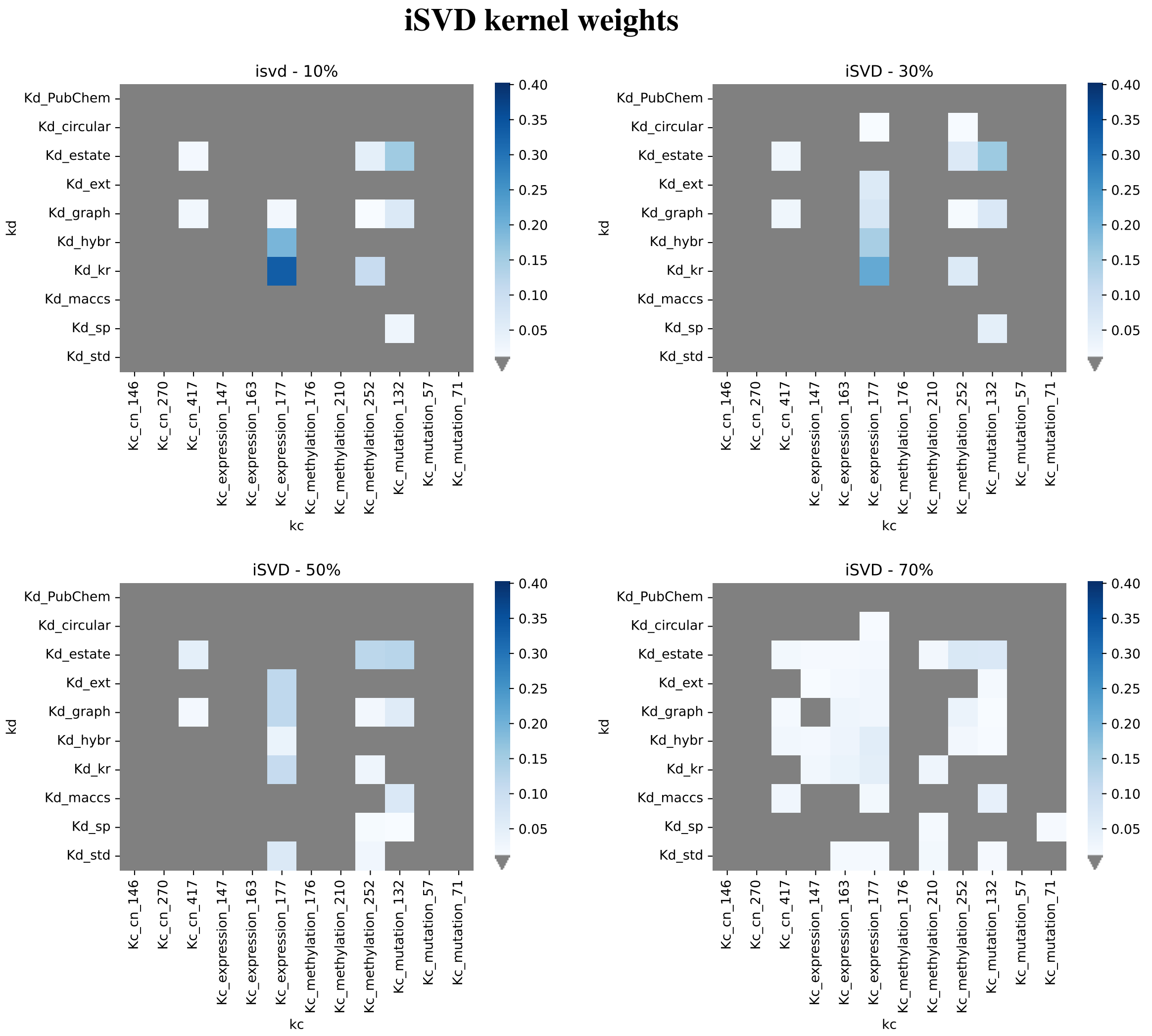

Figure 5. Mean kernel weights for iSVD imputation for 10%, 30%, 50% and 70% of missing values. Values lower than 0.01 are shown as gray.

edge about the effects that the increase in the percentage of missing values has on each technique applied, the results showed that supervised imputation techniques have better performances than the techniques of imputation by single value when the percentage of missing data is low, but it was also possible to observe that simpler techniques such as imputation by mean and median may be desirable in cases where the percentage is high.

However, the authors highlight that a more detailed investigation of the impact of imputation methods is needed. More specifically, with the addition of more datasets, as well as additional learning algorithms [Chen and Zhang 2021] and the use of the supervised imputation technique KNN with different values of k. The incorporation of data imputation steps into MKL algorithms is another possible venue for future works.

References

A. Cichonska, T. Pahikkala, S. S. H. J. A. A. M. H. T. A. J. R. (2018). Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics, Oxford University Press.*

Alex Rubinsteyn, S. F. fancyimpute: An imputation library for python.

Ammad-Ud-Din (2016). Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics, Oxford University Press*.

Andre Nascimento, Ricardo Prudêncio, I. C. (2016). A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics*.

C. Chong, D. S. (2007). New uses for old drugs. *Nature*.

Chen, J. and Zhang, L. (2021). A survey and systematic assessment of computational methods for drug response prediction. *Briefings in bioinformatics*, 22(1):232–246.

Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical text mining, Springer*.

F Aiolli, M. D. (2015). Easymkl: a scalable multiple kernel learning algorithm. *Neurocomputing*.

Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., and Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific reports*, 11(1):1–11.

Kirch, W. (2008). Pearson's correlation coefficient. *Encyclopedia of Public Health, Dordrecht: Springer Netherlands*.

Klekota, J. and Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525.

Kumar, R., Chen, T., Hardt, M., Beymer, D., Brannon, K., and Syeda-Mahmood, T. (2013). Multiple kernel completion and its application to cardiac disease discrimination. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 764–767. IEEE.

Kurucz, M., Benczúr, A. A., and Csalogány, K. (2007). Methods for large scale svd with missing values. In *Proceedings of KDD cup and workshop*, volume 12, pages 31–38. Citeseer.

Li, M., Xia, J., Xu, H., Liao, Q., Zhu, X., and Liu, X. (2021). Localized incomplete multiple kernel k-means with matrix-induced regularization. *IEEE Transactions on Cybernetics*.

Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., Kloft, M., Shen, D., Yin, J., and Gao, W. (2019). Multiple kernel k k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1191–1204.

Liu, Y. (2020). Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Processing Letters, Springer*.

M. Gonen, E. A. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*.

Murat Cokol, Ivan Iossifov, C. W. A. R. (2005). Emergent behavior of growing knowledge about molecular interactions. *Nat Biotechnol*.

Peter Csermely, Tamás Korcsmáros, H. J. K. G. L. R. N. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery a comprehensive review. *Pharmacol Ther*.

R Rivero, R Lemence, T. K. (2017). Mutual kernel matrix completion. *IEICE*.

S. P. Neill, R. M. H. (2018). Fundamentals of ocean renewable energy: generating electricity from the sea. *Academic Press*.

Tuikkala, J. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC bioinformatics, BioMed Central*.

Wei, R. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports, Nature Publishing Group*.

Yang, W. (2012). Genomics of drug sensitivity in cancer: a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research, Oxford University Press*.

Yuan, X., Han, L., Qian, S., Xu, G., and Yan, H. (2019). Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems*, 163:485–494.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1).

Zhu, X., Liu, X., Li, M., Zhu, E., Liu, L., Cai, Z., Yin, J., and Gao, W. (2018). Localized incomplete multiple kernel k-means. In *IJCAI*, pages 3271–3277.