SCHEDAE INFORMATICAE

XXXXXXXXX VOLUME nn Year

Regression SVM for incomplete data

Łukasz Struski Marek Śmieja Bartosz Zieliński Jacek Tabor
Faculty of Mathematics and Computer Science
Lojasiewicza 6, 30-348 Kraków
e-mail: *lukasz.struski@uj.edu.pl*

Abstract. The use of machine learning methods in the case of incomplete data is an important task in many scientific fields, like medicine, biology, or face recognition. Typically, missing values are substituted with artificial values that are estimated from the known samples, and the classical machine learning algorithms are applied. Although this methodology is very common, it produces less informative data, because artificially generated values are treated in the same way as the known ones. In this paper, we consider a probabilistic representation of missing data, where each vector is identified with a Gaussian probability density function, modeling the uncertainty of absent attributes. This representation allows to construct an analogue of RBF kernel for incomplete data. We show that such a kernel can be successfully used in regression SVM. Experimental results confirm that our approach capture relevant information that is not captured by traditional imputation methods.

Keywords: regression SVM, incomplete data, missing attributes, RBF kernel.

1. Introduction

The incomplete data problem exists in a wide range of scientific fields, like medical diagnosis, clinical trials [13], psychology [1], or face recognition [22]. Inappropriate treatment of incomplete data (in which certain feature values are missing for particular samples) may cause large error or even false classification [8], therefore the ability of handling such a data is fundamental. One of the reasons for missing values in medical data sets is bad health condition of the patient, which prevents from performing sensitive and time-consuming examinations. In psychology, a client can refuse to answer some of the questions. While, in face recognition, face can be partially occluded by the other objects. Since classical learning algorithms cannot be directly applied to incomplete data sets, their adaptations are necessary.

Usually, missing values are substituted with artificial values that are estimated from the known samples. As a result, the complete data set is obtained and classical machine learning methods can be applied [16]. The most straightforward candidates for imputation are the mean values or medians (generated separately for each feature based on known values). One could also use the mean value of k Nearest Neighbors (kNN, see [14]). Such imputation techniques are easy to applied by the practitioners, but they produce less informative data, because artificially generated values are treated in the same way as the known ones.

There are also non-deterministic imputation methods, which estimate a distribution of the incomplete data set and use it to sample values of unknown features [17]. Such a distribution can be generated with Expectation Maximization algorithm (EM, see [9]) under some assumptions on missing data, however only if data are Missing at Random (see the next section for details). This however is difficult to verify in practice, therefore another possibility is to apply chained equations, which generate multiple imputations. Such approach produces very good results, but on the same time it increases the computational time, as many variants of the same data set need to be generated and then analyzed (see [4] for more details).

Some of the classification and regression algorithms use data distribution directly, without imputation stage. Such algorithms were proposed, among others, for logistic regression [24], kernel methods [20, 23], or for the second order cone programming [18]. Moreover, a few algorithms use raw incomplete data, without generating data distribution and without imputation stage. One of such algorithms, proposed in [6], trains Support Vector Machine (SVM) by scaling the margin with respect to known features of incomplete samples. The other approach, presented by [10], constructs the embedding mapping of feature-value pairs together with a classification objective function.

Understanding the reasons why data are missing is important to correctly handle the remaining data. If data are Missing Completely at Random (MCAR) then there is no relationship between whether a data point is missing and any values in the data set. Missing at Random (MAR) means that the absence of a feature is not related to the missing data, but it is related to some of the observed data. In both cases, the probability distribution on data space X can be estimated with EM algorithm. In general, data might be neither MAR nor MCAR, but this case is more difficult to handle and will not be considered in this paper.

In [21], the authors propose to create an analogue of classical Radial Basis Function kernel (RBF), based on a Gaussian estimation of data distribution. Its basic idea relies on modeling uncertainty of missing values with probability measures. After applying some necessary transformations, such a probabilistic representation is pushed into classical scalar product in L^2 space, producing a kernel matrix in result.

In this paper, we examine this approach in the case of regression SVM method (r-SVM). R-SVM [19] predicts the values of target feature from the set of input features, while ignoring the errors smaller than a fixed distance $\varepsilon > 0$ (this provides higher stability of prediction). We test this approach on data sets from UCI repository [3], with artificially removed data. The experiments confirm that our approach is more accurate than traditional imputation methods.

2. Model

Let $(x_i, y_i)_i \subset X \times \mathbb{R}$, where $X \subset \mathbb{R}^N$, be training data. In the simplest case of linear functions $f = \langle w, x \rangle + b$, where $\langle \cdot, \cdot \rangle$ denotes a scalar product on X, r-SVM aims to minimize:

$$L(w, b) = \frac{1}{2}\|w\|^2 + C \sum_i (\xi_i + \xi_i^*)$$

subject to:

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \end{cases}$$

where ξ_i and ξ_i^* are the slack variables, which allow to define *soft margin* [19], see Figure 1.

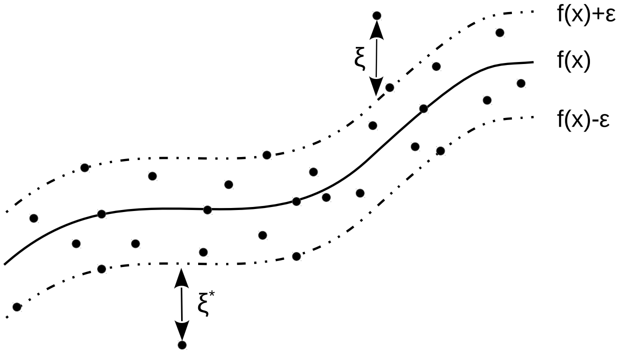


Figure 1.: A schematic diagram of SVR.

The constant $C > 0$ determines the trade-off between the flatness of f and the threshold up to which deviations larger than ε are tolerated. Usually, such optimization problem can be solved in the dual form, where the arbitrary kernel functions extends it to a nonlinear situation. Therefore, most of the effort goes into defining an appropriate kernel mapping for a particular problem. In remaining part of this section we define kernel mapping for incomplete data (based on [21]).

Let us assume that a data set $X \subset \mathbb{R}^N$ is incomplete. Since we do not know the values of vector $x \in X$ on some coordinates $J_x \subset \{1, \ldots, N\}$, we can define an affine subspace $x + V_x$, where $V_x := \mathrm{span}(e_j)_{j \in J_x}$ and $(e_j)_{j=1}^N$ denotes a canonical basis of

$\mathbb{R}^N$. More generally, we associate every missing data point x with an affine subspace $x + V_x$. We omit subscript x in V_x and J_x due to the clarity of the equations.

Following [21] our main idea is to obtain density estimation F of the data-set, and then represent the missing coordinates via F_{x+V}, the conditional density of F on $x + V$. Next we apply the regularization with gaussian kernel (that is the convolution of the resulting density with $N(0, \gamma I)$. Thus the final embedding into a Hilbert space L^2 is given by

$$\Phi : x + V \to F|_{x+V} * N(0, \gamma I).$$

Observe that in the case of complete data (that is when $V = \{0\}$), the above approach yields the classical Gaussian kernel, since $\Phi(x) = \delta_x * N(0, \gamma I) = N(x, \gamma I)$.

Since the method is strongly dependent on the initial density estimation for simplicity of computations and to avoid possible overfitting we therefore consider the class of Gaussian densities. Consequently, we assume that $F = N(m, \Sigma)$ is a Gaussian estimation of X. To model the uncertainty on missing attributes of x, we calculate a conditional density F_{x+V} of F on the affine subspace $x + V$ of $\mathbb{R}^N$. It is well known that the conditional density of Gaussian, is a Gaussian, and if we fix an orthonormal base Q in V (which can be identified with orthonormal projection onto V), then by [21]

$$F_{x+V}(x + y) = N(m_V, \Sigma_V)(y) \text{ for } y \in V,$$

where

$$\begin{aligned} \Sigma_V &:= (Q^T \Sigma^{-1} Q)^{-1}, \\ m_V &:= \Sigma_V [Q^T \Sigma^{-1}(m - x)]. \end{aligned}$$

This is a non-degenerate density in the space $x + V$ of dimension $\#J$. However, this conditional density can be identified with a degenerate Gaussian density $N(m^V, \Sigma^V)$ on $\mathbb{R}^N$, where [21]:

$$m^V := x + Q m_V, \qquad \Sigma^V := Q \Sigma_V Q^T.$$

This view on missing data as a degenerate normal density is better from our point of view, as it allows simple formulas for the regularization.

To define an analogue of RBF kernel, we first compute a convolution between Gaussian estimation of incomplete sample $N(m^V, \Sigma^V)$ and $N(0, \gamma I)$, where $\gamma > 0$, to avoid degenerated measures:

$$N(m^V, \Sigma^V) * N(0, \gamma I) = N(m^V, \Sigma^V + \gamma I). \tag{1}$$

Next, we apply a standard scalar product in L^2 space for two samples $x, y \in X$:

$$\begin{aligned} \langle N(m^{V_x}, \Sigma^{V_x}), N(m^{V_y}, \Sigma^{V_y}) \rangle_\gamma &= N(m^{V_x} - m^{V_y}, \Sigma^{V_x} + \Sigma^{V_y})(0) \\ &= \tfrac{1}{(2\pi)^{N/2} \det^{1/2}(\hat{\Sigma})} \exp(-\tfrac{1}{2} \| m^{V_x} - m^{V_y} \|^2_{\hat{\Sigma}}), \end{aligned} , \tag{2}$$

where $\hat{\Sigma} := 2\gamma I + \Sigma^{V_x} + \Sigma^{V_y}$. This produces a kernel matrix for r-SVM method.

This generalizes classical RBF kernel, because in case of complete samples x and y, we get $m^{V_x} = x$, $m^{V_y} = y$ and $\Sigma^{V_x} = \Sigma^{V_y} = 0$. In consequence $\hat{\Sigma} = 2\gamma$ and

$$\langle N(x, 0), N(y, 0) \rangle_\gamma = \frac{1}{(2\pi)^{N/2} \det^{1/2}(2\gamma)} \exp(-\tfrac{1}{4\gamma} \|x - y\|^2),$$

what is similar to classical RBF kernel, except for different parametrization and normalization.

3. Experimental results

We applied our probabilistic representation of incomplete data to r-SVM and compared it with various imputation-based techniques. We considered imputation with k-Nearest Neighbor (KNN-MV), k-Means Clustering Imputation (KMeans-MV), Support Vector Machines Imputation (SVMimpute-MV), and Multiple Imputation Chained Equations (Mice). In KNN-MV, the k nearest neighbors are found with Euclidean distance and then they are used to impute the missing values [5]. KMeans-MV treats instances from the same cluster as the nearest neighbors of each other and replace missing coordinates, in a way similar to KNN-MV [12]. SVMimpute-MV sets the decision attributes (target attribute) as the condition attributes (input attributes) and the condition attributes as the decision attributes, and uses SVM regression to predict the missing condition attribute values [11]. Mice samples missing values jointly from estimated probability distribution [4]. The parameters of the methods were the same as those presented in [15]: $k = 10$, for KNN-MV; $k = 10$, $\mathit{iteration} = 100$ and $\mathit{error} = 100$ for KMeans-MV; RBF kernel with $C = 1.0$, $\sigma = 0.001$ and no shrinking for SVMimpute-MV. We used KEEL software [2], which does not support parameter tuning for imputation methods. Therefore, we chose the values recommended by their respective authors.

For experiment, we selected six data sets from UCI repository and KEEL data set repository (see Table 1) and we considered two strategies of removing some of their values. The first strategy simulated MCAR mechanism and removed a fixed percentage of features randomly (10%, 20%, and so on, till 80%). In the second strategy, we defined a structural process of attributes removal satisfying MAR assumptions. More precisely, we sampled N points $x_1, \ldots, x_N$ of data set $X \subset \mathbb{R}^N$ and for every $x \in X$ we removed its i-th attribute with a probability $\exp(-t\|x - x_i\|_\Sigma))$, where $\|x\|_\Sigma$ denotes the Mahalanobis norm of x with respect to Σ. The values of t were selected to remove approximately $10\%, 20\%, 30\%, \ldots, 80\%$ of coordinates. Covariance matrix was computed as a sample covariance from training data.

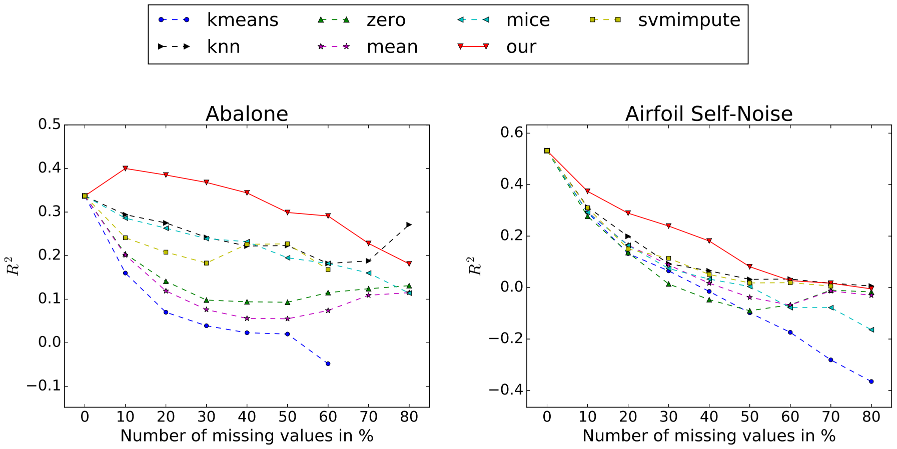
Data set	**#Instances**	**#Attributes**	γ
Abalone	4177	8	2^{-4}
Airfoil Self-Noise	1503	9	2^{-5}
Concrete Compressive Strength	1030	9	2^{-5}
Friedman	1200	5	2^{-4}
Laser	993	4	2^{-6}
Mortgage	1049	15	2^{-4}
Stock Prices	950	9	2^{-4}

Table 1.: Summary of data sets. The column γ presents values of parameter γ in Equation 1. We examined the range of $\gamma \in \{2^{-10}, 2^{-9}, \ldots, 2^{1}\}$ and selected γ, for which the highest R^2 score was obtained in case of complete data set. This strategy was not adjusted for any particular method used in the experimental section.

We applied 5-fold cross-validation procedure, where a data set was divided into 5 equal subsets. In each run, r-SVM with kerenl (2) was train on 4 normalized subsets and evaluated on the remaining fold. The results were averaged. We used $C = 1$, $\varepsilon = 0.1$ and the value of kernel parameter γ shown in Table 1. We used Coefficient of Determination (R^2 score) as a performance measure:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}},$$

where $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ is a total sum of squares (y_i are the actual values and $\bar{y}$ is the mean of these values) and $SS_{\text{res}} = \sum_i (f_i - y_i)^2$ is residual sum of squares (f_i are the predicted values). Usually, the R^2 score ranges from 0 to 1 and the best possible score is 1. However, the R^2 can be negative if the chosen model fits worse than a horizontal line. In such cases, it means that the chosen model fits the data poorly.



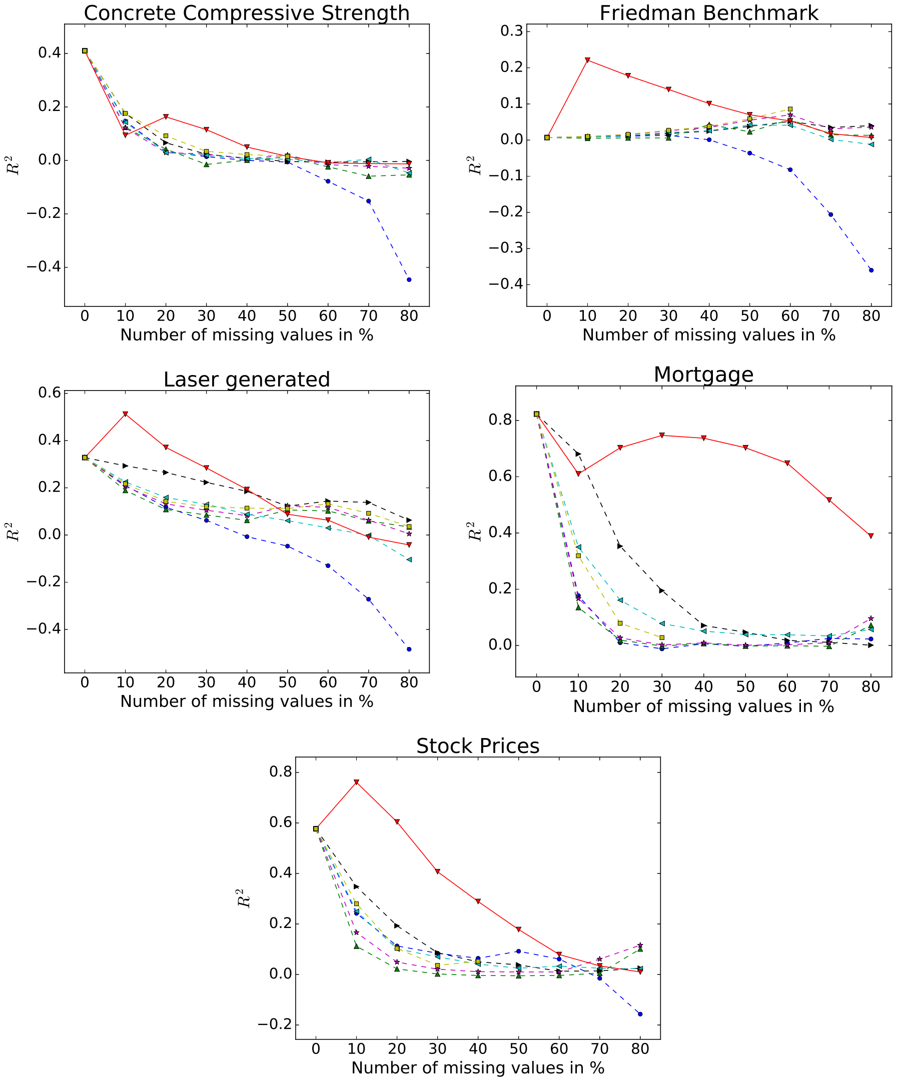


Figure 3.: R^2 score obtained when removing attributes using MCAR (the first strategy). The results for SVMimpute-MV are missing, because it requires at least one complete sample in data set.

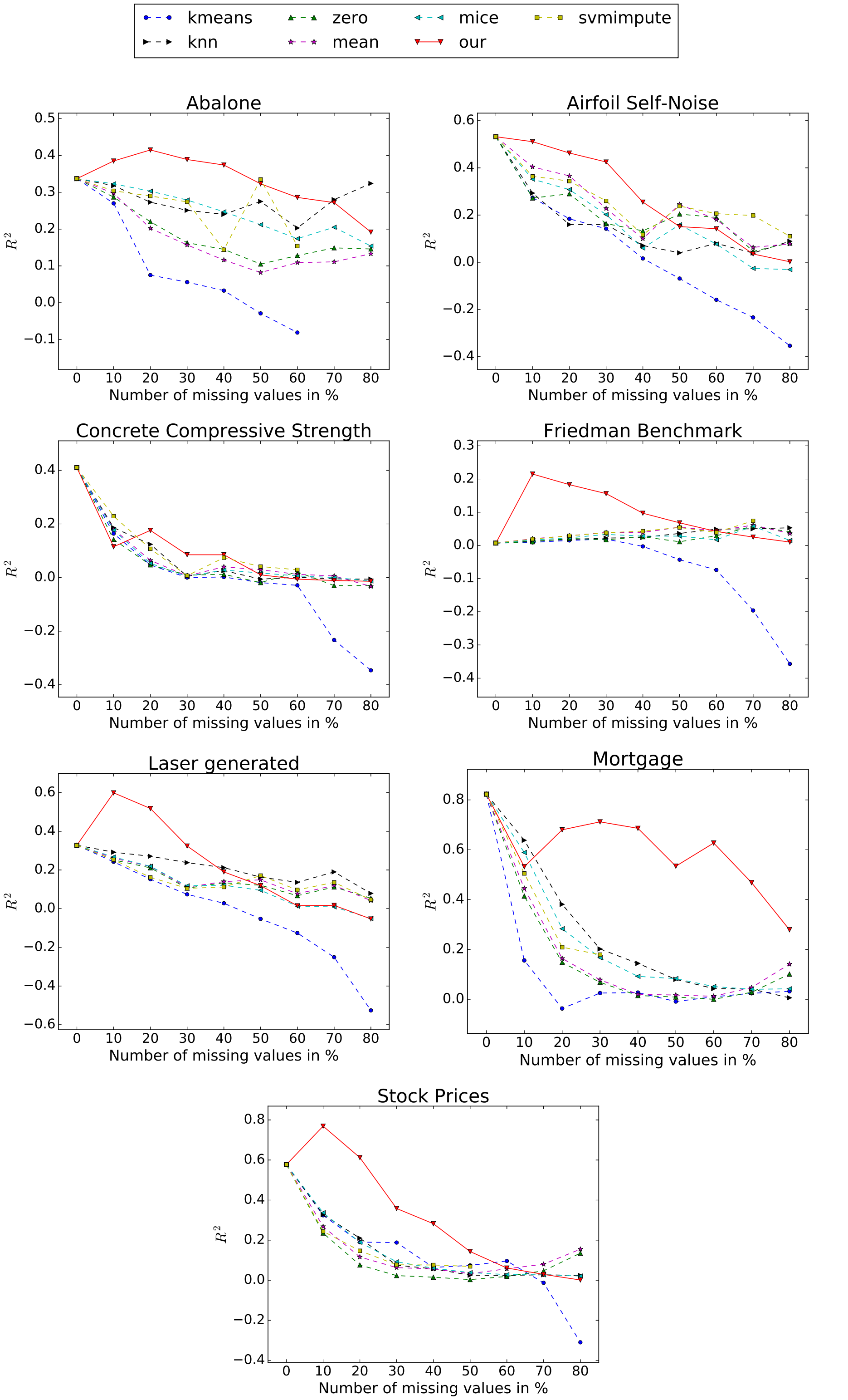


Figure 4.: R^2 score obtained when removing attributes using MAR (the second strategy).

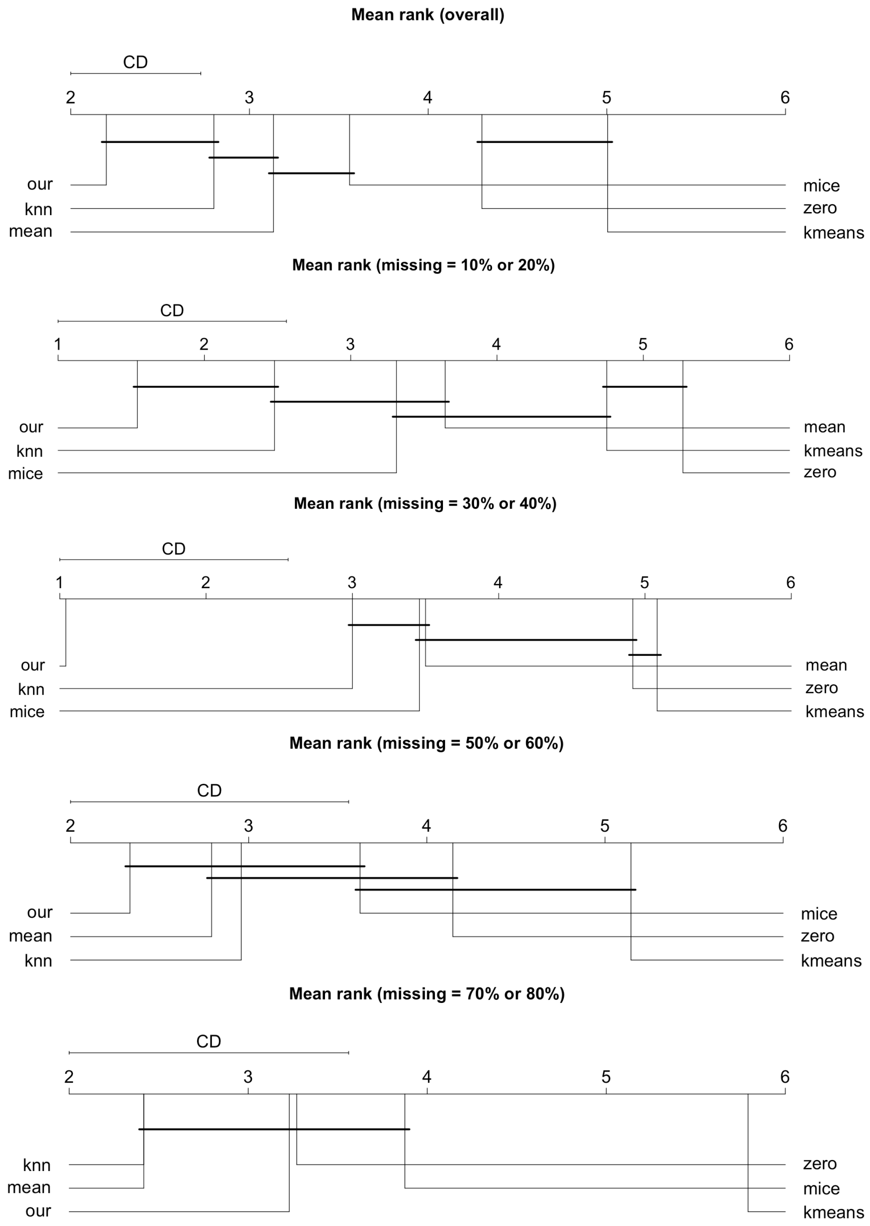
As expected, the prediction is usually more difficult, when more coordinates are missing. The exception are observed when applying our method to Friedman and Laser data sets, then R^2 score is worse for complete data than for data with 10%–50% of missing attributes. One can also observe that the performances of all methods in the case of MCAR and MAR removing strategies are similar. This behavior was expected in case of our method, because its performance strictly depends on density estimation's quality (since the use of EM in both strategies is theoretically justified, there should be no significant difference between their results).

Visual inspection confirm that our method gave the best prediction in case of most data sets. Its superiority is evident in case of Mortgage and Stock data, where only our method achieves the results comparable to the result of complete data set. For other data sets, the prediction task was more complicated and none of the methods exceed the level of $R^2 = 0.3$. One can observe that the advantage of our method was more evident for data with 10%-20% of missing attributes than for data with 70%-80% of missing data. It was also expected, as it is extremely difficult to give reliable estimation when there is so many missing values.

We analyzed the results of the classification task using a method proposed by Demšar [7], specifically using the Friedman test with Nemenyi post hoc analysis. It ranks the methods for each data set (and percentage of missing coordinates) separately, the best performing algorithm getting the rank of 1, the second best rank 2 etc. Each combination of data set and percentage of missing values is treated as a separate test, giving one rank measurement per method. The analysis then consists of two steps: (i) the null hypothesis is made that all methods perform the same and the observed differences are merely random (the hypothesis is tested by the Friedman test, which follows a χ^2 distribution); (ii) having rejected the null hypothesis the differences in ranks are analyzed by the Nemenyi test.

For a confidence level of $p = 0.05$ and given the 7 methods tested over 6 data sets with different percentage of missing coordinates, the "critical difference" was calculated as 0.998 (the difference in mean rank between a pair of methods must exceed 0.998 for the difference to be considered statistically significant). Figure 5 visualizes the results of this analysis using the CD (critical difference) diagram proposed by Demšar. The x-axis shows the mean rank over combinations of data set and percentage of missing values for each method. Methods are shown from left to right in increasing (first to last) rank order. Groups of methods for which the difference in mean rank is not significant are connected by horizontal bars.

As can be observed, the mean rank of our method is better than the others. However, the difference between our method and KNN-MV cannot be considered statistically significant in the case of all combinations of data set and percentage of missing values. Nevertheless, our method is significantly better than all the others when the percentage of missing values is at the level o 30% and 40%. This is visually confirmed by the results in Figure 3-4.



Mean rank (overall)

Mean rank (missing = 10% or 20%)

Mean rank (missing = 30% or 40%)

Mean rank (missing = 50% or 60%)

Mean rank (missing = 70% or 80%)

Figure 5.: Analysis of statistically significant differences in the regression results. The mean rank over all combinations of data set and percentage of missing values is plotted on the x-axis for each method. Methods which are not significantly different (for $p = 0.05$), in terms of mean rank, are connected. Some of the SVMimpute-MV results are missing, therefore we did not took it into consideration in this analysis.

4. Conclusion

We have presented a method that constructs an analogue of classical RBF kernel for incomplete data. The ability of working with such a data is of practical importance, because classical machine learning algorithms cannot be directly applied to data set with missing values.

When comparing our method with existing imputation techniques, the mean rank of our method is always better than the others. Moreover, the difference is significant, when the percentage of missing values is at the level o 30% and 40%.

This confirms that our approach capture relevant information that is not captured by traditional imputation methods, where the artificially generated values are treated in the same way as the known ones.

5. References

[1] Alan C Acock. What to do about missing values. 2012.

[2] Jesús Alcalá-Fdez, Luciano Sanchez, Salvador Garcia, Maria Jose del Jesus, Sebastian Ventura, Josep Maria Garrell, Jose Otero, Cristobal Romero, Jaume Bacardit, Victor M Rivas, et al. Keel: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, 2009.

[3] Arthur Asuncion and David J. Newman. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml/`, 2007.

[4] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[5] Gustavo EAPA Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

[6] Gal Chechik, Geremy Heitz, Gal Elidan, Pieter Abbeel, and Daphne Koller. Max-margin classification of data with absent features. *Journal of Machine Learning Research*, 9:1–21, 2008.

[7] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[8] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[9] Zoubin Ghahramani and Michael I Jordan. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*, pages 120–127. Citeseer, 1994.

[10] David Grangier and Iain Melvin. Feature set embedding for incomplete data. In *Advances in Neural Information Processing Systems*, pages 793–801, 2010.

[11] Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, and Chen Yumei. A svm regression based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 581–587. Springer, 2005.

[12] Dan Li, Jitender Deogun, William Spaulding, and Bill Shuart. Towards missing data imputation: a study of fuzzy k-means clustering method. In *International Conference on Rough Sets and Current Trends in Computing*, pages 573–579. Springer, 2004.

[13] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[14] Roderick J. A. Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[15] Julián Luengo, Salvador García, and Francisco Herrera. A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfns and eventcovering method. *Neural Networks*, 23(3):406–418, 2010.

[16] Patrick E McKnight, Katherine M McKnight, Souraya Sidani, and Aurelio Jose Figueredo. *Missing data: A gentle introduction*. Guilford Press, 2007.

[17] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC Press, 1997.

[18] Pannagadatta K Shivaswamy, Chiranjib Bhattacharyya, and Alexander J Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

[19] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[20] Alexander J Smola, SVN Vishwanathan, and Thomas Hofmann. Kernel methods for missing variables. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. Citeseer, 2005.

[21] Łukasz Struski, Marek Śmieja, and Jacek Tabor. Incomplete data representation for SVM classification. `https://arxiv.org/abs/1612.01480`, 2016.

[22] Andrew Wagner, John Wright, Arvind Ganesh, Zihan Zhou, Hossein Mobahi, and Yi Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012.

[23] David Williams and Lawrence Carin. Analytical kernel matrix completion with incomplete multi-view data. In *Proceedings of the ICML Workshop on Learning With Multiple Views*, 2005.

[24] David Williams, Xuejun Liao, Ya Xue, and Lawrence Carin. Incomplete-data classification using logistic regression. In *Proceedings of the International Conference on Machine Learning*, pages 972–979. ACM, 2005.