Missing Data Imputation within the Statistical learning Paradigm

Antonio D'Ambrosio

Abstract In the framework of missing data imputation, Rubin [7] formalized three types of missing data mechanisms upon definition of a missing data indicator matrix pointing out the missing-ness in the data matrix, assigning it a random variable with a conditional probability distribution given the data matrix depending on unknown parameters. Within Rubin's paradigm, missing data imputation can be understood as a model selection problem, such as to estimate the performance of different models in order to choose the best one which generate sample data. This paper formalizes a new missing data imputation paradigm [3]. Within *statistical learning paradigm* [8], missing data imputation can be understood as a model assessment problem, whatever is the probability model underlying sample data the goal is to minimize its prediction error (generalization error) on new data.

Key words: Missing data imputation, Data fusion, Statistical learning, Machine learning

1 Missing at random mechanism, Classical paradigm and Statistical learning paradigm

The proposed methodology for missing data imputation assumes *missing at random* mechanism [7, 5, 4]. Let $Z = [X|Y]$ be a multivariate random variable, in which X represents the complete submatrix a nd Y contains missing values. Let $\mathbf{R}$ be the missing data indicator matrix. Specifically, the elements $\mathbf{R}$ are considered as realizations of random variables characterized by a probability distribution function $f(R|Z,\psi)$ depending on the parameter vector ψ. Let $f(Z|\theta)$ be the probability distribution function of the random variables $Z = [X|Y]$ associated to the complete and

Antonio D'Ambrosio
Department of Mathematics and Statistics, University of Naples Federico II
e-mail: antdambr@unina.it

un-complete part of the data matrix (respectively, X and Y), depending on the parameter vector θ, such that $f(X|\theta) = \int f(Z|\theta)dY$. The joint distribution of R and Z can be expressed as follows:

$$f(R,Z|\theta,\psi) = f(Z|\theta)f(R|Z,\psi) \tag{1}$$

According to Rubin's definition, missing at random mechanism assumes that the distribution of **R** depends on the data **Z** only through the complete part **X**, thus it holds

$$f(R,Z|\psi) = f(R|X,\psi) \tag{2}$$

Thus, inference on parameters should be based on

$$f(R,X|\theta,\psi) \propto \int f(X,Y|\theta)f(R|X,Y,\psi)dY \tag{3}$$

If the (1) holds then from (2) and (3) it results

$$f(X|\theta,\psi) \propto f(R|X,\psi)f(X|\theta) \propto f(X|\theta) \tag{4}$$

As a result, under missing at random mechanism maximizing eq. (2) is equivalent to maximizing eq. (4), thus the inference on θ can be funded on the observed data ignoring the missing mechanism. Rubin's paradigm aims to provide an unbiased estimate of θ such to identify the probability distribution function which generates the sample data.
In the following, statistical learning theory [8, 9] is considered to provide an alternative and distinctive paradigm for missing data imputation. If the missing at random condition is satisfied then for any record i of the sample data it holds

$$f([X_i,Y_i]|\theta,\psi) \propto f(R|[X_i,Y_i],\psi)f([X_i,Y_i]|\theta) \propto f([X_i,Y_i]\ \theta) \tag{5}$$

Within statistical learning theory, the interest relies on the imputation of each missing data such that the imputed value is the nearest to the real one. Discrepancy measures and validation in terms of minimization of the generalization (prediction) error will be considered.
Two separate goals can be satisfied. Within Rubin's paradigm, missing data imputation can be understood as a model selection problem, such as to estimate the performance of different models in order to choose the best one which generate sample data. Within statistical learning paradigm, missing data imputation can be understood as a model assessment problem, whatever is the probability model underlying sample data the goal is to minimize its prediction error (generalization error) on new data.
Statistical Learning Theory describes a general model of supervised learning which is suitable in presence of large data dimensionality and unknown probability distribution which generate sample data. Three are the main components: first, a generator G of random vectors $x \in R^k$ drawn independently from a fixed but unknown probability distribution function $F(x)$; second, a supervisor S which returns an out-

put value y to every input vector x, according to a conditional distribution function $F(y|x)$, also fixed but unknown; third, a learning machine *LM* capable of implementing a set of functions $f(x,\theta)$ where $\theta \in \Theta$ is a set of parameters. The problem of learning is that of choosing from the given set of functions $f(x,\theta)$, $\theta \in \Theta$ which is the best to approximate the supervisor's response. This can be theoretically satisfied upon definition of a loss function $L(y, f(x,\theta))$, to measure the discrepancy between the output value y returned by S and the output $\tilde{y} = f(x,\theta)$ provided by *LM*, and its expected value, known as the functional risk $R(\theta) = \int L(y, f(x,\theta))\, dF(x,y)$, to be minimized over the set of parameters. Aim is to provide a generalization of the results derived from its experience using a training sample of n independent and identically distributed observations drawn according to $F(x,y) = F(x)F(y|x)$: $(x_1,y_1),\ldots,(x_n,y_n)$ and considering the Empirical Risk Minimization (ERM) of $R_{emp}(\theta) = n^{-1}\sum_{i=1}^{n} L(y_i, f(x_i,\theta))$ over the set of parameters. This provides a consistent estimate of θ, given the loss function and the unknown probability distribution, if it holds both

$$plim R_n(\theta) = inf R(\theta),\ plim R_{emp,n}(\theta) = inf R(\theta) \tag{6}$$

where $R_n(\theta)$ and $R_{emp,n}(\theta)$ are the functional risk and empirical risk functions given the data of the training sample. This allows to define the *key theorem of Statistical Learning Theory* [10]: let $L(y, f(x,\theta))$, $\theta \in \Theta$ be a set of functions that satisfy the condition $A \leq R(\theta) \leq B$ with A and B finite constant values, then for the ERM principle to be consistent, it is necessary and sufficient that the empirical risk $R_{emp}(\theta)$ converges uniformly, in probability, to the actual risk $R(\theta)$.
Vapnik and Chervonenkis have defined a measure of the learner ability defined as the cardinality of the largest set of points the algorithm can shatter. For each set of functions characterized by a given complexity it is possible to define the VC dimensionality d to be the largest value such that it exists a correspondence between the input and the output. It can be shown that the ERM method yields to a consistent estimate if the dimensionality d is finite and if, for a probability equal to $(1-\eta)$ with $\eta > 0$, it holds

$$R_n(\theta) \leq R_{emp,n}(\theta) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp,n}(\theta)}{B\varepsilon}}\right) \tag{7}$$

where B is the upper limit of $R(\theta)$ and $\varepsilon = 4\frac{d(ln\frac{2n}{d}+1) - ln(\frac{\eta}{4})}{n}$, with d the VC dimensionality of the class of functions to be used in the learning process. There is a trade-off between the training error and the complexity of the class of functions. If the ratio n/d is high then the ε is small, thus the empirical risk function $R_{emp,n}(\theta)$ tends to the functional risk function $R_n(\theta)$ for the training sample data. According to Vapnik's structural risk minimization (SRM) approach both the empirical risk and the second term of equation (7) need to be minimized. Thus, given the fit of a nested sequence of models of increasing VC dimensions $d_1 < d_2 < \ldots$, the model with the smallest value of the upper bound is finally chosen. Under this perspective, the imputation process consists in minimizing the structural risk such that for any

missing data the imputed value is the nearest to the real one. In other words, given $L_1 = L(y, f(x,\theta_1))$ and $L_2 = L(y, f(x,\theta_2))$, the imputation due to $f(x,\theta_1)$ is better than the one provided by $f(x,\theta_2)$ if $L_1 < L_2$.
Missing at random condition reflects indeed a *pattern* of missing data related to the observed data in the dataset. For that reason missing data imputation can be seen as a supervised learning process. If we agree that missing data imputation is a supervised learning process, then we *must* agree that the lower is the generalization error provided by the learning machine in imputing missing data, the better is the method used to solve the problem and the better is the solution achieved. If this is true, then for data mining purposes missing data imputation methods should be deterministic and no stochastic. It is worthwhile that the same paradigm is extended to data editing [6], and data fusion [1, 2].

References

1. CONVERSANO, C., and SICILIANO, R. (2009). Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering. *Journal of Classification*, 26(3), 361-379
2. D'AMBROSIO, A., ARIA, M., and SICILIANO, R. (2007). Robust Tree-based Incremental Imputation Method for Data Fusion. *LNCS 4273; Advances in Intelligent Data Analysis*, Springer-Verlag, pp 174-183.
3. D'AMBROSIO, A., ARIA, M., and SICILIANO, R. (2012). Accurate Tree-based Missing Data Imputation and Data Fusion within the Statistical Learning Paradigm. *Journal of Classification*, to appear.
4. LITTLE, J.R.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420), 1227-1237.
5. LITTLE, J.R.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.
6. PETRAKOS, G., CONVERSANO, C., FARMAKIS, G., MOLA, F., SICILIANO, R., and STAVROPOULOS, P. (2004) New ways to specify data edits. *Journal of Royal Statistical Society*, Series A, volume 167, part 2, 249-274.
7. RUBIN, D.B. (1976). Inference and Missing Data (with Discussion). *Biometrika* 63, pp.581-592.
8. VAPNIK, V.N. (1995). The Nature of Statistical Learning Theory, Springer Verlag.
9. VAPNIK, V.N. (1998). Statistical Learning Theory, Wiley.
10. VAPNIK, V.N., CHERVONENKIS, A. (1989). The necessary and sufficient conditions for consistency of the method of empirical risk minimization, *Pattern Recognition and Image Analysis*, pp. 284-305.