The Treatment of Missing Values and its Effect on Classifier Accuracy

Edgar Acuña and Caroline Rodriguez

University of Puerto Rico at Mayaguez, Puerto Rico
edgar@cs.uprm.edu, caroline@math.uprm.edu

Abstract: The presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset as a training sample. Several methods have been proposed to treat missing data and the one used most frequently deletes instances containing at least one missing value of a feature. In this paper we carry out experiments with twelve datasets to evaluate the effect on the misclassification error rate of four methods for dealing with missing values: the case deletion method, mean imputation, median imputation, and the KNN imputation procedure. The classifiers considered were the Linear Discriminant Analysis (LDA) and the KNN classifier. The first one is a parametric classifier whereas the second one is a nonparametric classifier.

1 Introduction

Missing data is a common problem in statistical analysis. Rates of less than 1% missing data are generally considered trivial, 1-5% are manageable. However, 5-15% requires sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation. Several methods have been proposed in the literature to treat missing data. Many of these methods were developed for dealing with missing data in sample surveys (cf. Kalton and Kasprzyk, 1986; Mundfrom and Whitcomb, 1998), and have some drawbacks when they are applied to classification tasks. Chan and Dunn (1972) considered the treatment of missing values in supervised classification using the LDA classifier but only for two classes problems considering a simulated dataset from a multivariate normal model. Dixon (1979) introduced the KNN imputation technique for dealing with missing values in supervised classification. Tresp et al. (1995) also considered the missing value problem in a supervised learning context for neural networks. The interest in dealing with missing values has continued with the statistical applications to new areas such as data mining (Grzymala-Busse and Hu, 2000) and microarrays (Hastie et al., 1999;

D. Banks et al. (eds.), *Classification, Clustering, and Data Mining Applications*

Troyanskaya et al., 2001). These applications include supervised classification as well as unsupervised classification (clustering). Bello (1995) compared several imputation techniques in regression analysis, a area related to classification.

Little and Rubin (2002) divide methods for treating missing data into three categories: Case/Pairwise Deletion, Parameter Estimation, and Imputation. The first is easiest and most commonly applied. The second uses maximum likelihood procedures and variants of the Expectation-Maximization algorithm to handle parameter estimation in the presence of missing data. These methods are generally superior to case deletion methods, because they utilize all the observed data and especially good when the probability mechanism leading to missingness can be included in the model. However, they suffer from several limitations, including: a strict assumption of a model distribution for the variables, such as a multivariate normal model, which has a high sensitivity to outliers, and a high degree of complexity (slow computation). The third method replaces missing values with estimated ones based on information available in the data set. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. There are many options varying from naive methods, like mean imputation, to some more robust methods based on relationships among attributes.

In this paper we compare four methods to treat missing values in supervised classification problems. We choose the case deletion technique (CD), the mean imputation (MI), the median imputation (MDI) and the k-nearest neighbor (KNN) imputation. The criteria to compare them are the effects on the misclassification rate in two classifiers: the Linear Discriminant Analysis (LDA) and the KNN classifier. The first is a parametric classifier and the second one is a nonparametric classifier. In Section 2 the four methods to treat missing values considered in this paper are described. In Section 3 we explain our experimental methodology and in Section 4 we present and discuss our results.

2 Four Methods for Missing Values

The following four methods are used in this paper to treat missing values in the supervised classification context. We also give a brief description of other methods not considered in this paper.

A. Case Deletion (CD). This method is also known as complete case analysis. It is available in all statistical packages and is the default method in many programs. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and deleting the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with a high degree of missing values. CD is less hazardous if it involves minimal loss of sample size (minimal missing data or a sufficiently large sample size) and there is no structure or pattern to the missing data. For other situations, where the sample size is insufficient or some structure exists in the missing data, CD has been shown to produce more biased estimates than alternative methods. CD should be applied only in cases in which data are missing completely at random (see Little and Rubin (2002)).

B. Mean Imputation (MI). This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. Let us consider that the value x_{ij} of the k-th class, C_k, is missing then it will be replaced by

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in C_k} \frac{x_{ij}}{n_k}, \tag{1}$$

where n_k represents the number of non-missing values in the j-th feature of the k-th class. In some studies the overall mean is used but we believe that this does not take in account the sample size of the class to which the instance with the missing values belongs. According to Little and Rubin (2002), among the drawbacks of mean imputation are (a) sample size is overestimated, (b) variance is underestimated, (c) correlation is negatively biased, and (d) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean. Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Surprisingly though, mean imputation has given good experimental results in data sets used for supervised classification purposes (Chan and Dunn, 1972; Mundfrom and Whitcomb, 1998).

C. Median Imputation (MDI). Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing values for a given feature are replaced by the median of all known values of that attribute in the class to which the instance with the missing feature belongs. This method is also a recommended choice when the distribution of the values of a given feature is skewed. Let us consider that the value x_{ij} of the k-th class, C_k, is missing. It will be replaced by

$$\hat{x}_{ij} = \text{median}_{\{i:x_{ij} \in C_k\}}\{x_{ij}\}. \tag{2}$$

In case of a missing value in a categorical feature we can use mode imputation instead of either mean or median imputation. These imputation methods are applied separately in each feature containing missing values. Notice that the correlation structure of the data is not being considered in the above methods. The existence of other features with similar information (high correlation), or similar predicting power, can make the missing data imputation useless, or even harmful.

D. KNN Imputation (KNNI). In this method the missing values of an instance are imputed by considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function. The algorithm is as follows:

---

(1) Divide the data set D into two parts. Let D_m be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called D_c.
(2) For each vector x in D_m:
  a) Divide the instance vector into observed and missing parts as $x = [x_o; x_m]$.
  b) Calculate the distance between the x_o and all the instance vectors from the set D_c. Use only those features in the instance vectors from the complete set D_c, which are observed in the vector x.
  c) Use the K closest instances vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k-nearest neighborhood. The median could be used instead of the median.

The advantages of KNN imputation are: (i) k-nearest neighbor can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors); (ii) it does not require one to create a predictive model for each attribute with missing data—actually, the k-nearest neighbor algorithm does not create explicit models; (iii) it can easily treat instances with multiple missing values; (iv) it takes in consideration the correlation structure of the data.

The disadvantages of KNN imputation are: (i) The choice of the distance function. It could be Euclidean, Manhattan, Mahalanobis, Pearson, etc. In this work we have considered the Euclidean distance. (ii) The KNN algorithm searches through all the dataset looking for the most similar instances. This is a very time consuming process and it can be very critical in data mining where large databases are analyzed. (iii) The choice of k, the number of neighbors. In similar fashion as it is done in Troyanskaya et al. (2001), we tried several numbers and decided to use k=10 based on the accuracy of the classifier after the imputation process. The choice of a small k produces a deterioration in the performance of the classifier after imputation due to overemphasis of a few dominant instances in the estimation process of the missing values. On the other hand, a neighborhood of large size would include instances that are significantly different from the instance containing missing values hurting their estimation process and therefore the classifier's performance declines. For small datasets k smaller than 10 can be used.

Other imputation methods are:

Hot deck Imputation. In this method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. In Random Hot deck, a missing value (the recipient) of a attribute is replaced by a observed value (the donor) of the attribute chosen randomly. There are also cold deck imputation methods that are similar to hot deck but in this case the data source to choose the imputed value must be different from the current data source. For more details see Kalton and Kasprzyk (1986).

Imputation using a prediction model. These methods consist of creating a predictive model to estimate values that will substitute the missing data. The attribute with missing data is used as the response attribute, and the remaining attributes are used as input for the predictive model. The disadvantages of this approach are (i) the model estimated values are usually more well-behaved than the true values would be; (ii) if there are no relationships among attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values; (iii) the computational cost is large since we have to build a many models to predict missing values.

Imputation using decision trees algorithms. All the decision tree classifiers handle missing values by using built-in approaches. For instance, CART replaces a missing value of a given attribute with the corresponding value of a surrogate attribute which has the highest correlation with the original attribute. C4.5 uses a probabilistic approach to handle missing data in both the training and the test samples.

Multiple imputation. In this method the missing values in a feature are filled in with values drawn randomly (with replacement) from a fitted distribution for that feature. Repeat this a number of times, say M=5 times. After that we can apply the classifier to each "complete" dataset and compute the misclassification error for each dataset by averaging the misclassification error rates to obtain a single estimate and also we can estimate the variance of the error rate. Details can be found in Little and Rubin (2002) and Shafer (1997).

3 Experimental Methodology

Our experiments were carried out using twelve datasets coming from the Machine Learning Database Repository at the University of California, Irvine. A summary of the characteristics of each dataset appears in Table 1. The number in parenthesis in the column *Features* indicates the number of relevant features for each dataset. The Missing Val. column contains the percentage of missing values with respect to the whole dataset and the Missing Inst. column contains the percentages of instances with at least one missing value. Considering these two values for *Hepatitis* we can conclude that its missing values are distributed in a large number of instances. The last two columns of Table 1 show the 10-fold cross-validation error rates for the the LDA and

KNN classifier, respectively. For the datasets with missing values these error rates correspond to the case deletion method to treat missing values.

Table 1. Information about the datasets used in this paper. Some features in the Ionosphere and Segment datasets were not considered in our experiment.

Dataset	n	Classes (no., size) (no., size)	Features	Missing Val.(%)	Missing Inst.(%)	LDA	KNN
Iris	150	3 (50,50,50)	4(3)	0	0	3.18	4.68
Hepatitis	155	2 (32,123)	19(10)	5.67	48.38	27.7	28.95
Sonar	208	2 (111,97)	60(37)	0	0	26.60	14.74
Heartc	303	2 (164,139)	13	0.15	1.98	16.51	19.42
Bupa	345	2 (145,200)	6(3)	0	0	35.04	36.46
Ionosphere	351	2 (225,126)	34(21)	0	0	16.59	13.23
Crx	690	2 (383,307)	15(9)	0.64	5.36	13.62	25.09
Breastw	699	2 (458,241)	9(5)	0.25	2.28	3.66	3.41
Diabetes	768	2 (500,268)	8(5)	0	0	24.59	27.37
Vehicle	846	4, all $\approx$ 200	18(10)	0	0	29.15	34.87
German	1000	2 (700,300)	20(13)	0	0	24.38	29.7
Segment	2310	7, all 330	19(11)	0	0	9.15	4.64

In the *Ionosphere* dataset we have discarded features 1 and 2 since feature 2 assumes the same value in both classes and feature 1 assumes only one value in one of the classes. For similar reasons in the *Segment* dataset we have not considered three features (3,4, and 5). Note that *Hepatitis* has a high percentage of instances containing missing values.

To evaluate more precisely the effect of missing values imputation on the accuracy of the classifier we worked only with the relevant variables in each dataset. This also sped up the imputation process. The relevant features were selected using the RELIEF, a filter method for feature selection in supervised classification, see Acuña et al. (2003) for more details. Batista et al. (2002) run a similar experiment but they choose only the three most important features and entered them one by one.

First we considered the four datasets having missing values. Each of them was passed through a cleaning process where features with more than 30% of missing values as well as instances with more than 50% of missing values were eliminated. We have written a program to perform this task that allows us to change these percentages as we want. This cleaning process is carried out in order to have minimize the number of imputations needed. After that is done we apply the four methods to treat missing values and once that is finished and we have a "complete" dataset we compute the 10-fold cross-validation estimates of the misclassification error for both the LDA and the KNN classifiers. The results are shown in Table 2.

Table 2. Cross-validation errors for the LDA and KNN classifiers using the four methods to deal with missing data.

Datasets	LDA				KNN			
	CD	MI	MDI	KNNI	CD	MI	MDI	KNNI
Hepatitis	27.7	31.50	32.07	30.83	28.95	38.32	37.67	39.23
*H*eartc	16.51	16.08	16.16	15.99	19.42	18.79	18.62	18.70
*C*rx	13.62	14.49	14.49	14.49	25.09	25.20	24.71	24.58
*B*reastw	3.66	3.72	3.66	3.96	3.41	3.84	3.88	3.61

Next, we considered the eight datasets without missing values and the "complete" versions of *Heartc*, *Breastw*, and *Crx*, obtained by case deletion. *Hepatitis* was not considered here because of its high percentage of instances containing missing values. In each of these 11 datasets we randomly insert a given percentage of missing values distributed proportionally according to the class size. We tried several percentages varying from 1% to 20%, but here, due to the lack of space, we only show the results for three of them. We recorded also also percentage of instances containing the missing values generated, but they are not shown in the table. After that we apply the four methods to treat missing values and compute 10-fold cross-validation estimates of the misclassification error rates for both the LDA and KNN classifiers. The results are shown in table 3.

4 Conclusions and Discussion

From Table 2 we can conclude that in datasets with a small number of instances containing missing values there is not much difference between case deletion and imputation methods for both types of classifiers. But this is not the case for datasets with a high percentage of instances with missing values, such as in *Hepatitis*.

From Tables 2 and 3 we can see that is not much difference between the results obtained with mean and median imputation. It is well known that most of datasets used here have features whose distributions contain outliers in both directions and their effects cancel out. Otherwise one could expect a better performance of the median imputation. From the same tables we can see that there is some difference between MI/MDI and KNN imputation only when a KNN classifier is used. However there is a noticeable difference between case deletion and all the imputation methods considered. Comparing the error rates from Tables 1 and 3 we can see that CD performs badly in *Sonar*, *Breast* and *German*, mostly due to the distribution of the missing values in a high percentage of instances. Overall KNN imputation seems to perform better than the other methods because it is most robust to bias when the percentage of missing values increases. In general, doing imputation does

Table 3. Cross-validation errors for LDA and KNN classifiers using the four missing data methods with missing rates shown in column MR.

Datasets	MR	LDA				KNN			
		CD	MI	MDI	KNNI	CD	MI	MDI	KNNI
Iris	1	2.89	3.82	3.72	3.64	4.82	4.81	4.70	4.86
	7	3.34	3.38	3.32	2.82	5.89	4.76	4.69	4.65
	13	3.82	2.97	3.16	3.04	4.28	2.28	2.65	3.44
Sonar	1	29.87	26.59	26.58	26.16	17.41	14.52	15.25	14.71
	3	31.63	25.48	25.91	26.14	24.06	11.50	12.24	12.99
	7	46.31	23.27	23.40	23.38	27.36	13.36	13.68	13.26
Heartc	5	12.96	14.84	14.11	15.66	16.44	18.25	18.32	18.53
	11	18.00	14.54	13.75	15.22	11.22	13.42	11.36	13.05
	21	11.75	12.94	10.64	13.64	17.12	12.41	10.23	12.59
Bupa	1	34.88	35.20	35.42	35.21	35.35	36.18	36.43	35.32
	3	36.50	36.23	36.66	35.70	35.98	37.22	37.02	36.37
	7	33.83	35.13	35.39	35.18	36.71	35.18	35.19	33.24
Ionosphere	1	15.57	16.04	16.16	16.17	14.63	12.52	12.44	12.87
	5	21.91	15.86	15.64	16.05	17.42	13.81	12.82	13.68
	9	27.87	15.28	15.13	15.83	18.97	13.81	12.82	13.68
Crx	3	15.05	13.18	13.16	13.35	24.60	24.93	25.39	25.65
	11	12.17	11.94	11.94	12.52	25.07	22.76	24.00	22.64
	21	16.44	10.82	10.71	10.71	34.70	18.97	18.24	23.97
Breastw	3	3.60	3.68	3.54	3.91	3.30	3.32	3.26	3.33
	11	4.68	3.46	3.59	3.78	3.34	2.82	2.82	2.86
	21	5.05	2.93	3.10	3.47	2.12	1.92	1.97	2.07
Diabetes	3	23.60	24.59	24.80	24.41	27.49	26.38	26.45	26.29
	9	24.09	24.09	24.24	24.42	25.35	25.82	24.56	26.05
	11	23.22	24.02	23.85	24.40	30.36	24.16	23.01	23.58
Vehicle	5	30.96	30.28	30.36	28.85	38.40	36.33	34.95	33.25
	13	30.91	34.49	34.81	28.81	40.30	33.78	32.83	32.41
	21	32.80	34.92	33.48	32.75	42.66	31.94	31.51	30.17
German	5	26.05	24.22	24.28	24.40	31.19	29.56	28.91	28.67
	13	26.00	23.88	22.70	23.96	35.92	28.03	28.93	27.61
	21	29.14	22.14	21.53	23.60	41.43	23.49	23.55	23.51
Segment	5	8.88	9.32	9.37	9.17	6.51	6.24	6.11	4.39
	13	8.29	9.35	9.44	8.84	9.41	7.04	6.60	5.31
	21	8.96	7.81	7.69	7.48	9.12	7.67	6.81	5.07

not seem to hurt too much the accuracy of the classifier, even if sometimes there is a high percentage of instances with missing values. This agrees with the conclusions obtained by Dixon (1979). We recommend that one can deal with datasets having up to 20% of missing values. For the CD method one can have up to 60% of the instances containing missing values and still have reasonable performance.

The R functions for all the procedures discussed in this paper are available in www.math.uprm.edu/˜edgar, and were tested in a DELL workstation with 3GB of memory RAM and a dual processor PENTIUM Xeon.

5 Acknowledgment

This work was supported by grant N00014-00-1-0360 from ONR and by grant EIA 99-77071 from NSF

References

1. Acuña, E., Coaquira, F. and Gonzalez, M. (2003). "A Comparison of Feature Selection Procedures for Classifiers Based on Kernel Density Estimation," in *Proceedings of the International Conference on Computer, Communication and Control Technologies*, Orlando, FL:CCCT'03, Vol I, pp. 468–472.
2. Batista G. E. A. P. A., and Monard, M. C. (2002). "K-Nearest Neighbour as Imputation Method: Experimental Results," Technical Report 186, ICMC-USP.
3. Bello, A. L. (1995). "Imputation Techniques in Regression Analysis: Looking Closely at Their Implementation," *Computational Statistics and Data Analysis*, **20**, 45–57.
4. Chan, P., and Dunn, O. J. (1972). "The Treatment of Missing Values in Discriminant Analysis," *Journal of the American Statistical Association*, **69**, 473-477.
5. Dixon J. K. (1979). "Pattern Recognition with Partly Missing Data," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-9, **10**, 617–621.
6. Grzymala-Busse, J. W., and Hu, M. (2000). "A Comparison of Several Approaches to Missing Attribute Values in Data Mining," in *Rough Sets and Current Trends in Computing 2000*, pp. 340–347.
7. Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M, Brown, P. and Bolstein, D. (1999). "Imputing Missing Data for Gene Expression Arrays," Techical Report, Division of Biostatistics, Stanford University.
8. Kalton, G., and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data," *Survey Methodology*, **12**, 1–16.
9. Little, R. J., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* second edn., John Wiley and Sons, New York.
10. Mundfrom, D. J., and Whitcomb, A. (1998). "Imputing Missing values: The Effect on the Accuracy of Classification," *Multiple Linear Regression Viewpoints*, **25**, 13–19.
11. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.
12. Tresp, V., Neuneier, R., and Ahmad, S. (1994). "Efficient Methods for Dealing with Missing Data in Supervised Learning," in *NIPS 1994*, eds. G. Tesauro, D. S. Touretzky, and T. K. Leen, Cambridge, MA:MIT Press, pp. 689–696.
13. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. Hastie, T., Tibshirani, R., Bostein, D. and Altman, R. B. (2001). "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, **17**, 520–525.