Effects of Missing Data Imputation on Classifier Accuracy

Rahul Samant,
SVKM'S NMIMS, Shirpur Campus, India;

Srikantha Rao,
TIMSCDR, Mumbai University, Kandivali,
Mumbai, India,

Abstract

This paper evaluates the impact of missing data imputation in a decision support system used in predicting the probability of occurrence of Hypertension & Diabetes. In this study we used four classifiers, viz. Naïve Bayesian, KNN and LDA (linear and quadratic) classifiers to classify patients records for diagnosis of hypertension and diabetes. Linear discriminate function and a logistic regression equation were developed using a set of thirteen symptom (input) variables. We replaced missing values in the dataset by artificially-generated values using different imputation techniques such as mean substitution, median value imputation and KNN imputation. The effect on the accuracy of the diagnosis predictions using the developed model with imputed values was determined. It is found that KNN imputations performed slightly better than other techniques.

1. Introduction

Most real-life knowledge-based applications encounter missing values in their database. Values can be missing for several reasons including incorrect data entry, erroneous or skipped measurements or equipment faults. Missing values cause problems such as loss of effectiveness, inability of the system to process data with missing values and biasing of the data compared to the original dataset. [1] Numerous methods have been adopted to treat missing data. Several of these methods were developed for dealing with missing data in sample surveys [2, 3] and have some disadvantages when they are applied to classification domain. Methods based on the k-nearest neighbor algorithm substitute the missing value with a value taken from 'k' cases that are most similar to the one with the missing value. To find more similar cases the weighted k-nearest neighbor method (wKNN) can be used [12]. Tresp et al [5] has considered the missing value problem in a supervised learning in context of neural networks. The interest in dealing with missing values has continued with the statistical applications to new areas such as Data Mining [6] and Microarrays [7, 8]. Imputation, i.e. the estimation of missing values by making an informed guess is very popular in knowledge based systems, especially in applications using clinical data [11].In general the methods for missing data has been divided into three categories [10], Case/Pair wise Deletion, Parameter estimation and Imputation techniques. The Case/Pair wise Deletion method is easiest and commonly used. In parameter estimation, maximum likelihood procedure is employed that use the variants of 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 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. While in imputation techniques, missing values are replaced 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 different 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 criterion to compare them is the effect on the percentage misclassification error of three classifiers: the Linear Discriminant Analysis (LDA), Naïve Bayesian (NB) classifier and the KNN classifier. The first two are parametric classifiers and the third one is a nonparametric classifier.

2. Methods for Missing value Treatment

The four methods used in this paper to treat missing valuesin the supervised classification context are described as follows.

2.1 Case Deletion (CD)

This method consists of discarding all instances (cases) with missing values for at least one feature.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. [10].

2.2 Mean Imputation (MI)

It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute.

$$mean \ = \frac{1}{n}\sum_{i=0}^{n} a_i$$

The drawbacks of mean imputation are (a) Sample size is over estimated, (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. [9] Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. However, mean imputation has given good experimental results in data sets used for supervised classification purposes. [4]

2.3 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 data for a given feature is replaced by the median of all known values of that attribute in the class where 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.

2.4 KNN Imputation (KNNI)

This method the missing values of an instance are imputed 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 Dm 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 Dc.

2. For each vector x in Dm:
a) Divide the instance vector into observed and missing parts as x = [xo; xm].
b) Calculate the distance between the xo and all the instance vectors from the set Dc. Use only those features in the instance vectors from the complete set Dc, 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.

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 creating 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., [8] we tried several numbers and decided to use k=7 based on the accuracy of the classifier after the imputation process.

3. Results and Discussion

The database used for analysis in this study has been compiled as a part of an earlier study entitled Early Detection Project (EDP) conducted at the Hemorheology Laboratory of the erstwhile Inter-Disciplinary Programme in Biomedical Engineering at the School (now Department) of Biosciences and Bioengineering, Indian Institute of Technology Bombay (IITB), Mumbai, India. Spanning over a period from January 1995 to April 2005, it compiled 1168 records, each with 13 parameters, which encapsulated the biochemical, hemorheological and clinical status of the individuals. Table 1 lists the summary of the characteristics of the dataset. Table 2. shows the 10-fold cross-validation error rates for the Naïve Bayesian, LDA (linear& quadratic) and KNN classifier, respectively.

Table 1. Summary of Dataset characteristics

Feature name	Missing Values (%)	Feature name	Missing Values (%)	Feature name	Missing Values (%)
AGE	3.68	SALB	20.20	RG	11.38
BSF	8.39	SP	17.03	BP1	18.06
BSP	27.48	CPV2	7.44	BP2	16.09
SC	6.50	CB2	6.50		
STG	19.34	HCT	1.88		

Table 2. Effect of imputation on classifier misclassification error

#	Classifier	% misclassification Error			
		Data set-1 Cleaned dataset	Data set-2 Mean imputed	Data set-3 KNN imputed	Data set-4 Median imputed
1.	NB	0.3	20	**19**	22
2.	KNN	3.9	35	**22**	35
3.	LDA-Linear	0.3	38	**36**	39
4.	LDA-Quad.	2.4	44	43	**40**

Table 1. Displays the extent of missing values present in the dataset for various features. Table 2 shows that, as expected, the cleaned dataset that all record dropped that had any feature data missing had smallest classification error. On the other hand missing data imputed by three different methods each shows higher classification error, as expected. The KNN method shows the smallest % classification error for the dataset used.

4. Conclusion

Numerous methods have been adopted to treat missing data. Several of these methods were developed for dealing with missing data in sample surveys and have some disadvantages when they are applied to classification problems. Methods based on the k-nearest neighbor algorithm substitute the missing value with a value taken from 'k' cases that are most similar to the one with the missing value. To find more similar cases the weighted k-nearest neighbor method (wKNN) can be used. The interest in dealing with missing values has continued with the statistical applications to new areas such as Data Mining and Microarrays .Imputation, i.e. the estimation of missing values by making an informed guess is very popular in knowledge based systems, especially in applications using clinical data. The KNN method of missing data imputing appears to perform the best when the measure is % classification accuracy for the four different classifiers exercised. The results using our data set consisting of real case data from the medical domain are promising and it would be very interesting to evaluate the method on a larger dataset and on data taken from different domains.

5. References

[1] A. Farhangfar, L. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, vol. 37, September 2007, pp. 692– 709.

[2] Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data Survey Methodology 12, 1-16.

[3] Mundfrom, D.J and Whitcomb, A. (1998). Imputing missing values: The effect on the accuracy of classification. Multiple Linear Regression Viewpoints. 25(1), 13-19.

[4] Chan, P. and Dunn, O.J. (1972). The treatment of missing values in Discriminant analysis. *Journal of the American Statistical Association*, 6, 473-477.

[5] Tresp, V., Neuneier, R. and Ahmad, S. (1995). Efficient methods for dealing with missing data in supervised learning. In G. Tesauro, D. S. Touretzky, and Leen T. K., editors, *Advances in NIPS 7.* MIT Press.

[6]Grzymala-Busse, J.W. and Hu, M. (2000). A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In *RSCTC'2000*, pages 340-347.

[7] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M, Brown, P. and Bolstein, D. (1999). Imputing missing data por gene expression arrays. *Techical Report Division of Biostatistics*, Stanford University.

[8] 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(6), 520-525.

[9] Little, R. J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* Second Edition. John Wiley and Sons, New York.

[10] O. Abdala and M. Saeed, "Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted k-nearest neighbors algorithm," Computers in Cardiology, vol. 31, Sept. 2004, pp. 693–696.

[11] J. Barnard and X. Meng, "Applications of multiple imputation in medical studies: From AIDS to NHANES," *Statistical Methods in Medical Research*, vol. 8, 1999, pp. 17–36.

[12] O. Abdala and M. Saeed, "Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted k-nearest neighbors algorithm," *Computers inCardiology*, vol. 31, Sept. 2004, pp. 693–696.