An Analysis and Comparison of Various Missing Data Imputation Tools and Techniques

Vijayakumar Kuppusamy [#1], Ilango Paramasivam [#2]
[#] School of Computer Science and Engineering, VIT UniversityVellore, India
[1] kvijayakumar@vit.ac.in
[2] pilango@vit.ac.in

Abstract—The missing data and noisy data are common in a data set and the finding the effect it causes on the accuracy is very important to be determined. In statistics, missing data, or such values, occur when no data value is assigned for a field in a dataset. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data given or taken from warehouses. Missing data reduce the representativeness of the sample and can therefore distort/deviate inferences & conclusions about the population. This study aims at calculating the effect of missing values on Naïve Bayes algorithm by using two data sets that are lymphoma and breast cancer. The values are skipped in certain order of both the data set and accuracy is computed and results were compared in a table. Naïve Bayes is based on probalistic model.

Keyword - Missing Data, Imputation,, Data Mining, Open Source Tools.

I. Introduction

Missing data is an issue in multivariate data because a case will be skipped from the analysis if it is missing data for a variable included in the analysis process [1], [2]. The following are the reasons for missing data.

1. Missing data can occur because of lack of response: no information is provided for several items/objects or no information is provided for a whole unit.
2. Dropout is a type of missingness that occurs mostly when studying development over a period of time. In this type of study the measurement is seen repeated after a certain period of time. Missingness occurs when participants leave before the test ends and one or more such measurements are missing.
3. Sometimes missing values are caused by the researcher itself.

II. Literature Review

Jonathan Sterne and colleagues did study describing about the use and guidance about multiple imputation approach to dealing with them in their paper [3], [4]. Brick, J. Michael, and Graham Kalton did a research on handling missing data on survey in Statistical method in 1996 and calculating heir effects [5]. Naïve Bayesian has used in medical diagnosis. Russell and Norvig was the first to study about Naïve Bayesian and they have mentioned in their first book. Rish, Irina in 2001 who worked on an empirical study of the Naïve Bayesian [6]. Altman, Douglas G., and J. Martin Bland did a article in British medical journal describing the effects of missing data [7].

A. Types of missing data

Understanding the reasons in depth why data are missing can help with analyzing the remaining data. If values are missing at random, the data sample may still be typical of the population. But if the values are missing systematically, analysis may be harder comparatively [3], [8].

MCAR : Values in a data set are missing completely at random if the events that lead to any particular data-item being missing are independent both of observed variables and of unobserved parameters of interest, and takes place entirely at random [9].

MAR : Missing at random occurs when the missing-ness is not random, but where missing-ness can be fully accounted for by the variables where there is complete information. MAR is an assumption that is indispensable to verify statistically, we must depend on its firm reasonableness [10].

MNAR : Missing not at random (also known as unavoidable nonresponse) is data that is neither MAR nor MCAR (i.e. the value of the variable that's missing is related to the reason why it's missing) [11].

III. METHODOLOGY

The naıve Bayesian classification gives the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. The naïve Bayesian classifier is the most correct in comparison with all other classifiers [6]. Bayesian classification is the statistical classifiers and for each new sample they provide that the sample belongs to a class. For ex: - sample john (age=27, income=high, student=no, credit_rating=fair).It uses the concept of joint conditional probability distributions, which makes class conditional independencies to be defined between subsets of variables and also it have a graphical model of relationships, by which learning can be performed. A belief network has directed acyclic graph and set of conditional probability tables' components, Each and every node in directed acyclic graph represents a random variable which can be discrete- or continuous-valued. They may be actual attributes given in the data or can be "hidden variables" understood to form a relationship. In this directed acyclic graph each and every arc represents probabilistic dependence. If an arc is drawn from a node A to a node B, then A is a parent or immediate predecessor of B, and B is a descendant of A. Every variable in the graph is independent of its non-descendants, gives its parents [12]. The reasons behind choosing Bayesian network among all the various classification techniques are:

1. The probabilistic nature of the Bayesian network is calculation the probabilities for hypothesis, among the most practical approaches to certain types of problems.
2. The incremental nature of Bayesian network i.e. each and every training example can be incrementally increase or decrease the probability that a hypothesis is correct.
3. Various types of past knowledge can be combined with observed data.

Let assumed that there is a sample called A, the probability of a hypothesis h, P(h|A) follows the Bayes theorem stated mathematically as the following equation

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

During the training of a belief network, various scenarios are possible. The network topology is constructed by human experts or inferred from the data and the network variables may be observable by some training tuples. The hidden data case is which is also called as missing values. Several algorithms exist for learning the missing values from the training data. The problem is one of discrete optimization. For solutions, Bayesian classification is the best choice. When the pattern of missing is known and the variables of training tuples is given are observable, then training the dataset is candid. It consists of computing the continuous probability table (CBT) entries, as is similarly done when computing the probabilities involved in naıve Bayesian classification [11].

In Bayesian network, Naïve Bayesian classifier is also used in which we assume that attributes are conditionally independent.

$$p(C_K \mid x_1......x_n) = \frac{1}{Z} P(C_K) \prod_{i=1}^{n} p(x_i \mid C_K)$$

IV. Most commonly used Tools for Missing Data Imputation

This Section describes the popular tools used in Missing Data Imputation. The various tools which includes proprietary and open source.

TABLE I – PSS Vs Weka

PRORIETARY	OPEN - SOURCE
PSS - Statistical Package for the Social Sciences (SPSS)	Weka
∑ α ÷	WEKA The University of Waikato
Developer : IBM Corporation Operating system: Windows, Linux on z Systems, Linux / UNIX and Mac Platform : Java Type : Statistical analysis, Data mining, Text analytics, Data collection, Collaboration & Deployment	Developer(s):University of Waikato Operating system:Windows, OS X, Linux Platform:IA-32, x86-64; Java SE Type:Machine learning
Overview : •Descriptive statistics: supports Cross tabulation, Frequencies, Descriptives, Explore, and Descriptive Ratio Statistics •Bivariate statistics: supports Means, t-test, ANOVA and Correlation (bivariate, partial, distances), Nonparametric tests. •Prediction for identifying groups: Factor analysis, cluster analysis (two-step, K-means, hierarchical)	Overview : • Available for free which uses the General Public License (GNU). • Portability, since it is fully implemented in the Java programming language and thus runs on any modern computing platform. • A comprehensive collection of data preprocessing and modelling techniques. • comfort of use due to its Graphical User Interfaces (GUI).
The SPSS analgorithm known as fully conditional specification (FCS) or chained equations imputation The basic idea is to impute one incomplete variables at a time, using the filled-in variable from one step as a predictor in all subsequent steps It uses linear regression for continuous variables and logistic regression for categorical variables [13], [14].	It supports various regular data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualisation, and feature selection [15].

TABLE II – SAS Vs PSPP

PRORIETARY	OPEN - SOURCE
sas	GNU PSPP http://www.gnu.org
Developer(s):SAS Institute Written in: C Operating system:Windows, IBM mainframe, Unix/Linux, OpenVMS Alpha Type:Numerical analysis	Developer(s):GNU Project Written in:C Operating system:GNU Type:Statistics
SAS (Statistical Analysis System) is a application package developed by SAS Institute for advanced analytics, multivariate analyses and predictive analytics[16].	This software provides a set of tools which includes frequencies, cross-tabs comparison of means (t-tests and one-way ANOVA); linear, logistic regression, reliability (Cronbach's alpha, not failure or Weibull), and re-ordering data, non-parametric tests [17].

TABLE II – Kxen Vs R

PRORIETARY	OPEN - SOURCE
kxen	R
Developer(s):KXEN Inc. Operating system Windows, Linux, Unix Type Predictive analytics	Paradigm Multi-paradigm: Array, object-oriented, imperative, functional, procedural, reflective Developer : R Core Team GNU General Public License
InfiniteInsight is a predictive modelling suite developed by KXEN that assists analytic professionals, and business executives to extract facts from data. Among other functions, InfiniteInsight is used for variable importance, classification, regression, and product recommendation, as described and expressed by the JDM API group. InfiniteInsight aimed to allow the prediction of a behaviour or a value, the forecast of a time series or the understanding of a group of individuals with similar behavior [18].	R is a Statistical tool which support programming language and also comprehensive software environment. The R language is broadly used among statisticians, data miners and data analysis [4]. Polls, surveys of data miners, and studies of scholarly literature databases R is easily extensible through functions, and the R community is noted for its active contributions in terms of packages. [19].

V. Fuzzy Unordered Rules Induction Algorithm

The original data set is first classified in two groups. The data having missing values in their attributes are kept in one group and those without any missing values are placed in the other group. The classifier is trained with the complete data sets, and later the missing data is given to the model for predicting the missing attribute values [20]. The process is looped for the entire set of attributes that have missing values. At the end of training, this training dataset and imputed missing value datasets are combined to make the complete data. The final dataset is then produced to the selected classifier for classification [21].

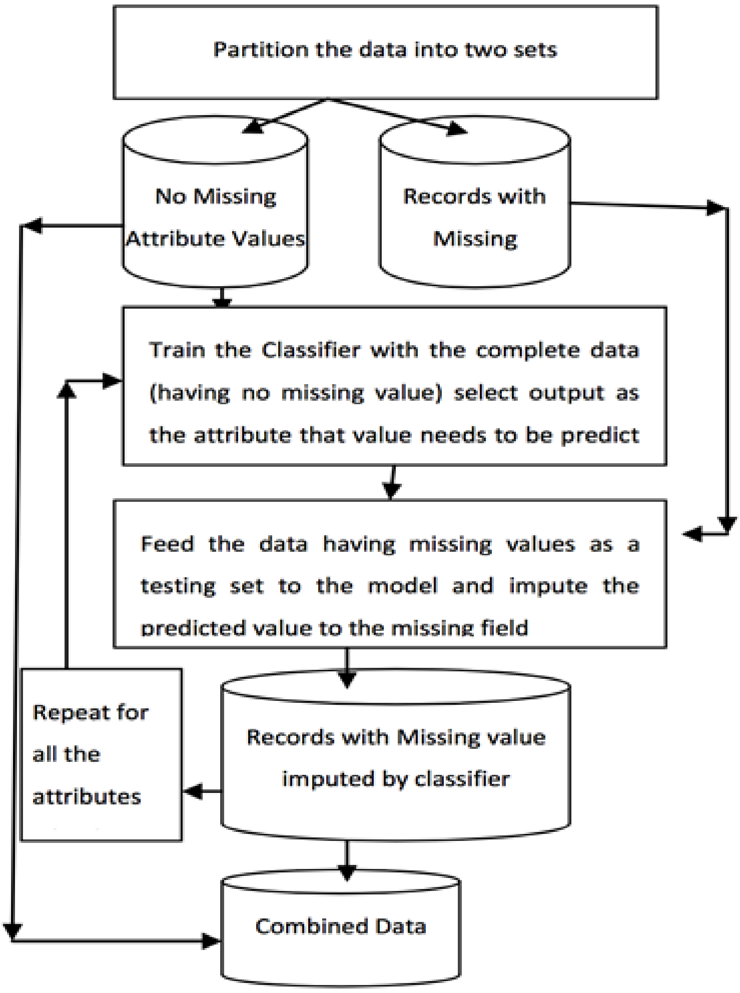
Partition the data into two sets
No Missing Attribute Values
Records with Missing
Train the Classifier with the complete data (having no missing value) select output as the attribute that value needs to be predict
Feed the data having missing values as a testing set to the model and impute the predicted value to the missing field
Repeat for all the attributes
Records with Missing value imputed by classifier
Combined Data

Fig. 1. Fuzzy unordered Rules Induction Algorithm

The data sets were taken from UCI-repository. The data set is of breast cancer and other is of lymphograpy. First data set has 699 rows and 9 attributes whereas other dataset has 142 rows and 18 attributes. The data set was converted to CSV (comma delimiter file) and imported to PHPMyAdmin in form of RDBMS tables. The IDE used to create is Net Beans 8.0 and JDK 6.0. The language used is JSP (java Servlet packages).

VI. RESULT ANALYSIS

The data set is varied as reducing 5%, 10%, 20%, 30% and 45%. For 5% rows are deleted and for greater than 20% some attributes are deleted. Both of them showed that the accuracy initially increases and start decreasing after a certain point. The accuracy is calculated by no. of values matching the correct label by the total number of values.

TABLE IV – Imputed Results

Missing Data (%)	Imputed Results in Data Set	
	Lymphoma	Breast Cancer
5	127/140	96
10	124/140	620/650
30	82/110	534/570
40	72/90	416/450

VII. CONCLUSION

Imputation missing value is one of the major tasks of data pre-processing when performing data mining. Simply removing the records which contain missing value from the original datasets can bring more problems than solutions. A suitable method for imputing the missing value can help to produce good quality datasets for better analysing trials. Mean/mode imputation, fuzzy unordered rule generation algorithm for imputation, decision tree imputation and other machine learning algorithms are used for imputing the missing value and the final datasets are classified using K-Mean clustering. The experiment shows that performance is improved when the fuzzy unordered rule induction algorithm is used to predict missing attribute values. According to the results and observations it can be seen that initially the accuracy increases up to a certain point and then it started decreasing gradually. As Naïve Bayes algorithm is based on probabilistic model and all the values are considered while testing hence reducing some values can have a positive impact but up to some limit.

REFERENCES

[1] B. Efron, “Missing data, imputation, and the bootstrap,” J. Am. Stat. Assoc., vol. 89, no. 426, pp. 463–475, 1994.
[2] T. J. Cleophas and A. H. Zwinderman, “Missing data imputation,” in Clinical Data Analysis on a Pocket Calculator, Springer, 2016, pp. 93–97.
[3] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls,” Bmj, vol. 338, p. b2393, 2009.
[4] M. L. Brown and J. F. Kros, “Data mining and the impact of missing data,” Ind. Manag. Data Syst., vol. 103, no. 8, pp. 611–621, 2003.
[5] J. M. Brick and G. Kalton, “Handling missing data in survey research,” Stat. Methods Med. Res., vol. 5, no. 3, pp. 215–238, 1996.
[6] I. Rish, “An empirical study of the naive Bayes classifier,” in IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, vol. 3, no. 22, pp. 41–46.
[7] D. G. Altman and J. M. Bland, “Missing data,” Bmj, vol. 334, no. 7590, p. 424, 2007.
[8] D. A. Bennett, “How can I deal with missing data in my study?,” Aust. N. Z. J. Public Health, vol. 25, no. 5, pp. 464–469, 2001.
[9] R. J. A. Little, “A test of missing completely at random for multivariate data with missing values,” J. Am. Stat. Assoc., vol. 83, no. 404, pp. 1198–1202, 1988.
[10] P. E. Cheng, “Nonparametric estimation of mean functionals with data missing at random,” J. Am. Stat. Assoc., vol. 89, no. 425, pp. 81–87, 1994.
[11] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, “Review: a gentle introduction to imputation of missing values,” J. Clin. Epidemiol., vol. 59, no. 10, pp. 1087–1091, 2006.
[12] D. Lowd and P. Domingos, “Naive Bayes models for probability estimation,” in Proceedings of the 22nd international conference on Machine learning, 2005, pp. 529–536.
[13] M. J. Norušis, IBM SPSS statistics 19 guide to data analysis. Prentice Hall Upper Saddle River, New Jersey, 2011.
[14] A. C. Acock, “Working with missing values,” J. Marriage Fam., vol. 67, no. 4, pp. 1012–1028, 2005.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, 2009.
[16] S. A. S. Institute, SAS user’s guide: statistics, vol. 2. Sas Inst, 1985.
[17] J. Yagnik and others, “Pspp A Free And Open Source Tool For Data Analysis,” 2014.
[18] F. F. Soulié, “Data Mining in the real world,” in Workshop on Data Mining for Business Applications, 2006.
[19] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data. John Wiley & Sons, 2014.
[20] M. M. Rahman and D. N. Davis, “Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-Mean clustering on real cardiovascular data,” in Proceedings of the World Congress on Engineering, 2012, vol. 1, no. 1, pp. 391–394.
[21] X. Dong, “Analysis of Impact of Missing Data in The Study if Racial Differences in Endometrial Cancer Survival,” University of Pittsburgh, 2009.

AUTHOR PROFILE

Vijayakumar K received his M.Tech Degree in 2006 from VIT University, Vellore - 632014, India. He is pursuing Ph.D in Data Mining & Warehousing.

Dr. Ilango Paramasivam received his Ph.D Degree in 2009 from National Institute of Technology, Tiruchirappalli - 620015, India. His research interest includes Data Mining & Warehousing, Web Mining, Machine Learning and Information Security.