ISSN 2047-3338

Effect of Missing Value Methods on Bayesian Network Classification of Hepatitis Data

Nazziwa Aisha, Mohd Bakri Adam and Shamarina Shohaimi

***Abstract*—Missing value imputation methods are widely used in solving missing value problems during statistical analysis. For classification tasks, these imputation methods can affect the accuracy of the Bayesian network classifiers. This paper study's the effect of missing value treatment on the prediction accuracy of four Bayesian network classifiers used to predict death in acute chronic Hepatitis patients. Missing data was imputed using nine methods which include, replacing with most common attribute, support vector machine imputation (SVMI), K-nearest neighbor (KNNI), Fuzzy K-means Clustering (FKMI), K-means Clustering Imputation (KMI), Weighted imputation with K-Nearest Neighbor (WKNNI), regularized expectation maximization (EM), singular value decomposition (SVDI), and local least squares imputation (LLSI). The classification accuracy of the naive Bayes (NB), tree augmented naive Bayes (TAN), boosted augmented naive Bayes (BAN) and general Bayes network classifiers (GBN) were recorded. The SVMI and LLSI methods improved the classification accuracy of the classifiers. The method of ignoring missing values was better than seven of the imputation methods. Among the classifiers, the TAN achieved the best average classification accuracy of 86.3% followed by BAN with 85.1%.**

***Index Terms*—Bayesian Network Classifiers, Missing Data, Imputation, Hepatitis Dataset, Classification and Data Mining**

I. Introduction

MANY real life data sets contain missing values. These missing values may result from equipment errors, data entry procedure, and incorrect measurements. The presence of missing values may make data analysis difficult, lead to loss of efficiency and cause bias resulting from difference between complete and missing data [1]. In classification tasks, the problem of missing data becomes more important. Missing data in the training or test set or in both affects the prediction accuracy of the classifier learned [2]. In order to carry out data mining, the data needs to be preprocessed; a stage that involves cleaning and preparing the data [3].

Until recently missing data analysis has been dominated by pair wise and list wise deletion. These methods will bias results if the remaining cases are not representative of the entire sample. There are more statistically principled methods of handling missing data which have been shown to perform better than ad hoc methods. These include maximum likelihood methods and imputation methods. An extensive body of technical literature exists on maximum likelihood methods [4], [5], [6]. For multiple imputations, a review of imputation methods can be seen in [7], [8], [9].

The interest in dealing with missing values has continued with the statistical applications to other areas such as data mining [10]. These applications include supervised classification as well as unsupervised classification (clustering). Lakshminarayan [11] states that the choice between unsupervised and supervised classification techniques should be influenced by the motivation for solving the missing data problem. The following literature lists some data mining methods that have been used as imputation methods. Chan [12] applies the linear discriminant analysis classifier on a simulated data set from a multivariate normal model to solve missing values in supervised classification. Hathaway [13] employs the fuzzy c-means algorithm and modifies it to act as useful tool for clustering real s-dimensional data that is incomplete. Brand [14] uses an incremental singular value decomposition to develop an efficient and unusually robust subspace-estimating flow-based tracker, and to handle missing points in structure- from-motion factorization. In data mining of micro array data, Zhang [15], proposes a sequential local least squares to impute missing values of micro array data and [16], estimates missing value for DNA micro array gene expression data by support vector regression imputation (this is a combination of logistic regression and support vector machines). On performing an extensive research investigating low and high amounts of missing data sets, Batista [17] found that k-nearest neighbor performed consistently better than embedded methods such as C4.5 and CN2.

These data mining methods that have been adopted as imputation methods have been compared to classical methods to evaluate their relevance in classification tasks. This

This work was supported in part by Ministry of Higher Education Malaysia who gave the scholarship.

A. Nazziwa. Author is with Mathematics Department of University Putra Malaysia, 43400 Serdang, Malaysia on leave from the Islamic University In Uganda, P.O.Box2555, Mbale, Phone: +60126815391; (Email: aishanazziwa@yahoo.ca).

M. B. Adam is with the Mathematics Department of University Putra Malaysia, 43400 Serdang, Malaysia, (Email: bbakri@science.upm.edu.my).

S. Shohaimi is with the Department of Biology, Universiti Putra Malaysia, 43400, Serdang Malaysia, (Email: sshohaimi@yahoo.com).

relevance is measured in terms of improvement in the prediction and classification accuracy. Data mining methods have been found to be competitive with classical methods [18]. A wider study evaluating the effect of fourteen different imputation approaches on classification finds that imputation methods could improve the accuracy of the classifiers [19].

Imputation is therefore an important data preparation task. The substitution of missing values should however not insert biases into the data set. Studies that asses the bias inserted by imputation methods usually use the prediction capability of imputation methods as a measure of bias [20]. These studies involve simulation of missing entries for some attributes whose values are known. The artificially missing values are then imputed and compared with the original values. Even though this evaluation is useful, it does not allow the influence of imputed values in the ultimate modeling task (e.g., in classification) to be inferred. Hruschka [21] notes that imputation cannot be properly evaluated apart from the modeling task. Alternative approaches are therefore needed to assess the bias. Since Bayesian network modeling involves specification of the structure and parameters of the model, we assess the bias inserted by the imputation methods using the classification accuracy.

In this study we classify Hepatitis patients into one of two outcomes (live or die). We develop four Bayesian network classifiers to predict whether patients with acute chronic Hepatitis will live or die. Since the data has many missing values, we treat the missing values in 10 ways to find which imputation method will improve prediction of outcome of acute chronic hepatitis. Our goal is to find the effect of imputed data on the classification accuracy of the Bayesian Network classifiers (BNC's). We compare the imputation methods using the classification accuracy obtained with the different imputation methods. The hepatitis data set from the UCI data repository, which has 48% natural missing values, is used. Most of the studies using this data set have employed the classification tree algorithm, which can be trained on data with missing values [22]

II. BAYESIAN NETWORK CLASSIFIERS

A classifier is a function that assigns a class label to an instance. In classification, the goal of a learning algorithm is to construct a classifier when given a training set with class labels. Let $U = (X,...,X_{n1},C)$ where $X_1,...,X_n,C$ are the attributes and C is the class variable. According to Bayes rule, the probability of an instance $u = (x_1,...,x_n)$ being class c is,

$$P(c/u) = \frac{P(u/c)P(c)}{p(u)} \qquad (1)$$

U is classified as the class $C = "+"$ if and only if

$$fbc(u) = \frac{P(C = "+" | u)}{P(c = "-" | u)} \geq 1$$

Where fbc is called the Bayesian classifier.

Naive Bayes and Augmented naive Bayes

Consider a graph structure where the class variable is the root as in the Fig.1 top, that is $Pa_c = \varnothing$, and each attribute has the class variables as its unique parent, namely $Pa_{x_i} = C$ $1 \leq i \leq n$. For this type of graph structure (1) above results in

$$P(X_1,...,X_n,C) = P(C)\prod_{i=1}^{n} P(X_i | C)$$

From the definition of conditional probability, we get

$$P(C | X_1,...,X_n) = \alpha P(C)\prod_{i=1}^{n} P(X_i | C).$$

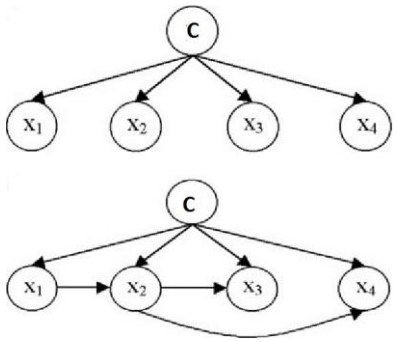

Fig. 1: The top figure shows the naive Bayes and the bottom figure shows the augmented naive Bayes. The class variable is "C" and X_1... X4 are the attributes

Where α is a normalizing constant. This is the definition of the naive Bayes (NB). The NB is the simplest form of useful Bayesian network classifier in which all attributes are independent given the class (conditional independence). While a NB classifier performs well on some tasks, often additional links and nodes provide a richer, more accurate model. The structure of the NB has been extended to meet these criteria. The Augmented naive Bayes (ANB) is an extension of NB that allows dependencies between variables. When the maximum number of parents a node in the Bayes net can have is set to two, we get a tree augmented Bayes network (TAN), and three parents is a Bayes net augmented Bayes network (BAN). When no restriction on the number of parents is enforced we get the general Bayes network classifier (GBN). The joint probability distribution of ANB is;

$$P(x_1,...,x_{n,}c) = P(c)\prod_{i=1}^{n} P(x_i | (Pa_{x_i}), c)$$

Learning Bayesian network classifiers requires a complete data set. We therefore apply imputation methods to our data set to find out which method will give us the best classification accuracy.

III. IMPUTATION METHODS USED IN THIS PAPER

The missing value methods used in this paper are:

1) Replacing with most common attribute value for nominal attributes, and average value for numerical attributes (MCOM) [26]. Only the instances with the same class as the reference instance are considered.
2) Imputation using the K-Nearest Neighbor (KNNI) [25], [26]. Every time a MV is found in a current instance, KNNI computes the k nearest neighbors and a value from them is imputed. For nominal values, the most common value among all neighbors is used, and for numerical values, the average value is used. The Euclidean distance is used to measure the proximity between instances.
3) K-means Clustering Imputation (KMI) [27]. The data set is divided into groups based on the similarity of objects and to minimize the intra-cluster dissimilarity. Data objects that belong to the same cluster are taken to be nearest neighbors of each other, and KMI applies a nearest neighbor algorithm to replace MV, in a similar way to KNNI.
4) Imputation with Fuzzy K-means Clustering (FKMI) [28], [13]. The degree to which a data object belongs to a certain cluster is described by a membership function. When updating membership functions and centroids, FKMI only takes into ac- count complete attributes. In this process, the data object cannot be assigned to a concrete cluster represented by a cluster centroid (as is done in the basic K-mean clustering algorithm), because each data object belongs to all K clusters with different membership degrees. FKMI replaces non-reference attributes for each incomplete data object based on the information about membership degrees and the values of cluster centroids.
5) Singular value decomposition Imputation (SVDI) [29]. This method simply learns a set of basic functions or Eigen-values from the complete data, and then imputes the missing values for an attribute by regressing its non-missing entries on the Eigen values, and use the regression function to predict the expression values at the missing locations.
6) Support Vector Machines Imputation (SVMI) [30]. In this approach, SVMI first selects the instances in which there are no missing attribute values. In the next step, the method sets one of the input attribute, some of those values that are missing, as the output attribute, and the output attributes as the input attributes by contraries. An SVMI regression is then used to predict the output attribute values.
7) Local Least Squares Imputation (LLSI) [31]. This method represents a target instance that has MVs as a linear combination of similar instances. Rather than using all available instances in the data, only similar instances based on a similarity measure are used. The method has the local connotation. There are two steps in the LLSI. The first step is to select k genes by the Pearson correlation coefficient. The second step is regression and estimation, regardless of how the k genes are selected. A heuristic k parameter selection method is used by the authors.
8) Regularized Expectation-Maximization (EM) [32]. The EM imputation is based on the Expectation-Maximization algorithm proposed by [6]. It uses the iterative procedure of the EM algorithm to calculate the sufficient statistics and estimate the parameters. The missing values will be produced in the process.
9) Weighted Imputation with K-Nearest Neighbor (WKNNI) [33]. The method selects the instances with similar values (in terms of distance) to a considered one, so it can impute as KNNI does. However, the estimated value now takes into account the different distances from the neighbors, using a weighted mean or the most repeated value according to the distance.

These nine methods are compared with instance deletion or Ignore Missing (IGN). Using this method, all instances with at least one MV are discarded from the data set.

IV. MATERIALS AND METHODS

The original hepatitis data is obtained from the UCI data repository [34]. It consists of 155 instances with 19 attributes. The attributes are categorical integer and real valued. The class label is grouped into two: Die with 32 instances, live with 123 instances. The data consists of 48.39 % missing instances. We have no information about the mechanism that generated the missing values and we take the assumption that they are distributed in a random way. This data set has been used widely for classification task [22], [23], [24]. Missing data in the hepatitis data set is treated using ten selected methods. The NB, TAN, BAN and GBN are then trained on the imputed data sets to predict death in patients with acute chronic hepatitis. For all the classification algorithms, we use the 10-fold cross validation on the same partitions to perform fair comparison scheme. Since these classifiers do not work with numerical attributes, we discretised the numerical attributes using the well known discretising method proposed by [35]. Missing value imputation is carried out using KEEL software [36] and the actual modeling task is carried out in WEKA [37]. Table 1, shows the parameters used by the imputation methods. The imputation methods are evaluated based on the classification accuracy, and the structure of the Bayesian network.

V. RESULTS AND DISCUSSION

Table 2 shows the classification accuracy achieved by the BNC's. The accuracy of the classifiers is different with different imputation methods. The bold values in the table represent the best classification accuracy across an imputation method while the underlined values show which imputation method is best for that classifier. The NB performed best with SVMI imputation; accuracy of 89% meaning the NB predicted the right outcome i.e. die or leave in 138 patients and the

wrong one in 17 cases, followed by KNN imputation with accuracy of 88.4 %. The BAN also performed best with SVMI imputation with accuracy of 89.7%. Although GBN performed poorest among the classifiers, when missing values are ignored and with SVMI imputation it performed the best. On ranking the imputation methods, TAN is least sensitive and benefits the most with imputation methods. It performed the best on seven out of ten imputation methods. SVMI is the best imputation method, with all the classifiers having the best accuracy with the SVMI imputation. The SVMI has also performed well in other studies [38]. On average TAN achieved classification accuracy of 86.3% followed by BAN with 85.1%.

The bold faced figures are the best accuracy across an imputation method and the underlined figures represent the best performance for a classifier.

The objective of this study was to find the effect of missing value treatment on Bayesian network classification of acute chronic Hepatitis. We use as our benchmark the accuracy of

Table 1: Parameters used by the imputation methods

Method	Parameters
SVMI	Kernel=RBF C=1.0 Epsilon=0.001 Shrinking=No
EM	Stagnation tolerance =0.0001 Inflation factor=1 Regression type=multiple ridge regression
SVDI	stagnation tolerance=0.005 Inflation factor=1 Regression type=multiple ridge regression Singular vectors=10
LLSI	Max number of nearest neighbor=200
KNNI,WKNNI	K=10
KMI	K=10 Iterations=100 Error=100
FKMI	K=3 Iterations=100 Error=100 m=1.5

Table 2: Accuracy of the Bayesian Classifiers

Classifier	NB	TAN	BAN	GBN
EM	84.5	85.8	**85.8**	80.6
FKM	82.5	**85.2**	82.6	81.9
IGN	87.5	87.5	86.3	**90.0**
KMI	81.9	83.9	**84.5**	81.2
KNNI	81.3	**85.2**	83.2	80.6
LLSI	88.4	**90.3**	89.0	87.7
MCOM	83.2	**83.9**	80.6	81.2
SVDI	85.8	**86.5**	85.8	84.5
SVMI	89.0	90.3	89.7	**90.9**
WKNNI	82.6	**83.9**	83.9	80.6
Average	84.6	86.3	85.1	83.9

BNC's achieved by ignoring of missing values i.e. case deletion. Results show that SVM and LLSI improved the accuracy of BNC's i.e. were better than case deletion. SVM was found to be a good imputation method in other studies [30]. Apart from SVM and LLSI, all the other methods used in this study did not improve the accuracy of BNC's. Although studies have found that imputation methods performed better than ignoring missing data [39] and that they improved classifier performance [40] [19], in this study, ignoring missing data was better than MCOM, KMI, SVDI, FKM, WKNNI, and EM. These results agree with Grzymala [41] who made a comparison of nine imputation methods on ten input data files and found that ignoring missing values was better than eight other methods.

VI. CONCLUSION

Imputation of missing values using support vector machine and local least squares improves classification accuracy of the Bayesian network classifiers. Instance deletion/ ignoring missing values is better than some methods of treating missing values for this data set. Among the Bayesian Network classifiers, the Tree Augmented Naive Bayes is the best classifier for the hepatitis data.

REFERENCES

[1] J. Wang, Data mining: opportunities and challenges. Irm Press, 2003.

[2] G. Corani and M. Zaffalon, "Learning Reliable Classifiers from Small or Incomplete Data Sets: The Naive Credal Classifier 2," Journal of Machine Learning Research, vol. 9, pp. 581–621, 2008.

[3] J. Han and M. Kamber, Data mining: concepts and techniques. Morgan Kaufmann, 2006.

[4] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data, vol. 4. Wiley New York, 1987.

[5] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art.," Psychological methods, vol. 7, no. 2, p. 147, 2002.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp. 1–38, 1977.

[7] P. D. Allison, "Multiple imputation for missing data," Sociological Methods & Re- search, vol. 28, no. 3, pp. 301–309, 2000.

[8] A. R. T. Donders, G. J. M. G. V. D. Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," Journal of clinical epidemiology, vol. 59, no. 10, pp. 1087–1091, 2006.

[9] 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: British Medical Journal, vol. 338, 2009.

[10] J. Grzymala Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in Rough sets and current trends in computing, pp. 378–385, Springer, 2001.

[11] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," Applied Intelligence, vol. 11, no. 3, pp. 259—-275, 1999.

[12] S. Chan and O. J. Dunn, "The treatment of missing values in discriminant analysis. The sampling experiment," Journal of the American Statistical Association, vol. 67, no. 338, pp. 473–477, 1972.

[13] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 31, no. 5, pp. 735–744, 2001.

[14] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," Computer Vision ECCV 2002, pp. 707–720, 2002.

[15] X. Zhang, X. Song, H. Wang, and H. Zhang, "Sequential local least squares imputation estimating missing value of microarray data," Computers in Biology and Medicine, vol. 38, no. 10, pp. 1112–1120, 2008.

[16] X. Wang, A. Li, Z. Jiang, and H. Feng, "Missing value estimation for DNA micro array gene expression data by Support Vector Regression imputation and orthogonal coding scheme," BMC bioinformatics, vol. 7, no. 1, p. 32, 2006.

[17] G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," Applied Artificial Intelligence, vol. 17, no. 5-6, pp. 519–533, 2003.-

[18] E. R. Hruschka and N. F. F. Ebecken, "Bayesian networks for imputation in classification problems," Journal of Intelligent Information Systems, vol. 29, no. 3, pp. 231–252, 2007.

[19] J. Luengo, S. Garc´ıa, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," Knowledge and Information Systems, pp. 1–32, 2012.

[20] S. Youting, B. N. Ulisses, and D. Edward, "Impact of Missing Value Imputation on Classification for DNA Microarray Gene Expression Data A Model-Based Study," EURASIP Journal on Bioinformatics and Systems Biology, 2009.

[21] E. R. Hruschka, A. J. T. Garcia, and R. Estevam, "On the influence of imputation in classification: practical issues," Journal of Experimental & Theoretical Artificial Intelligence, vol. 21:1, no. June 2012, pp. 43—-58, 2009.

[22] W. Duch and K. Grudzinski, "Ensembles of similarity-based models," in Proceedings of the International Symposium on" Intelligent Information Systems X, pp. 75–85, 2001.

[23] P. Kontkanen, P. Myllym¨aki, T. Silander, H. Tirri, and P. Grunwald, "On predictive distributions and Bayesian networks," Statistics and Computing, vol. 10, no. 1, pp. 39–54, 2000.

[24] W. Duch and R. Adamczak, "Statistical methods for construction of neural net- works," in International Congress on Neural Information Processing, pp. 629–642, 1998.

[25] A. T. Hudak, N. L. Crookston, J. S. Evans, D. E. Hall, and M. J. Falkowski, "Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data," Remote Sensing of Environment, vol. 112, no. 5, pp. 2232–2245, 2008.

[26] P. Jonsson and C. Wohlin, "An evaluation of k-nearest neighbor imputation using likert data," in Software Metrics, 2004. Proceedings. 10th International Symposium on, pp. 108–118, IEEE, 2004.

[27] Z. Liao, X. Lu, T. Yang, and H. Wang, "Missing Data Imputation: A Fuzzy K- means Clustering Algorithm over Sliding Window," in Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, vol. 3, pp. 133–137, IEEE, 2009.

[28] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation. A study of fuzzy k-means clustering method," in Rough Sets and Current Trends in Computing, pp. 573–579, Springer, 2004.

[29] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," 1999.

[30] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru, and C. Yumei, "A SVM regression based approach to filling in missing values," in Knowledge-Based Intelligent Information and Engineering Systems, p. 179, Springer, 2005.

[31] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," Bioinformatics, vol. 21, no. 2, pp. 187–198, 2005.

[32] T. Schneider, "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values," Journal of Climate, vol. 14, no. 5, pp. 853-871, 2001.

[33] S. Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," Expert Systems with Applications, vol. 28, no. 4, pp. 667–671, 2005.

[34] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases. University of California, Irvine, Dept. Of Information and Computer Sciences, 1998," Datasets is available at http://www. ics. uci. edu/˜ mlearn/ML Repository. Html, 2005.

[35] K. B. Irani, "Multi-Interval Discretization of Continuous Valued Attributes for Classification Learning," In: Proceedings of 13th international joint conference on uncertainly in artificial intelligence (IJCAI93), pp. 1022–1029, 1993.

[36] J. Alcal´a-Fdez, L. S´anchez, S. Garc´ıa, M. J. Del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, and Others, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," Soft Computing-A Fusion of Foundations, Methodologies and Applications, vol. 13, no. 3, pp. 307–318, 2009.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.

[38] H. Wang and S. Wang, "Data mining with incomplete data," Encyclopedia of Data Warehousing and Mining, pp. 293–296, 2005.

[39] J. M. Jerez, I. Molina, P. J. Garcıa-Laencina, E. Alba, N. Ribelles, M. Martın, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem.," Artificial intelligence in medicine, vol. 50, pp. 105–15, Oct. 2010.

[40] 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," Lecture Notes in Engineering and Computer Science, vol. 2197, no. 1, 2012.

[41] J. W. Grzymala-busse, M. Hu, and N. York, "A Comparison of Several Approaches to Missing Attribute Values in Data Mining," in Rough sets and current trends in computing, pp. 378–385, 2001.