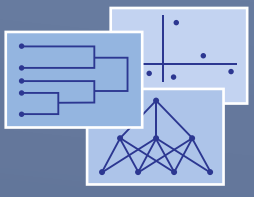
Studies in Classification, Data Analysis, and Knowledge Organization

Francesco Palumbo
Angela Montanari
Maurizio Vichi *Editors*

Data Science

Innovative Developments in Data Analysis and Clustering



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

More information about this series at http://www.springer.com/series/1564

Francesco Palumbo • Angela Montanari •
Maurizio Vichi
Editors

Data Science

Innovative Developments in Data Analysis and Clustering

Springer

Editors
Francesco Palumbo
Department of Political Sciences
University of Naples Federico II
Napoli, Italy

Angela Montanari
Department of Statistical Sciences Paolo Fortunati
Alma Mater Studiorum, University of Bologna
Bologna, Italy

Maurizio Vichi
Department of Statistical Sciences
Sapienza University of Rome
Rome, Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-319-55722-9 ISBN 978-3-319-55723-6 (eBook)
DOI 10.1007/978-3-319-55723-6

Library of Congress Control Number: 2017942955

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

On 4 July 1985 in Cambridge (UK), six classification societies gave birth to the International Federation of Classification Societies; the *Società Italiana di Statistica* (SIS) played an active propulsive role in the IFCS constitution. During the 30 years of the IFCS life, many other classification societies from all around the world have joined the IFCS. Through the active participation of its members, the SIS has actively and enthusiastically contributed to the IFCS growth. In 1997 the first conference of the Classification and Data Analysis Group of the SIS was hosted by the University of Chieti-Pescara. Following this long story of presence, in the occasion of IFCS's 30th birthday, under the IFCS presidency of Maurizio Vichi, the Classification and Data Analysis Group of the SIS and the Department of Statistical Sciences P. Fortunati of the *Alma Mater Studiorum* University of Bologna were proudly willing to organize the IFCS conference in Bologna. The conference organizer was Angela Montanari (University of Bologna) and Francesco Palumbo (University of Naples Federico II) served as chair of the scientific program committee. The conference was held between 5 and 8 July 2015.

Scholars from many different countries attended the conference. The commitment of the local organizing committee and the earnestness of the Scientific Program Committee ensured a successful and worthwhile conference. We are grateful to the members of the Scientific Program Committee: A. Cerioli (ClaDAG, Italy), D. Choi (KCS, Korea), C. Cuevas Covarrubias (SOLCAD, Mexico), N. Dean (BCS, UK), A. Ferligoj (SSS, Slovenia), P. Giudici (ClaDAG, Italy), C. Hennig (BSC, UK), T. Imaizumi (JCS, Japan), B. Lausen (GfKl, UK), P. McNicholas (CS, Canada), M. Nadif (SFC, France), A. Okada (JCS, Japan), I. Papadimitriou (GSDA, Greece), J. Pociecha (SKAD, Poland), A. Sbihi (MCA, Marocco), B. Scotney (IPRCS, Ireland), F. Sousa (CLAD, Portugal), D. Steinley (CS, USA), I. Van Mechelen (VOC, Belgium), and J. Vermunt (VOC, the Netherlands). IFCS has long-standing tradition of cooperation and exchange with other scientific statistical societies; in occasion of the IFCS conference in Bologna, V. Esposito Vinzi (France) and P. Groenen (the Netherlands) were invited to join the Scientific Program Committee as delegates of the ISBIS and IASC societies, respectively.

More than 200 contributions were organized into specialized sessions, contributed paper sessions, and one poster session. Moreover, five keynote lectures were given by eminent colleagues on different topics of data analysis and classification. The opening plenary session was devoted to the IFCS birthday celebration and a special session celebrated the 25th anniversary of the publication of the book on Generalized Additive Models by Hastie and Tibshirani.

Thanks to the collaboration with the publisher Springer and to its interest and attention to the IFCS activities during these 30 years, according to the longly consolidated tradition, the present post proceeding volume has been edited after the conference.

The scientific community unanimously considers *data science* as one of the most promising fields where to direct scientific research in the next years. However, already in occasion of the fifth IFCS conference, which was held in the year 1996 in Kobe (Japan), the related proceedings volume was entitled *Data Science, Classification, and Related Methods* (Hayashi et al. eds; Springer Japan, publisher). To emblematize the line of continuity along the IFCS, in occasion of the 30th birthday conference, we have decided to entitle the volume *Data Science: Innovative Developments in Data Analysis and Clustering*. The volume is a collection of full papers submitted after the conference. Papers were selected after a peer-review process, according to the high-quality standards of the series.

The volume is made of 27 contributions organized in three parts including contributions on:

- Classification methods for high-dimensional data
- Clustering methods and applications
- Multivariate methods and applications

Bologna, Italy — Angela Montanari
Napoli, Italy — Francesco Palumbo
Roma, Italy — Maurizio Vichi
November 2016

Acknowledgments

We are indebted to many people who allowed the success of the IFCS 2015 conference with their commitment. This book represents the final outcome of all the work done for the organization of the conference, during the days of the conference and after the end of it. First, we are grateful to the Department of Statistical Sciences of the University of Bologna who hosted the conference. In particular our thanks are addressed to the members of the organizing committee: L. Anderlucci, S. Bianconcini, S. Cagnone, L. De Angelis, G. Galimberti, A. Lubisco, M. Lupparelli, P. Monari, L. Stracqualursi, and C. Viroli. Two more additional thanks are for Laura Anderlucci, who has taken care of the conference web site, and for Paola Monari who, discreetly but effectively, has made all her experience and shrewdness in the organization available for the success of the conference.

We are also indebted to our colleagues that have collaborated in the review process of this volume:

Andrea Cerioli,
Pasquale Dolce,
Patrick Groenen,
Tadashi Imaizumi,
Antonello Maruotti,
Fionn Murtagh,
Akinori Okada,
Giancarlo Ragozini,
José Fernando Vera,
Vincenzo Esposito Vinzi,
Claudio Conversano,
Leonardo Grilli,
Christian Hennig,
Berthold Lausen,
Paul McNicholas,
Mohamed Nadif,
Domenico Piccolo,
Iven Van Mechelen,
Rosanna Verde,
Domenico Vistocco.

Last but not least, we are also indebted with SAS Institute, Springer, APT Servizi Regione Emilia Romagna and Ascom Bologna for their financial support to the conference.

Contents

Contributors

Patrizia Agati Department of Statistical Sciences, University of Bologna, Bologna, Italy

Olivier Alata Hubert Courien Laboratory, UMR 5516, Jean Monnet University, Saint-Étienne, France

Slavka Bodjanova Texas A&M University-Kingsville, Kingsville, TX, USA

Eva Boj Facultat d'Economia i Empresa, Universitat de Barcelona, Barcelona, Spain

Anne-Laure Boulesteix Department of Medical Informatics, Biometry and Epidemiology, University of Munich, München, Germany

Stefania Capecchi Department of Political Sciences, University of Naples Federico II, Naples, Italy

Gunnar Carlsson Department of Mathematics, Stanford University, Stanford, CA, USA

Theodore Chadjipadelis Aristotle University of Thessaloniki, Aristotle University Campus, Thessaloniki, Greece

Marcello Chiodi Department of Economics, Management and Statistics, University of Palermo, Palermo, Italy

Teresa Costa Facultat d'Economia i Empresa, Universitat de Barcelona, Barcelona, Spain

Renato Cordeiro de Amorim School of Computer Science, University of Hertfordshire, Hatfield, UK

Patrizia de Luca Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università di Trieste, Trieste, Italy

Francesca Di Salvo Department of Economics, Management and Statistics, University of Palermo, Palermo, Italy

Vincent Egan Department of Psychiatry and Applied Psychology, University of Nottingham, Nottingham, UK

Abdellatif El Idrissi LabTIC Laboratory, ENSA-Tangier, Tangier, Morocco

Elaine Fehrman Men's Personality Disorder and National Women's Directorate, Rampton Hospital, Retford, Nottinghamshire, UK

Mario Fordellone Sapienza University of Rome, Rome, Italy

Josep Fortiana Facultat de Matemàtiques, Universitat de Barcelona, Barcelona, Spain

Krzysztof Gałuszka Department of Finance, University of Economics in Katowice, Katowice, Poland

Álvaro Gómez-Losada Department of Statistics and Operational Research, University of Seville, Seville, Spain

Alexander N. Gorban Department of Mathematics, University of Leicester, Leicester, UK

Tomasz Górecki Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Myriam Hatz Department of Medical Informatics, Biometry and Epidemiology, University of Munich, München, Germany

Khadidja Henni Department of Computer Science, University of Sciences and Technologies Oran "Mohamed Boudia" USTO-MB, Bir El Djir, Algeria

Jaroslav Hlinka Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

National Institute of Mental Health, Klecany, Czech Republic

Lynette A. Hunt University of Waikato, Hamilton, New Zealand

Tadashi Imaizumi School of Management and Information Sciences, Tama University, Tokyo, Japan

Jan Kalina Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

National Institute of Mental Health, Klecany, Czech Republic

Jadwiga Kostrzewska Department of Statistics, Cracow University of Economics, Cracow, Poland

Maciej Kostrzewski Department of Econometrics and Operational Research, Cracow University of Economics, Cracow, Poland

Agnieszka Kozera Poznań University of Life Sciences, Poznań, Poland

Mirosław Krzyśko Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Daniela Lucini BIOMETRA Department, University of Milan, Milano, Italy

Aleksandra Łuczak Poznań University of Life Sciences, Poznań, Poland

Vladimir Makarenkov Département d'informatique, Université du Québec à Montréal, Montreal, QC, Canada

Geoffrey J. McLachlan Department of Mathematics, University of Queensland, St Lucia, QLD, Australia

Facundo Mémoli Department of Mathematics, The Ohio State University, Columbus, OH, USA

Department of Computer Science, The Ohio State University, Columbus, OH, USA

Evgeny M. Mirkes Department of Mathematics, University of Leicester, Leicester, UK

Boris Mirkin Department of Data Analysis and Machine Intelligence, National Research University Higher School of Economics, Moscow, Russia

Department of Computer Science and Information Systems, Birkbeck University of London, London, UK

Odysseas Moschidis University of Macedonia, Thessaloniki, Greece

Ahmed Moussa LabTIC Laboratory, ENSA-Tangier, Tangier, Morocco

Awaz K. Muhammad Department of Mathematics, University of Leicester, Leicester, UK

Fionn Murtagh University of Derby, Derby, UK

Goldsmiths University of London, London, UK

Miki Nakai Department of Social Sciences, College of Social Sciences, Ritsumeikan University, Kyoto, Japan

Atsuho Nakayama Tokyo Metropolitan University, Hachioji-shi, Japan

Shu Kay Ng School of Medicine and Menzies Health Institute Queensland, Griffith University, Nathan, QLD, Australia

Massimo Pagani BIOMETRA Department, University of Milan, Milano, Italy

Francesco Palumbo Federico II University of Naples, Napoli, Italy

Pietro Patricola Department of Economics, Management and Statistics, University of Palermo, Palermo, Italy

Barbara Pawełek Department of Statistics, Cracow University of Economics, Cracow, Poland

Domenico Piccolo Department of Political Sciences, University of Naples Federico II, Naples, Italy

Alejandro Ribeiro Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

Rosaria Romano University of Calabria, Cosenza, Italy

Gabriella Schoier Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università di Trieste, Trieste, Italy

Santiago Segarra Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

Rosaria Simone Department of Political Sciences, University of Naples Federico II, Naples, Italy

Nadia Solaro Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Milano, Italy

Luisa Stracqualursi Department of Statistical Sciences, University of Bologna, Bologna, Italy

Nadia Tahiri Département d'informatique, Université du Québec à Montréal, Montreal, QC, Canada

Brigitte Vannier Receptors, Regulation and Tumor Cells (2RCT), University of Poitiers, Poitiers, France

Waldemar Wołyński Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Feliks Wysocki Poznań University of Life Sciences, Poznań, Poland

Lynda Zaoui LSSD Laboratory, Department of Computer Science, University of Science and Technology, Oran, Algeria

Part I
Classification Methods for High Dimensional Data

Missing Data Imputation and Its Effect on the Accuracy of Classification

Lynette A. Hunt

Abstract Multivariate data sets frequently have missing observations scattered throughout the data set. Many machine learning algorithms assume that there is no particular significance in the fact that a particular observation has an attribute value missing. A common approach in coping with these missing values is to replace the missing value using some plausible value, and the resulting completed data set is analysed using standard methods. We evaluate the effect that some commonly used imputation methods have on the accuracy of classifiers in supervised leaning. The effect is assessed in simulations performed on several classical datasets where observations have been made missing at random in different proportions. Our analysis finds that missing data imputation using hot deck, iterative robust model-based imputation (IRMI), factorial analysis for mixed data (FAMD) and Random Forest Imputation (MissForest) perform in a similar manner regardless of the amount of missing data and have the highest mean percentage of observations correctly classified. Other methods investigated did not perform as well.

1 Introduction

Many of the multivariate data sets collected today would have unobserved or missing observations scattered throughout the data set. These missing values can have no particular pattern of occurrence. Despite the frequent occurrence of missing data, many machine learning algorithms assume that there is no particular significance in the fact that a particular observation has an attribute value missing:—the value is simply unknown, and the missing value is handled in a simple way.

With many classification algorithms, a common approach that is used is to replace the missing values in the data set with some plausible value and the resulting completed data set is analysed using standard algorithms. The procedure that replaces the missing values using some value is known as imputation.

L.A. Hunt
University of Waikato, Hamilton, New Zealand
e-mail: lah@waikato.ac.nz

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_1

Some algorithms treat the missing attribute as a value in its own right, whilst other classifiers such as Naïve Bayes ignore the missing data:—If an attribute is missing, the likelihood is calculated on the observed attributes and there is no need to impute a value. Decision trees, such as C4.5 and J48 [12, 13, 23], cope with missing values for an observation by notionally splitting the observation into pieces with a weighted split, and sending part down each branch of the tree.

However, the effect the treatment of missing values has on the performance of classifiers is not well understood. The estimation of the missing values can introduce additional biases into the data depending on the imputation method used and affect the classification of the observations.

This paper analyses the effect that several commonly used methods of imputation have on the accuracy of classification when classifying data that has a known classification. In Sect. 3, we review the mechanisms that can lead to data being missing, and in Sect. 2, we review the basic strategies for handing missing data. In Sect. 5, the imputation methods considered in this paper are examined and in Sect. 6, the effect the imputation methods have on the classification accuracy of several data sets is assessed.

2 Missing Data Mechanisms

The treatment of missing data indicators as random variables which were subsequently assigned a distribution was proposed by Rubin [16]. Depending on the distribution of the indicator, three basic mechanisms were defined by Rubin [16]:

1. Missing completely at random (MCAR).
   If the missingness does not depend on the values of the data, either missing or observed, then the data are MCAR.
2. Missing at random (MAR). If the missingness depends only on the data that are observed but not on the components that are missing, the data are MAR.
3. Not missing at random (NMAR). If the distribution of the missing data depends on missing values in the data matrix, then the mechanism is NMAR.

Knowledge of the mechanism that led to the values being missing is important in choosing an appropriate analysis to use for the data [10]. Hence it is important to consider how the classifier handles the missing data to avoid bias being introduced into the knowledge induced from that classifier.

3 Strategies for Handling Missing Data

There are several basic strategies that can be used to deal with missing data in classification studies. Some of these methods were developed in the context of sample surveys and can have some disadvantages in classification.

3.1 Complete Case Analysis

Complete case analysis (also known as elimination) is an approach in which observations that have any missing attributes are deleted from the data set. This strategy may be satisfactory with small amounts of missing data. However with large amounts of data, it is possible to lose considerable sample size. The critical concern with this strategy is that it can lead to biased estimates as it requires the assumption that the complete cases are a random subsample of the original observations. The completely recorded cases frequently differ from the original sample.

3.2 Available Case Analysis

Available case analysis is another approach that can be used. As this procedure uses all observations that have values for a particular attribute, there is no loss of information as all cases are used. However, the sample base changes from attribute to attribute depending on the pattern of missing data, and hence any statistics calculated can be based on different numbers of observations. The main disadvantage to this approach is that the procedure can lead to covariance and correlation matrices that are not positive definite, see, for example, [7]. This approach is used, for example, by Bayesian classifiers.

3.3 Weighting Procedures

Weighting Procedures are another approach to dealing with missing data. This approach is frequently used in the analysis of survey data. In survey data, the sampled units are weighted by their design weight which is inversely proportional to the probability of selection. Weighting procedures for non-response modify the weights in an attempt to adjust for non-response as if it were part of the sample design.

3.4 Imputation Procedures

Imputation procedures in which the missing data values are replaced with some value is another commonly used strategy for dealing with missing value. These procedures result in a hypothetical ‘complete’ data set that will cause no problems with the analysis. Many machine learning algorithms are designed to use either a complete case analysis or an imputation procedure.

Imputation methods often involve replacing the missing values with estimated values based on information that is in the data set. Many of the imputation methods are restricted to coping with one type of variable (i.e. either categorical or continuous) and make assumptions about the distribution of the data or subsets of variables. The performance of classifiers with imputed data in unreliable, and it is hard to distinguish situations in which the methods work from those in which they fail. When imputation is used, it is easy to forget that the data is incomplete [6]. However, imputation methods are commonly used in classification algorithms. There are many options that are available for imputation.

Imputation using a model-based approach is another popular strategy for handling missing data. A predictive model is created to estimate the values to be imputed for the missing values. With regression imputation, the attribute with missing data is used as the response attribute, and the remaining attributes are used as input for the predictive model. Maximum likelihood estimation using the EM algorithm [5] is one of the recommended missing data techniques in the methodological literature. This method assumes that the underlying model for the observed data is Gaussian.

Rather than imputing a single value for each missing data value, multiple imputation procedures are also commonly used. With this method, the missing values are imputed with values drawn randomly (with replacement) from a fitted distribution for that attribute. This is repeated a number, N, of times. The classifier is applied to each of the N "complete" data sets and the misclassification error is calculated. The misclassification error rates are averaged to provide a single misclassification error estimate and also estimate variances of the error rate.

Iterative regression imputation is not restricted to data having a multivariate normal distribution and can cope with mixed data. For the estimation, regression methods are usually applied in an iterative manner where each iteration uses one variable as an outcome and the remaining variables as predictors. If the outcome has any missing values, the predicted values from the regression are imputed. Iterations end when all variables in the data frame have served as an outcome.

4 Methods Used to Deal with Missing Values

The methods used in this paper for imputing the missing values are now described.

4.1 Mean and Median Imputation

Imputation of the missing value by either the mean, median or mode for the attribute are commonly used imputations. These types of imputation ignore any relationships between the variables. For mean imputation, it is well known that this method of imputation will underestimate the variance covariance matrices for that data [17].

The authors [10] also point out that with mean imputation the distribution of the "new values" is an incorrect representation of the population values as the shape of the distribution is distorted by adding values at the mean. Both mean and median imputation can only be used on continuous attributes. For categorical data, the mode is often imputed whilst using either mean or median imputation.

4.2 Hot Deck Imputation

Hot deck imputation is another imputation method that is commonly used, especially in survey samples, and it can cope with both continuous and categorical attributes. Hot deck imputation involves replacing the missing values using values from one or more similar instances that are in the same classification group. There are various forms of hot deck imputation commonly used. Random hot deck imputation involves replacing the missing value with a randomly selected value from the pool of potential donor values. Other methods known as deterministic hot deck imputation involve replacing the missing values with those from a single donor, often the nearest neighbour that is determined using some distance measure. Hot deck imputation has an advantage in that it does not rely on model fitting for the missing value that is to be imputed and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model. Further details on hot deck imputation can be found, for example, in [2].

4.3 kth Nearest Neighbour Imputation

The *k*th nearest neighbour algorithm is another method that can be used for imputation of missing values. This approach can predict both categorical and continuous attributes and can easily handle observations that have multiple missing values. This approach takes into account the correlation structure of the data. The algorithm requires the specification of number of neighbours, k, and the distance function that is to be used. The algorithm searches through the entire data set looking for most similar instances.

4.4 Iterative Model-Based Imputation

EM based stepwise regression imputation was proposed by Templ et al. [20] as a method for handling missing data. This technique for coping with missing values is an iterative model-based imputation (IRMI) that uses standard and robust methods. This algorithm has the advantage that it can cope with mixed data. In the first step of the algorithm, the missing values are initialised either using mean or KNN

imputation. The attributes are sorted according to the original amount of missing values. After the attributes are sorted, we have

$$M(x_1) \geq M(x_2) \geq M(x_3) \geq \cdots \geq M(x_p) \tag{1}$$

where $M(x_j)$ represents the amount of missing values for attribute j and where x_j is now the jth column of the data matrix. The algorithm proceeds iteratively with one variable acting as the response variable and the remaining variables as the predictors in each step of the algorithm.

The authors [20, 22] compared their algorithm with that of IVEMARE [14], an algorithm that also performs iterative regression imputation. IRMI has advantages over IVEWARE with regard to the stability of the initial values, the robustness of the imputed values and the lack of requirement at least one fully observed variable [22]. With IRMI imputation, the user can also use least trimmed squares (LTS) regression (see, for example, [15]), MM estimation ([24] and M estimation [8]).

4.5 Factorial Analysis for Mixed Data Imputation

Imputation of missing values for mixed categorical and continuous data using the principal component method "factorial analysis for mixed data" (FAMD)' was proposed by Josse and Husson [9]. See also [3]. This algorithm imputes the missing values using an iterative FAMD algorithm that uses the EM algorithm or a regularised FAMD algorithm where the method is regularised.

4.6 Random Forest Imputation

Random forest imputation was proposed by Stekhoven and Bühlmann [19] as a method to deal with missing values with mixed type data. This approach uses an iterative imputation scheme in which a random forest is trained on the observed values in the first stage, the missing vales are predicted and then the algorithm proceeds iteratively. The algorithm begins by making an initial guess for the missing values in the data matrix. This data matrix is the imputed data matrix. The guesses for the missing values could be obtained using mean imputation or some other imputation method. In the first stage, a random forest is trained on the observed values. The missing values are then predicted using the random forest that was trained on the observed values, and the imputed matrix is updated. This procedure is repeated until the difference between the updated imputed data matrix and the previous imputed data matrix increases for the first time for both the categorical and the continuous types of variables. For continuous variables, the performance of the imputation is assessed using a normalised root mean squared error [11], and for categorical values, the proportion of falsely classified entries over the categorical

missing values is used. Good performance of the algorithm gives a value that is close to 0 and bad performance gives a value close to 1. The algorithm proposed by Stekhoven and Bühlmann [19] is implemented in the R package MissForest [18]. This package also gives an estimate of the imputation error that is based in the out-of-bag error estimate from random forest.

5 The Analysis

Four classical machine learning data sets listed in Table 1 were taken. Note that the prostate cancer data set of [4] listed in [1] contained the information collected from 506 individuals; however, there were some missing values for some observations. This paper reports a complete case classification of the 12 pretrial attributes where individuals who had missing values in any of the pretrial attributes were omitted from further analysis, leaving 475 of the original 506 individuals.

For each data set, missing values were created such that the probability p of an attribute being missing was independent of all other data values where $p = 0.10$, 0.20, 0.30 and 0.50. This was repeated 20 times. The missing values generated in this fashion are missing completely at random, and the missing data mechanism is ignorable [10]. As some of the amounts of missing data are fairly extreme, it should be a good test of the types of imputation and the effect on the accuracy of a classifier.

The missing values in each data set were imputed using mean imputation, median imputation, k nearest neighbour (kNN) imputation with $k = 5$, hot deck imputation (HotD), iterative regression imputation (IRMI), the principal component method "factorial analysis for mixed data" (FAMD) and random forest imputation (MissForest). The missing values were imputed using the R packages VIM [21], HotDeckImputation, missMDA and MissForest. For data sets containing mixed data, the mode was imputed for the categorical attributes when using mean imputation for the continuous attributes.

The resulting "complete" data sets were analysed using the WEKA [23] experimenter with ten repetitions of tenfold cross-validation using several commonly used machine learning classifiers listed in Table 2. The mean percent of observations

Table 1 Datasets analysed

Datasets	Number of observations	Number of attributes	Type of attributes	Number of classes
Fisher's Iris data	150	4	Continuous	3
Pima Indian data	768	8	Mixed	2
Prostate cancer data	506[a]	12	Mixed	2
Wine data	178	13	Continuous	3

[a] This paper analyses the 475 complete cases

Table 2 Classifiers used

Classifier	Function
JRip	Uses a Ripper algorithm for fast efficient rule induction
J48	Implements C4.5 [12] decision tree learning
Naïve Bayes	Implements the standard probabilistic Naïve Bayes classification
Logistic	Builds linear logistics regression models
IBk	*k* nearest neighbours classifier
Logistic model trees (LMT)	Builds logistic model trees

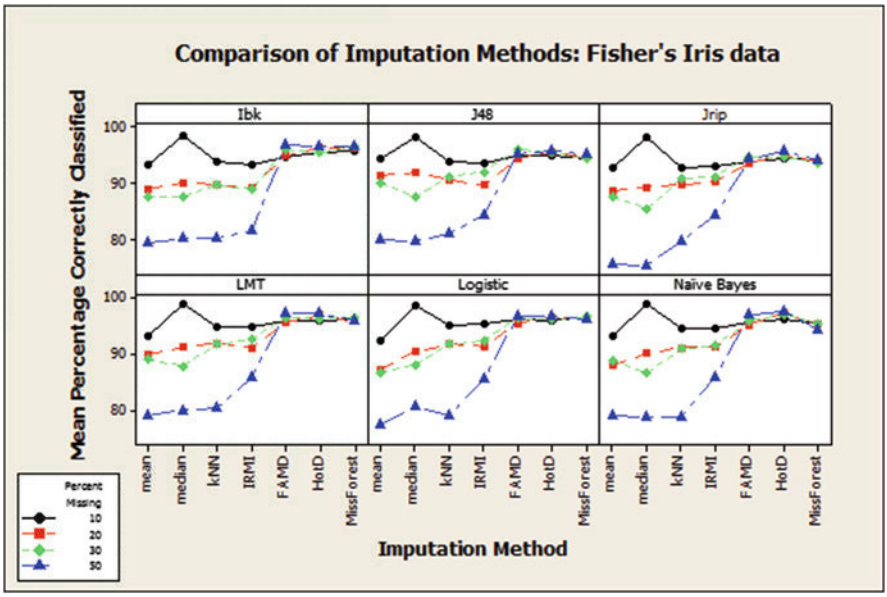


Fig. 1 Comparison of the imputation methods for Fisher's Iris data

correctly classified was recorded for each of the four amounts of missingness and each of the datasets analysed (see Figs. 1, 2, 3, 4).

It can be seen in Fig. 1 that, as the percentage of missing values increased, imputing the missing values in Fisher's Iris data using mean, median, *KNN* and IRMI imputation resulted in a decrease in the percentage of observations that were correctly classified. However with this data set, using FAMD, MissForest and Hot Deck imputation gave similar percentages of the observations correctly classified regardless of the amount of missingness in the data. The percentage correctly classified using FAMD, MissForest and Hot Deck imputation was similar to that for the complete data.

Figure 2 shows that mean, median and *KNN* imputation had similar percentages of observations correctly classified for each of the four missing data percentages, with lower percentages correctly classified as the amount of missingness in the data

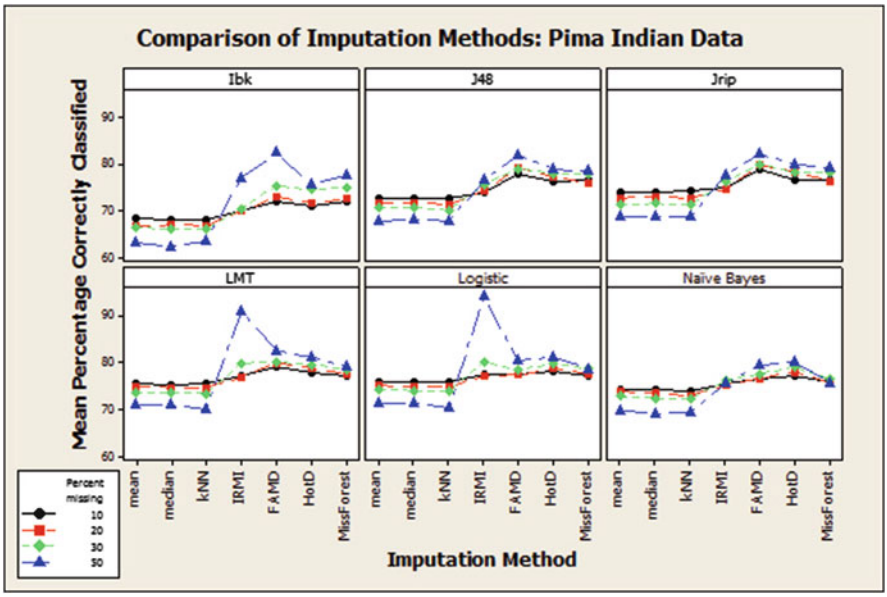


Fig. 2 Comparison of the imputation methods for Pima Indian data

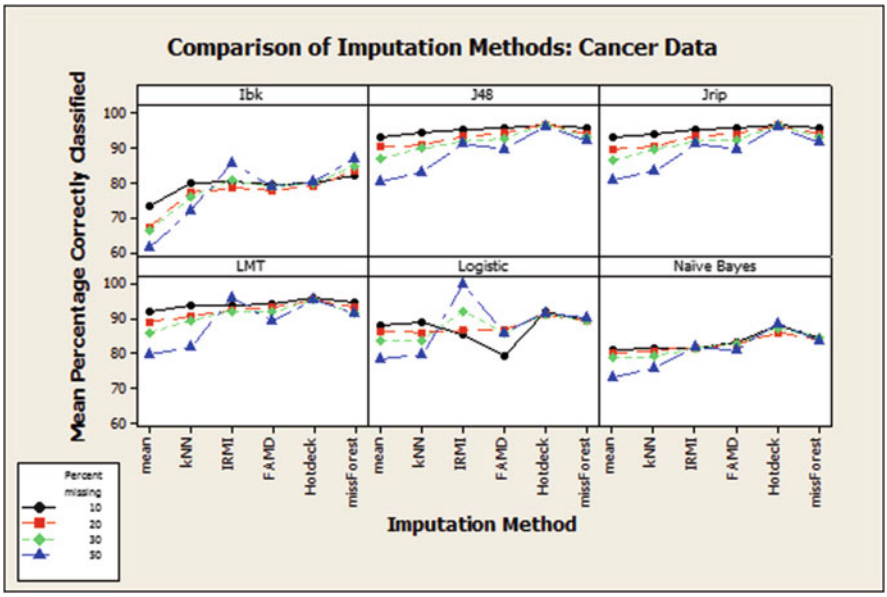


Fig. 3 Comparison of the imputation methods for prostate cancer data

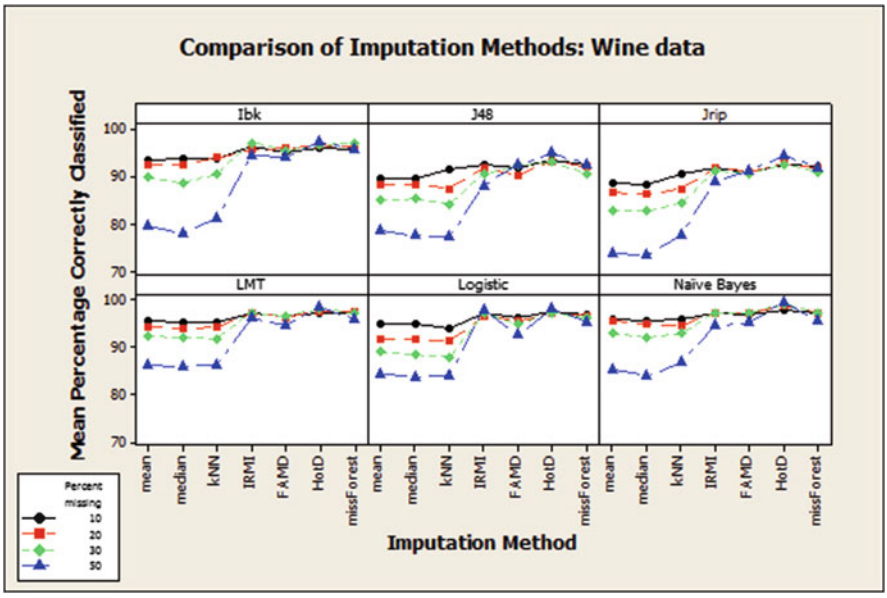


Fig. 4 Comparison of the imputation methods for wine data

increased for all the classifiers applied. FAMD, MissForest and Hot Deck imputation had similar mean percentages of the observations correctly classified regardless of the amount of missingness in the data. Overall for the Pima Indian data, the imputations using IRMI, FAMD, MissForest and Hot Deck gave consistently the highest mean percentage of observations correctly classified, with the percentage correctly classified similar to that for the complete data.

It can be seen in Fig. 3 that FAMD, MissForest and Hot Deck imputation had similar mean percentages of the observations correctly classified regardless of the amount of missingness in the data with the percentage of observations correctly classified similar to that for the complete data. Figure 4 shows that IRMI, FAMD, MissForest and Hot Deck imputations had similar mean percentages of the observations correctly classified regardless of the amount of missingness in the data, with the percentage correctly classified similar to that for the complete data.

6 Discussion

For all datasets analysed, we see that mean, median and kNN imputation have a similar mean percentage of observations correctly classified. We also see that the percentage of observations correctly classified when using the mean, median or kNN to impute the missing observations decreased as the percentage of missing data increased. In general, hot deck, IRMI, FAMD and missing Forest imputation

had the highest mean percentage of observations correctly classified, and performed in a similar manner regardless of the amount of missing data imputed.

The investigations have shown that the type of method used for imputing missing values in a data set can have an effect on the accuracy of classification. Future research needs to be undertaken on the effect of imputation on the accuracy of classification on data that has more than three classes.

References

1. Andrews, D.F., Herzberg, A.M.: Data. A Collection of Problems from Many Fields for the Student and Research Worker. Springer, New York (1985)
2. Andridge, R.R., Little, R.J.A.: A review of hot deck imputation for survey nonresponse. Int. Stat. Rev. **78**, 40–64 (2010)
3. Audigier, V., Husson, F., Josse, J.: A principal components method to impute mixed data. Adv. Data Anal. Classif. **10**, 5–26 (2016). doi:10.1007/s11634-014-0195-1
4. Byar, D.P., Green, S.B.: The choice of treatment for cancer patients based on covariate information: application to prostate cancer data. Bull. Cancer Paris **67**, 477–488 (1980)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)
6. Dempster A.P., Rubin D.B.: Introduction, Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography. Madow, W.G., Olkin, I., Rubin, D.B. (eds.), pp. 3–10. Academic, New York (1983)
7. Everitt, B.S., Dunn G.: Applied Multivariate Data Analysis. Edward Arnold, London (2001)
8. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
9. Josse, J., Husson, F.: Handling missing values in exploratory multivariate data analysis methods. J. de la Soc. Fr. de Stat. **153**(2), 1–21 (2012)
10. Little R.J.A., Rubin, D.B.: Statistical Analysis of Missing Data. Wiley New York (1987, 2002)
11. Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics **19**(16), 2088–2096 (2003). doi:10.1093/bioinformatics/btg287
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
13. Quinlan, J.R.: Improved use of continuous attributes in C4.5. J. Artif. Intell. Res. **4**, 77–90 (1996)
14. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol. **27**(1), 85–9 (2001)
15. Rousseeuw, P., Van Driessen, K.: Computing LTS regression for large data sets. Data Min. Knowl. Disc. **12**, 29–45 (2006). doi:10.1007/s10618-005-0024-4
16. Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–593 (1976)
17. Santos, R.: Effects of imputation on regression coefficients. Proc. Sect. Surv. Res. Methods Am. Stat. Assoc. 140–145 (1981)
18. Stekhoven, D.J.: Using the missForest package. https://stat.ethz.ch/education/semesters/ss2013/ams/.../missForest-1.2.pdf (2012)
19. Stekhoven, D.J., Bühlmann, P.: MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2011)
20. Templ, M., Kowarik, A., Filzmoser, P.: EM-based stepwise regression imputation using standard and robust methods. Research Report cs-2010-3, Department of Statistics and Probability Theory, Vienna University of Technology (2010)
21. Templ, M., Alfons, A., Kowarik, A., Prantner, B.: VIM: Visualization and imputation of missing values (2011). http://CRAN.R-project.org/package=VIM. R package version 3.0.0

22. Templ, M., Kowarik, A., Filzmoser, P.: Iterative stepwise regression imputation using standard and robust methods. Comput. Stat. Data Anal. **55**, 2793–2806 (2011)
23. Witten, I.H., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)
24. Yohai, V.J.: High breakdown-point and high efficiency estimates for regression. Ann. Stat. **15**, 642–656 (1987)

On Coupling Robust Estimation with Regularization for High-Dimensional Data

Jan Kalina and Jaroslav Hlinka

Abstract Standard data mining procedures are sensitive to the presence of outlying measurements in the data. Therefore, robust data mining procedures are highly desirable, which are resistant to outliers. This work has the aim to propose new robust classification procedures for high-dimensional data and algorithms for their efficient computation. Particularly, we use the idea of implicit weights assigned to individual observation to propose several robust regularized versions of linear discriminant analysis (LDA), suitable for data with the number of variables exceeding the number of observations. The approach is based on a regularized version of the minimum weighted covariance determinant (MWCD) estimator and represents a unique attempt to combine regularization and high robustness, allowing to down-weight outlying observations. Classification performance of new methods is illustrated on real fMRI data acquired in neuroscience research.

1 Robustness and Regularization of Classification Methods

Classification methods (classifiers) have the aim to automatically assign new data to one of K groups ($K \geq 2$) based on decision rules constructed over a training data set. Sensitivity (non-robustness) of standard classifiers to the presence of outlying measurements (outliers) in the data has been repeatedly reported as a serious problem [3] and robust classification methods have been proposed as alternatives, which are resistant to the presence of outliers [8].

Linear discriminant analysis (LDA) as a standard (supervised) classification method assumes the data in each group to come from a Gaussian distribution, while the covariance matrix Σ is the same across groups. Its pooled estimator denoted by S is singular for high-dimensional data with $n < p$ or even $n \ll p$. For such data, which commonly appear in a variety of applications (e.g., in medicine, molecular

J. Kalina (✉) • J. Hlinka
Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

National Institute of Mental Health, Topolová 748, 250 67 Klecany, Czech Republic
e-mail: kalina@cs.cas.cz; hlinka@cs.cas.cz

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_2

genetics, chemometrics, or econometrics), regularized versions of LDA have been proposed to avoid the curse of dimensionality. They have become popular tools with a clear comprehensibility.

One common approach to regularized LDA is known as shrunken centroid regularized discriminant analysis (SCRDA) [4]. In this context, regularization brings benefits from both the computational and statistical point of view [13], which is true for $n < p$, as well as for $n > p$ with a relatively small n [5]. Its results may be superior to approaches based on a prior dimensionality reduction performed by selection of the most relevant variables.

However, regularized versions of LDA are sensitive to the presence of outlying values in the data. Unfortunately, most robust versions of LDA, which have been proposed within the framework of robust statistics, are computationally infeasible for $n < p$ [3, 8]. Xanthopoulos et al. [19] estimated high-dimensional covariance matrices allowing for measurement errors in the observed data. The resulting estimates are robust (insensitive) only to noise, but not robust to the presence of outliers. Robust procedures for high-dimensional data have been considered for regression models, including the proposal of a canonical correlation analysis [18] or partial least squares [7]. In the context of estimating a covariance matrix Σ of multivariate data, nonparametric correlation coefficients have been investigated by Croux and Öllerer [2] under the assumption that Σ^{-1} is sparse, which allows interesting applications in the area of graphical modeling. None of these approaches however exploits the idea of coupling the robustness with regularizing the estimated covariance matrix.

This paper exploits principles of robust statistics with the aim to propose new robust classification methods for high-dimensional data. We work with methods which are robust in terms of the breakdown point, which can be characterized as a global measure of robustness of an estimator against severe outliers in the data [9]. Methods with a high breakdown point are commonly denoted as highly robust. We presented a detailed overview of regularized versions of LDA in [11], however without considering robustness aspects. On the other hand, our previous work [10] on robust classification methods cannot be applied to high-dimensional data. Only the current paper exploits a unique coupling of regularization for $n \ll p$ and statistical robustness, which is based on implicit weighting, and thus ensures a high breakdown point.

In Sect. 2 of this paper, several new robust regularized methods for high-dimensional data are proposed based on down-weighting less reliable observations. The following Sects. 3 and 4 illustrate various methods on two real data sets and bring a detailed discussion of the results. Finally, Sect. 5 concludes the paper.